Binaural rendering technology over loudspeakers and headphones

Dingding Yao¹,², Huaxing Xu³, Junfeng Li¹,²,* Risheng Xia¹,² and Yonghong Yan¹,²,⁴

¹Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China
²University of Chinese Academy of Sciences, Beijing 100049, China
³School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China
⁴Xinjiang Key Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China

e-mail: lijunfeng@hccl.ioa.ac.cn

Abstract: To provide listeners with an immersive listening experience, binaural rendering technology has become an important research topic, especially with the rising prominence of virtual/augmented reality in recent years. In this paper, we introduce our recent works on binaural rendering technology over headphones and loudspeakers. The first work is on crosstalk cancellation (CTC), which is critical for loudspeaker-based binaural rendering. An improved free-field CTC method is first presented, in which the head effects are formulated using an attenuation factor and a phase difference factor. A stochastic robust approximation method is then suggested to further improve its robustness against the perturbation perception caused by the listener’s head movement or rotation. The second work is on elevation perception of sound images, which is critical for headphone- and loudspeaker-based binaural rendering methods. By analyzing the elevation-dependent head-related transfer function (HRTF), a parametric elevation control approach is presented in which the key perceptual cues (i.e., spectral peaks and notches) are modeled using digital filters and controlled according to learned rules. The effectiveness and performance of the suggested algorithms are verified by subjective and objective experiments.

Keywords: Binaural rendering, Crosstalk cancellation, Stochastic robust approximation, Elevation control

PACS number: 43.66.Lj, 43.66.Pn, 43.66.Qp [doi:10.1250/ast.41.134]

1. INTRODUCTION

Under the assumption that our sound perception is controlled by the sound pressures at the ears, in theory, binaural rendering technology can simulate any virtual sound environment, which provides listeners with an immersive and authentic listening experience [1,2]. This technology has engaged scientific interest for many decades, particularly owing to the rising prominence of virtual/augmented reality in the last several years [3,4].

Basically, developments in binaural rendering technology have been evolving in two directions: headphone-based binaural rendering and loudspeaker-based binaural rendering. Because of their perfect channel separation, headphones are normally used to reproduce binaural signals. Yet for some applications such as TV and spatialized teleconferencing, headphones might not be practical or comfortable and therefore loudspeakers are also favorable.

In this paper, two practical key issues (i.e., crosstalk cancellation and elevation perception rendering) in binaural rendering over loudspeakers and headphones are addressed and associated proposed solutions are presented. The first part considers two practical concerns in crosstalk cancellation (CTC), needed in binaural rendering over loudspeakers. Firstly, an improved free-field CTC method incorporating head effect is proposed. Secondly, a stochastic approximation robust CTC method about head misalignments is presented. In the second part, contributions related to elevation perception of sound images, which is critical for both headphone- and loudspeaker-based binaural rendering methods, are presented. Additionally, the effectiveness and performance of the suggested algorithms are verified by subjective and objective experiments.

* e-mail: lijunfeng@hccl.ioa.ac.cn
2. CROSS-TALK CANCELLATION

When delivering binaural signals through loudspeakers, there is a crosstalk signal arriving at each ear from the other loudspeaker. Consequently, the major difficulty is to eliminate the unwanted crosstalk, i.e., the so-called CTC problem. Ideally, the expected signals obtained at the listener’s ears are delayed copies of the input signals. Mathematically, designing the CTC matrix is equivalent to approximating the inversion of the transfer function matrix. Even though a number of techniques for implementing CTC have been developed in the past decades, there are still some pending issues such as inversion constraints and robustness to head misalignment.

2.1. Proposed Improved Free-Field CTC Method

To perform the CTC method, it is necessary to determine the transfer functions between the loudspeakers and the listener’s ears. Ideally, the individual head-related transfer function (HRTF) (in the time domain, referred to as head-related impulse response, HRIR) is the first choice, as it provides a good performance [5]. However, owing to the complex and time-consuming measurement process, it is not practical for every listener. Instead, an HRTF measured using a manikin or from some database [6] is usually used, although the actual performance of CTC systems degrades to some extent owing to the large differences among individuals. Additionally, its use is also at the cost of a high computational load due to convolution with a long order. Because of its low computational load due to its simplicity, the free-field transfer function provides a good alternative, especially for real-time applications [7]. However, when not considering the listener’s head effects, its performance is limited.

