B-format for binaural listening of higher order Ambisonics

Article *in* The Journal of the Acoustical Society of America · May 2013

DOI: 10.1121/1.4805446 · Source: PubMed

Kotaro "" Sonoda

University of Nagasaki

Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Signal Processing in Acoustics
Session 1pSPc: Miscellaneous Topics in Signal Processing in Acoustics (Poster Session)

1pSPc3. B-format for binaural listening of higher order Ambisonics

Ryouichi Nishimura* and Kotaro Sonoda

***Corresponding author's address: National Institute of Information and Communications Technology, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, 619-0288, Kyoto, Japan, ryou@nict.go.jp**

B-format is a four-channel signal capable of rendering a sound scene with spatial information. It can be regarded equivalent to first-order Ambisonics. Ambisonics requires a high order to contain precise spatial information, and higher-order Ambisonics requires an exponentially large amount of data. This limitation comes from the fact that the original aim of Ambisonics is to reproduce the whole sound field. However, as mobile devices are prevalent, users often listen to sound media through earphones. Because nowadays users can hold sound contents individually, one can assume that sound contents could be produced adaptively to each user. Here we propose a method to make B-format signals more suitable for individual binaural listening. We assume that the production side can capture a sound scene with higher-order Ambisonics, because it may be processed for enterprise applications. Under this assumption, the binaural signal is once generated from the higher-order Ambisonics, and then its B-format signal is obtained by inversely processing the signal, assuming the first-order Ambisonics. Computer simulations show that interaural phase differences (IPDs) are improved at a frequency region where IPD dominantly affects sound localization. Results of hearing tests are also discussed.

Published by the Acoustical Society of America through the American Institute of Physics

[DOI: 10.1121/1.4800849]
Received 22 Jan 2013; published 2 Jun 2013

INTRODUCTION

As mobile devices become widely available, it is becoming common that people use headphones or earphones to listen to music or watch TV programs or movies. Listening through headphones often creates sound images inside the head. Whereas this phenomenon might help listeners immerse themselves into a scene of multimedia contents, some might complain that such sound has a lack of spatial information, or spaciousness. This problem is solvable by adapting the audio signal to the listener, considering person's own head-related transfer functions (HRTFs). This process, called individualization, and its effects on spatial impression of the sound were reported by some researchers [1, 2]. There are also attempts to clarify the physical mechanism of generating peaks and notches of HRTFs because they are important for human sound localization but are different from person to person [3].

Audio contents are generally made in such a way as to sound good when played back through loudspeakers. This is true partly because it is impossible for packaged media like CDs or DVDs to provide individualized contents. As a result, when listening through headphones, users must listen to the contents with a lower quality than the ideal quality, especially in terms of spatial impression. Nowadays, one can purchase multimedia contents in the format of digital data through the internet. Under such an environment, it is possible to sell audio media after adaptation to each customer to provide a better spatial impression in headphone listening than that provided by mass-produced packaged media.

B-format is an audio format which can create a sound scene with a spatial impression that is better than the conventional stereo format [4]. However, the quality is not better than that of higher-order Ambisonics, because B-format is theoretically equivalent to first-order Ambisonics. Noisternig *et al.* proposed a method to convert an Ambisonic signal into a binaural signal [5, 6]. In this method, assuming a virtual loudspeaker array, a binaural signal is generated by convolving the driving signal of each loudspeaker with a head-related impulse response for that direction and then summing up all of the convolved signals. Similarly to loudspeaker listening, the higher the order of the Ambisonics that is used, the better the spatial impression of the generated binaural signal. We utilized this feature to produce individualized B-format signals to improve spatial impression in headphone listening. Fig. 1 presents an image of the whole application service considered here. It is assumed that, at the production side, a sound scene is captured with multiple microphones for higher-order Ambisonics. This assumption is reasonable because the contents can be used not only for personal use but also for enterprise applications. For personal use, contents should be delivered in a small number of channels because of the limitation in communication bandwidth and storage capacity. In this respect, we developed a method to produce a B-format signal that is individualized to each user and which is made of higher-order Ambisonics for a good spatial impression when listening through headphones.