We proposed a newly improved free-field CTC method based on a rigid spherical head model [8] and inspired by the idea given in [9], attempting to incorporate some acoustical effects of the listener’s head. In the proposed method, the transfer function between the loudspeaker and the listener’s ears is split into a pair of transfer functions: a head-independent component characterizing transmission from the loudspeaker to the nearer ear and a head-dependent component representing the head effects. Furthermore, the latter is subdivided into an attenuation factor $R$ and a phase difference factor $d_m$ owing to sound wave propagation around the rigid head between the two “ears.” The attenuation factor $R$, acoustically representing the overall values of interaural level differences (ILDs), is defined as the interaural ratio of root mean square (RMS) values of HRIRs calculated using the rigid sphere model [9]. Finally, as shown in Fig. 1, the modified transfer function is rewritten as

$$H_m(w) = \begin{cases} 
\frac{1}{r_1} e^{-jk_1d_m} R_{11} & \text{for ipsilateral ear} \\
\frac{1}{r_1} e^{-jk_1d_m} R_{21} e^{-jk_0} & \text{for contralateral ear}
\end{cases}, \quad (1)$$

where the subscript $m$ denotes the “modified free-field,” $R_{11}$, $R_{21}$ represent the corresponding attenuation factors, and $e^{-jk_0}$ denotes the phase difference between the two ears resulting from interference by the head, in which $k$, $\omega$, and $c_0$ are the wave number, angular frequency, respectively. Assuming that a head is a sphere of radius $r$, the acoustic path difference $d_m$ between two ears can be written as [10]

$$d_m = r(\theta + \sin(\theta)), \quad (2)$$

where $\theta$ denotes the azimuthal angle of the sound source and $\theta = 0$ indicates the frontal direction.

2.1.1. Experimental evaluation

In practice, CTC is solved using a prespecified transfer function matrix, and the CTC filters are calculated as the inverse of transfer function matrix. In this section, CTC based on an improved free-field transfer function is compared with CTC based on a free-field transfer function and CTC based on generic HRTFs in terms of sound source localization and sound quality evaluation. Hereafter, the three methods are referred to as IFM, FM and HFM, respectively.

A sound source localization experiment was carried out with 11 participants (aged 22 to 27; 9 males, 2 females) with normal hearing in an anechoic chamber under the “stereo dipole” setup. Figure 2 shows the results of the localization test for the three methods in terms of target azimuthal angles versus perceived angles for pink noise, woman’s speech, and music (a popular Chinese song) and the average result for the three stimuli. It can be observed that HFM has the worst performance in most cases. This could be explained by the fact that the HRTFs used in our

![Fig. 1 Simplified geometry of a two-loudspeaker system with the listener’s head modeled as a rigid sphere.](image-url)
test were generic ones. Since the HRTF varies significantly among people, generic HRTFs may introduce incorrect unmatched localization cues (FM and IFM methods do not) to some listeners (we call them “bad” listeners). In other words, generic HRTFs may not fit the “bad” listener’s head and ear shape very well and consequently fail to produce good spatialized sound for these listeners. In contrast to FM, the results of IFM seem to show a slightly better performance, except at a very few angles where the performance is almost the same. This result demonstrated that by integrating common perceptual localization cues encoded in the rigid-sphere head model into a free-field transfer function, the performance of localization can be enhanced. The listening results of the three methods were further examined using the three-way analysis of variance (ANOVA). No statistically significant difference among the CTC method, stimulus, and direction angle \( F(16, 160) = 1.050, p = 0.4077 \) was observed. From a statistical point of view, the proposed free-field CTC method integrating head effects on the whole achieves a performance comparable to those of state-of-the-art methods.

A sound quality evaluation experiment was performed with 11 participants (aged 22 to 27; 9 males, 2 females) in accordance with the ITU-R BS.1534 MUSHRA (MUltiple Stimulus with Hidden Reference and Anchors) Recommendation [11]. The MUSHRA scores with 95% confidence intervals for the different signals are given in Fig. 3. The reference signal is the original sound source signal and the anchor signal is obtained by filtering the reference signal (i.e., the original sound source signal) with a low-pass filter with a cutoff frequency of 3.5 kHz [11]. Evidently, most of the listeners were able to find the reference and anchor signals, confirming the validity of the test. Among the three methods, the results of HFM were the worst. Most likely, this is partly due to nonindividualized HRTFs at high frequencies and partly due to the nonflat frequency characteristics of HRTFs, especially at high frequencies [12]. Moreover, IFM outperforms the FM. As such, the subjective tests clearly demonstrate that the proposed IFM yields much smaller spectral coloration than the other methods.
Regarding the localization performance, there is little perceptual difference between these three methods, although the proposed method shows a slight increase in performance. However, for the sound quality test, the proposed method outperforms the other two methods by a much greater margin. Combining the results of the two experiments, the experiments proved the effectiveness of the proposed approach in terms of both CTC and preservation of audio quality.