FORMULATION

In a spherical coordinate system, sound pressure p within $r < r_l$ in a direction of (θ, ϕ) in a source-free field is determined as a solution to the Helmholtz equation, as

$$p(kr, \theta, \phi) = -\sum_{n=0}^{\infty} \sum_{m=-n}^{n} ik\phi_{nm} Y_n^m(\theta, \phi) h_n(kr_l) j_n(kr) \tag{1}$$

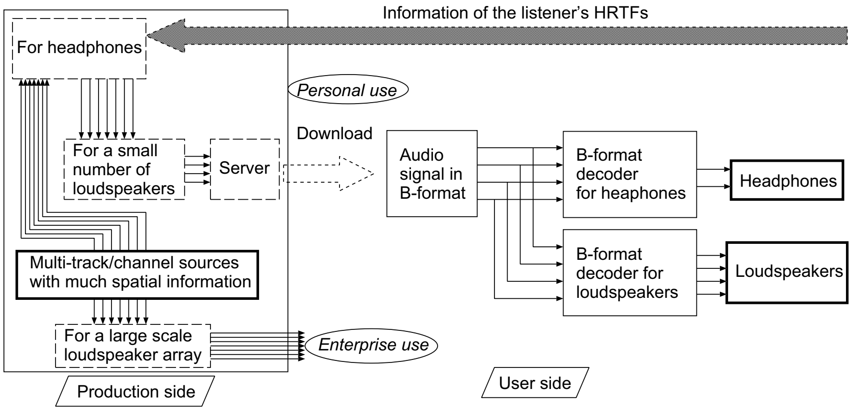

Information of the listener's HRTFs
For headphones
Personal use
Download
For a small number of loudspeakers
Server
Audio signal in B-format
B-format decoder for heaphones
Headphones
B-format decoder for loudspeakers
Loudspeakers
Multi-track/channel sources with much spatial information
For a large scale loudspeaker array
Enterprise use
Production side
User side

FIGURE 1: Application image where a sound scene is captured with multiple microphones for higher-order Ambisonics. Then, for personal use, it is delivered in the form of B-format after adaptive customization to an individual user.

where $j_n(x)$ is the spherical Bessel function, and $i = \sqrt{-1}$. In (1), $Y_n^m(\theta, \phi)$ is the spherical harmonic function defined as

$$Y_n^m(\theta, \phi) \equiv \sqrt{(2n+1)\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta)e^{im\phi} \tag{2}$$

where $P_n^m(x)$ is the associated Legendre function of order n and degree m [7]. We make the assumption that a binaural signal is represented as

$$\boldsymbol{b}(n) = \begin{bmatrix} b_l(n) \\ b_r(n) \end{bmatrix}, \tag{3}$$

where $b_l(n)$ and $b_r(n)$ respectively denote signals for the left and right ear. Similarly, a B-format signal is represented as shown below.

$$\boldsymbol{a}(n) = \begin{bmatrix} a_W(n) & a_X(n) & a_Y(n) & a_Z(n) \end{bmatrix}^T \tag{4}$$

Therein, superscript "T" represents a matrix transpose and $a_{W,X,Y,Z}(n)$ correspond to one zero and three first-order Ambisonics components. In addition, frequency representations of (3) and (4) are denoted, respectively, by $\boldsymbol{B}(w)$ and $\boldsymbol{A}(w)$. The matrix in (4) may have a larger number of components if higher-order Ambisonics is applied, although such a signal is no longer called B-format. Using these notations, a binaural signal is produced using the method proposed in [5] from an Ambisonic signal, as

$$\boldsymbol{B}(\omega) = \boldsymbol{H} \cdot \boldsymbol{Y}^+ \cdot \boldsymbol{A}(\omega), \tag{5}$$

where

$$\boldsymbol{Y}(\theta_0, \cdots, \theta_{L-1}, \phi_0, \cdots, \phi_{L-1}) = \begin{bmatrix} Y_0^0(\theta_0, \phi_0) & Y_0^0(\theta_1, \phi_1) & \cdots & Y_0^0(\theta_{L-1}, \phi_{L-1}) \\ Y_1^{-1}(\theta_0, \phi_0) & Y_1^{-1}(\theta_1, \phi_1) & \cdots & Y_1^{-1}(\theta_{L-1}, \phi_{L-1}) \\ Y_1^1(\theta_0, \phi_0) & Y_1^1(\theta_1, \phi_1) & \cdots & Y_1^1(\theta_{L-1}, \phi_{L-1}) \\ Y_1^0(\theta_0, \phi_0) & Y_1^0(\theta_1, \phi_1) & \cdots & Y_1^0(\theta_{L-1}, \phi_{L-1}) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}, \tag{6}$$