2.2. Proposed Stochastic Approximation Method for CTC

Generally, CTC is optimized to achieve optimum cancellation for a given transfer function matrix corresponding to a nominal listener’s position. However, practically, the performance degrades significantly when the head is laterally misaligned. Previous research considering the dynamic range loss and the condition number has shown that an arrangement of multiple loudspeakers (circular array [13] or linear array [14]) is more robust to such perturbations than a two-loudspeaker setup. However, even for reproduction with such multiple loudspeakers, it is still imperative to provide some inherent robustness in the design of CTC filters. In this part, we attempt to achieve this goal on the basis of statistical modeling.

In the time domain, the CTC problem can be described in matrix form as follows [15]:

$$ Ac = b, $$

where $A$ represents the transfer function matrix consisting of the transfer function between the loudspeakers and the listener’s ears, $c$ is the vector of coefficients of the CTC filters and $b$ is the ideal desired pure time-delay response. Consequently, CTC can be interpreted as an optimization problem with different forms (for example, Minimax [16], LMS [17]). Statistically, changes in the transfer function caused by all the uncertain disturbances or errors such as head misalignment can be modeled as random variables subject to a certain distribution.

Assuming that $A$ is a random variable taking values in $R^{m \times n}$ with mean $\bar{A}$, $A$ can be described as $A = \bar{A} + U$, where $U$ is a random matrix with zero mean and $\bar{A}$ represents the average value of $A$ corresponding to the nominal position. Using the expected value as the objective function in the $l_2$ norm, we consider

$$ \min E[\|Ac - b\|_2^2]. $$

(4)

Also the following can be derived:

$$ E[\|Ac - b\|_2^2] = E[(\bar{A}c - b + Uc)^T(\bar{A}c - b + Uc)] = (\bar{A}c - b)^T(\bar{A}c - b) + E[c^TU^TUc] $$

$$ = \|\bar{A}c - b\|_2^2 + c^TPc, $$

(5)

where $P = E[U^TU]$ corresponds to the mathematical expectation of the autocorrelation matrix of the perturbation matrix $U$. Therefore, the statistical robust approximation problem has the form of a regularized least-squares problem [18,19] (we call it stochastic LMS hereafter),

$$ \min \|\bar{A}c - b\|_2^2 + \|P^{1/2}c\|_2^2, $$

(6)

with the analytical solution

$$ c_{opt} = (A^T \bar{A} + P)^{-1} A^T b. $$

(7)

2.2.1. Modeling the random perturbation matrix

Without loss of generality, we model the perturbation $\xi_i^L (i = 1, 2)$ (modeling $\xi_i^R (i = 1, 2)$ similarly) as a statistical variable with zero mean and variance $\sigma_i^2$ ($i = 1, 2$). The perturbed transfer function is expressed as $u_i = \xi_i^L h_i^L (i = 1, 2)$. Here, the subscripts $i = 1, 2$ represent the left and right loudspeakers and the superscripts L, R represent the listener’s left and right ear, respectively. Together, $h_i^L$ denotes the perturbation on the transfer function from the left ($i = 1$) or right ($i = 2$) loudspeaker to the listener’s left ear. The meanings of other symbols can be similarly deduced. We consider the total uncertainty regarding the listener’s head mismanagement in practice and further assume that all the perturbation random variables are independent and identically distributed (IID) with zero mean and variance $\sigma$. Finally, after some manipulation, the matrix $P$ is expressed as

$$ P = \sigma^2 \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix}, $$

where $P_{11} = (\bar{A}^L)^T \bar{A}^L + (\bar{A}^R)^T \bar{A}^R$ and $P_{22} = (\bar{A}^L)^T \bar{A}^L + (\bar{A}^R)^T \bar{A}^R$.

2.2.2. Simulations and analysis

Under two arrangement conditions with a set of small head horizontal misalignments ($x_a = 1, 2, 3, 4 \text{ cm}$), we compared the proposed stochastic LMS method with the traditional LMS method utilizing channel separation (CHS) as the evaluation measure. All of the design parameters were the same as in [15]. Specifically, the loudspeakers were separated by a distance of 0.1 m. Two span angles ($10^\circ, 20^\circ$) (with the head placed symmetrically between the loudspeakers) were simulated. The CHSs were computed above 200 Hz because of inherent difficulties at low frequencies [15]. In this case, we only consider the reproduction of the left ear’s signal owing to symmetry, i.e., ideally the left ear’s signal is a pure-delay impulse signal and no signal is produced at the right ear. Mathematically, CHS is expressed as

$$ CHS = 20 \log \left| \frac{b_L(k)}{b_R(k)} \right|, $$

where $k$ denotes different discrete frequencies and $b_L(k)$ and $b_R(k)$ represent the frequency responses of left and
right ears’ signals, respectively. The average CHS is defined as

$$CHS = \frac{1}{n_L - n_H + 1} \sum_{k=n_L}^{n_H} CHS(k), \quad (10)$$

where $n_L$ and $n_H$ are the entire frequency ranges of interest.

The average CHSs designed in accordance with the two methods with different $\sigma$ are shown in Fig. 4 for the $10^\circ$ arrangement. Specifically, the solid and dashed lines represent the CHSs for the stochastic LMS method and the traditional LMS method, respectively. It can be observed that in the vicinity of $\sigma = 0.1$, the average CHS of the proposed stochastic LMS CTC method is higher than that of the corresponding traditional LMS method, demonstrating that the proposed method is robust against small movements of the listener’s head. Furthermore, the average CHSs of the two methods with the variance $\sigma = 0.1$ are illustrated in Table 1 along with the improvement, which clearly confirms the enhanced robustness of the proposed method against head misalignment perturbations.

For further comparison, the arrangement angle was increased to $20^\circ$. Following a similar process, we repeated the experiment with the other parameters kept the same. Figure 5 presents the results for the average CHSs of the stochastic LMS method and the traditional LMS method. It can still be seen that, in the vicinity of $\sigma = 0.15$, the average CHS of the proposed stochastic LMS crosstalk cancellation method is higher than that of the traditional LMS method, in good agreement with the first experiment. Likewise, when the variance $\sigma$ is 0.15, the average CHSs of the two methods are illustrated in Table 2 along with the improvement. This case also demonstrates the enhanced robustness of the stochastic LMS method.

### 3. ELEVATION PERCEPTION

Regardless of the use of headphones or loudspeakers, the first requirement in binaural rendering is to synthesize binaural signals in accordance with the expected spatial sound scene. To synthesize an immersive and natural sound, the binaural room impulse response (BRIR) is widely applied. However, most of the measured BRIR databases do not have elevation angles because of the limitation of the recording equipment [20]. The controllability of the elevation of the reproduced source is very important. To address this issue, we have developed a more efficient elevation control method divided into two parts, as shown in Fig. 6. The left part is for spectral feature extraction and modeling, while the right part is the modification process used to extend the horizontal-plane BRIR to the elevation of $\theta$.

#### 3.1. Extraction of Perceptually Dominant Spectral Cues

It is well known that spectral cues play the dominant role in vertical localization and that the notches in HRTFs are the main cue for perception of elevation [10,21,22]. Furthermore, it has also been revealed that the frequency of

---

**Table 1** CHS values with $\sigma = 0.10$ for $10^\circ$ arrangement.

| Misalignment | Traditional | Proposed | % improvement |
|--------------|-------------|----------|---------------|
| 1 cm         | 23.33 dB    | 24.05 dB | 3.07          |
| 2 cm         | 17.09 dB    | 18.80 dB | 9.94          |
| 3 cm         | 13.17 dB    | 15.23 dB | 15.64         |
| 4 cm         | 10.17 dB    | 12.41 dB | 21.99         |

**Table 2** CHS values with $\sigma = 0.15$ for $20^\circ$ arrangement.

| Misalignment | Traditional | Proposed | % improvement |
|--------------|-------------|----------|---------------|
| 1 cm         | 18.49 dB    | 19.20 dB | 3.78          |
| 2 cm         | 13.73 dB    | 15.35 dB | 11.72         |
| 3 cm         | 11.61 dB    | 13.39 dB | 15.27         |
| 4 cm         | 9.35 dB     | 11.13 dB | 19.02         |
the notch varies smoothly with the elevation, whereas the
spectral peaks do not show this smooth trend.

By extracting the peaks and notches, the HRTF can be
decomposed into peaks and notches [23]. Specifically, the
first two spectral peaks and two spectral notches (herein-
after N1, N2, P1, and P2) are selected as spectral features.
Then, each peak or notch can be described by a second-
order notch or peak filter [24] defined by three parameters:
central frequency \( f_c \), bandwidth \( B \), and gain \( G \) [25]. A
spectral notch parameterized by a second-order notch filter
is mathematically defined as

\[
H_{\text{notch}}(z) = \frac{1 + (1 + k)\frac{H_0}{2} + d(1 - k)z^{-1} + (\frac{k}{2} - (1 + k)H_0 z^{-2})}{1 + d(1 - k)z^{-1} - k z^{-2}},
\]

where

\[
\begin{align*}
V_0 &= 10^{\frac{G}{20}} \\
H_0 &= V_0 - 1 \\
d &= -\cos\left(\frac{2\pi}{\frac{f_c}{f_0}}\right) \\
h &= \frac{\tan\left(\frac{\pi}{\frac{B}{f_c}}\right) - V_0}{\tan\left(\frac{\pi}{\frac{B}{f_c}}\right) + V_0}
\end{align*}
\]

A spectral peak parameterized by a second-order peak filter
is mathematically defined as

\[
H_{\text{peak}}(z) = \frac{V_0(1 - h)(1 - z^{-2})}{1 + 2dhz^{-1} + (2h - 1)z^{-2}},
\]

where

\[
\begin{align*}
h &= \frac{1}{1 + \tan\left(\frac{\pi}{\frac{B}{f_c}}\right)} \\
d &= -\cos\left(\frac{2\pi}{\frac{f_c}{f_0}}\right) \\
V_0 &= 10^{\frac{G}{20}}
\end{align*}
\]

and \( f_0 \) is the sampling frequency. \( f_c, B, \) and \( G \) in
the equations above are the parameters of the corresponding
peak or notch.

To reconstruct the HRTF, two distinct filter blocks are
used [26], one accounting for peaks and one for notches.
The general model for the reconstruction is shown in
Fig. 7.

### 3.2. Parameterization of Peaks and Notches

The center frequencies of notches \( N_1 \) (a) and \( N_2 \) (b)
and peaks \( P_1 \) (c) and \( P_2 \) (d) along with elevations for 21
subjects are plotted in Fig. 8. Although there are small
differences among individuals, the figure clearly shows
that the center frequencies of notches \( (N_1 \) and \( N_2)\) tend
to increase with elevation and that those of peaks \( (P_1 \) and \( P_2)\)
almost remain the same. Each peak or notch can be
formulated as a function of \( f_c, B, \) and \( G \). For a given
elevation \( \theta \), each parameter \( Q \in \{f_c, B, G\} \) can be
estimated using the following “formula”:

\[
\tilde{Q}(\theta, A) = \sum_{i=0}^{K} a_i \theta^i,
\]

where \( m \in \{N_1, N_2, P_1, P_2\} \) is the spectral cue, \( A = \{a_0, a_1, \cdots, a_K\} \) denotes the linear regression coefficients,
and \( K \) is the order of linear regression.

Furthermore, considering the independence among the
different parameters \( (f_c, B, \) and \( G) \) for each spectral cue,
the regression coefficients $A$ can be obtained in the sense of the least-square error given by

$$
\hat{A} = \sum_{r \in \Omega} \sum_{\theta \in \Theta} (\Omega(\theta) - \hat{Q}(\theta, A)),
$$

where $Q(\theta)$ denotes the parameters of the measured HRTFs at the elevation of $\theta$, and $\Omega$ and $\Theta$ are the sets of HRTFs and elevations, respectively.

The obtained linear regression coefficients were used to formulate the elevation control rules, which describe the changes in parameters of the spectral cues and elevation, and will be used for elevation control in the following section.

3.3. Modification of Peaks and Notches

According to Asano et al. [27], the macroscopic patterns of the HRTF dominate vertical localization. Thus, we modified the peaks and notches of the macroscopic patterns while keeping the microscopic patterns unchanged to retain the acoustic properties of the original BRIR. For a given horizontal-plane BRIR, the first 2 ms of the BRIR was selected as the direct component, which was regarded as the HRIR [20]. Firstly, the peaks (P1 and P2) and notches (N1 and N2) were extracted, and the associated parameters ($f_c$, $B$, and $G$) were obtained. Secondly, these parameters of the direct components were substituted into the reconstruction model to synthesize the reconstructed spectrum. By subtracting the original spectrum, we obtained the residual $d$ (the microscopic fluctuations).

Assuming that $\tilde{Q}(0)$ denotes the parameters of the direct measured BRIR components with elevation $0^\circ$, the parameters $\tilde{Q}(\theta)$ of the direct BRIR components with elevation $\theta$ can be given by

$$
\tilde{Q}(\theta) = \tilde{Q}(0) + [\tilde{Q}(\theta, A) - \tilde{Q}(0, A)],
$$

where the second term in the right side describes the differences in the spectral parameters between the target elevation and the measured elevation ($0^\circ$) of the HRTFs. By adding the resynthesized pinna spectrum with the microscopic fluctuations $d$, the direct component of the BRIR is modified to elevation $\theta$. Lastly, by adding the modified direct components and the remaining part of the original BRIR, the final BRIR at elevation $\theta$ is obtained.

3.4. Experimental Evaluation

Nine young adults (6 males and 3 females) with normal hearing served as paid volunteers. Four sources (gunshot, male speech, a piece of guitar music and Gaussian noise) were chosen as stimuli. In the experiment, two BRIR databases were selected and the HRTF database of Center for Image Processing and Integrated Computing (CIPIC) was employed to analyze the spectral features specific to elevation across all subjects.

![Fig. 9 Responses to resynthesized BRIR.](image)

The sound localization results averaged across all subjects, stimuli, and azimuth directions are plotted in Fig. 9, in which the circles represent the responded elevation distributed for different target elevations and the diameter of each circle is proportional to the number of responses with a resolution of $5^\circ$. The results show that, for most of the test elevations, the perceived elevations approximate the target elevations. For a few cases, such as the elevations of $67.5^\circ$ and $80^\circ$, the localization performance decreased. Additionally, analyzing the distribution of the subject responses at different azimuths also shows that the majority of the responded azimuths approximate the target azimuth, with some front-back confusion.

Using BRIR type, stimulus type, and direction as the three within-subject factors, a three-way ANOVA was conducted. The results revealed significant effects of direction $[F(34, 272) = 4.523, p < 0.001]$ and no significant effects of BRIR type $[F(1, 8) = 4.322, p = 0.0640]$ or stimulus type $[F(3, 24) = 4.242, p = 0.3311]$. There are significant interactions between BRIR type and direction $[F(34, 272) = 1.932, p < 0.01]$ and between stimulus type and direction $[F(102, 816) = 1.382, p < 0.01]$.

4. CONCLUSIONS

Binaural rendering is the reproduction of realistic spatial sound images using headphones or loudspeakers and can highly improve the listening experience in immersive entertainment applications. In this paper, we introduce our recent works on binaural rendering technology over headphones and loudspeakers, namely, CTC and elevation perception rendering. Specifically, three contributions are presented. Firstly, an improved free-field CTC method incorporating head effects was proposed. The results of subjective listening tests comparing its performance with that of state-of-the-art algorithms show smaller
spectral coloration of the proposed method while maintaining comparable sound localization. Secondly, we describe a stochastic approximation CTC method with the aim of increasing the robustness against head misalignment. Thirdly, we proposed an elevation control approach for binaural reproduction by parametrically modifying the direct components of BRIRs. The listening test results demonstrated that the proposed method yields good elevation perception for most positions.

ACKNOWLEDGEMENTS

This research is partially supported by the National Key Research and Development Program (No. 2017YFB1002803) and the National Natural Science Foundation of China (Nos. 11804309, 11590770-4, 11722437, 61650202, U1536117, 61671442, 11674352, 11504406, 61601453).

REFERENCES

[1] H. Miller, “Fundamentals of binaural technology,” Appl. Acoust., 36, 171–218 (1992).
[2] B. Xie, Head-Related Transfer Function and Virtual Auditory Display (J. Ross Publishing, Plantation, Fla., 2013).
[3] Y. Huang, J. Chen and J. Benesty, “Immersive audio schemes,” IEEE Signal Process. Mag., 28, 20–32 (2011).
[4] K. Sunder, J. He, E. L. Tan and W. S. Gan, “Natural sound rendering for headphones: Integration of signal processing techniques,” IEEE Signal Process. Mag., 32, 100–113 (2015).
[5] P. Majdak, B. Masiero and J. Fels, “Sound localization in individualized and non-individualized crosstalk cancellation systems,” J. Acoust. Soc. Am., 133, 2055–2068 (2013).
[6] V. R. Algazi, R. O. Duda, D. M. Thompson and C. Avendano, “The CIPIC HRTF database,” Proc. Applications of Signal Processing to Audio and Acoustics, pp. 99–102 (2001).
[7] S. Cecchi, A. Primavera, M. Virgult, F. Bettarelli, J. Li and F. Piazza, “An efficient implementation of acoustic crosstalk cancellation for 3D audio rendering,” Proc. Signal and Information Processing (ChinaSIP), pp. 212–216 (2014).
[8] R. O. Duda and W. L. Martens, “Range dependence of the response of a spherical head model,” J. Acoust. Soc. Am., 104, 3048–3958 (1998).
[9] T. Momose, M. Otani, M. Hashimoto, M. Kayama and K. Itoh, “Adaptive amplitude and delay control for stereophonic reproduction that is robust against listener position variations,” J. Audio Eng. Soc., 63, 90–98 (2015).
[10] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, revised ed. (MIT Press, Cambridge, Mass., 1997).
[11] G. Stoll and F. Kozamernik, “A method for subjective listening tests of intermediate audio quality,” ITU Working Party (2001).
[12] A. Sitzle, “Selection and tuning of HRTFs,” Proc. 112th Audio Eng. Soc. Conv., pp. 1–14 (2002).
[13] T. Takeuchi and P. A. Nelson, “Optimal source distribution for binaural synthesis over loudspeakers,” J. Acoust. Soc. Am., 112, 2786–2797 (2002).
[14] J. Zheng, T. Zhu, J. Lu and X. Qiu, “A linear robust binaural sound reproduction system with optimal source distribution strategy,” J. Audio Eng. Soc., 63, 725–735 (2015).
[15] D. B. Ward, “Joint least squares optimization for robust acoustic crosstalk cancellation,” IEEE Trans. Speech Audio Process., 8, 211–215 (2000).
[16] H. I. Rao, V. J. Mathews and Y. C. Park, “A minimax approach for the joint design of acoustic crosstalk cancellation filters,” IEEE Trans. Audio Speech Lang. Process., 15, 2287–2298 (2005).
[17] P. Nelson, H. Hamada and S. J. Elliott, “Adaptive inverse filters for stereophonic sound reproduction,” IEEE Trans. Signal Process., 10, 1621–1632 (1992).
[18] L. E. Ghaoui and H. Lebret, “Robust solutions to least-squares problems with uncertain data,” SIAM J. Matrix Anal. Appl., 18, 1035–1064 (1997).
[19] S. Boyd and L. Vandenberghe, Convex Optimization (Cambridge University Press, Cambridge, 2004), pp. 332–333.
[20] A. Karapetyan, F. Fleischmann and J. Plogsties, “Elevation control in binaural rendering,” Proc. 140th Audio Eng. Soc. Conv. (2016).
[21] M. B. Gardner and R. S. Gardner, “Problem of localization in the median plane: Effect of pinnae cavity occlusion,” J. Acoust. Soc. Am., 53, 400–408 (1973).
[22] V. R. Algazi, C. Avendano and R. O. Duda, “Elevation localization and head-related transfer function analysis at low frequencies,” J. Acoust. Soc. Am., 109, 1110–1122 (2001).
[23] V. C. Raykar, R. Duraiswami and B. Yegnarayana, “Extracting the frequencies of the pinna spectral notches in measured head related impulse responses,” J. Acoust. Soc. Am., 118, 364–374 (2005).
[24] S. Spagnol, M. Geronazzo and F. Avanzini, “Structural modeling of pinna-related transfer functions,” Proc. Int. Conf. Sound and Music Computing (SMC 2010), pp. 422–428 (2010).
[25] U. Zölzer, DAFX: Digital Audio EEffects (John Wiley, New York, 2011).
[26] V. R. Algazi, R. O. Duda and P. Satarzadeh, “Physical and filter pinna models based on anthropometry,” Proc. 122nd Audio Eng. Soc. Conv. (2007).
[27] F. Asano, Y. Suzuki and T. Sone, “Role of spectral cues in median plane localization,” J. Acoust. Soc. Am., 88, 159–168 (1990).