$$\boldsymbol{H}(\theta_0, \cdots, \theta_{L-1}, \phi_0, \cdots, \phi_{L-1}, \omega) = \begin{bmatrix} H_l(\theta_0, \phi_0, \omega) & H_l(\theta_1, \phi_1, \omega) & \cdots & H_l(\theta_{L-1}, \phi_{L-1}, \omega) \\ H_r(\theta_0, \phi_0, \omega) & H_r(\theta_1, \phi_1, \omega) & \cdots & H_r(\theta_{L-1}, \phi_{L-1}, \omega) \end{bmatrix}, \tag{7}$$

and where superscript "+" denotes the pseudo-inverse of the matrix. In (6) and (7), $H_l(\theta_i, \phi_i, \omega)$ is the head-related transfer function for the left ear with respect to the i-th virtual loudspeaker placed in the direction of (θ_i, ϕ_i), and L is the number of virtual loudspeakers. Similarly, $H_r(\theta, \phi, \omega)$ is the same but for the right ear. An individualized B-format signal is therefore obtained by performing this process inversely, after generating the binaural signal using (5) employing the user's own HRTFs in (7).

$$\hat{\boldsymbol{A}}(\omega) = \boldsymbol{Y}_b \cdot \boldsymbol{H}_b^{+} \cdot \boldsymbol{B}(\omega) \tag{8}$$

where $\boldsymbol{Y}_b$ is a matrix consisting of the upper four rows of (6), and where $\boldsymbol{H}_b$ is created exactly in the same way as (7). Directions of virtual loudspeakers are not necessarily the same as (6) and (8) but should be consistent between $\boldsymbol{Y}_b$ and $\boldsymbol{H}_b$. Finally, the time domain representation of the individualized B-format signal is obtained by performing inverse Fourier transform of (8).

The whole procedure described in this section can be depicted as a flowchart shown in Fig. 2, which also includes the paths used in performance evaluation discussed in the next section.

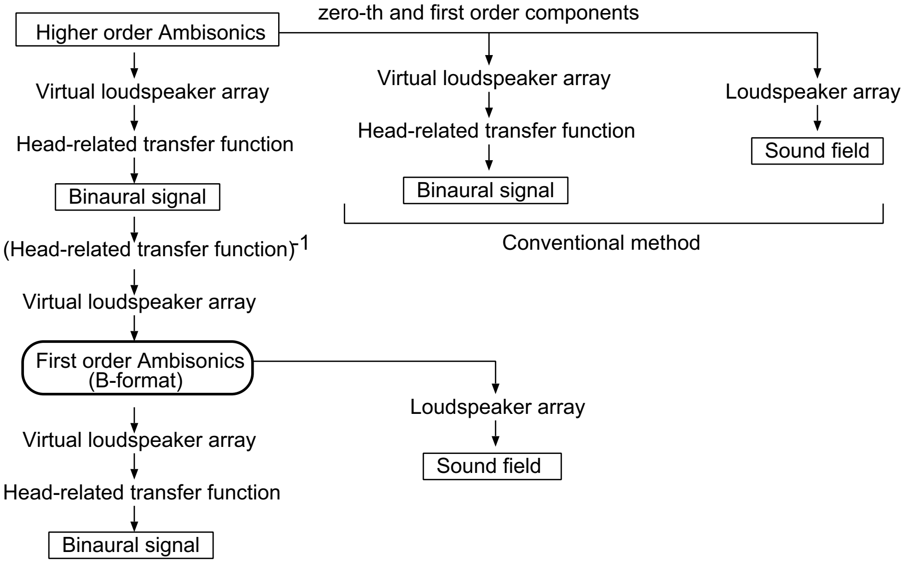

Higher order Ambisonics
zero-th and first order components
Virtual loudspeaker array
Head-related transfer function
Binaural signal
Virtual loudspeaker array
Head-related transfer function
Binaural signal
Loudspeaker array
Sound field
Conventional method
(Head-related transfer function)$^{-1}$
Virtual loudspeaker array
First order Ambisonics (B-format)
Loudspeaker array
Sound field
Virtual loudspeaker array
Head-related transfer function
Binaural signal

FIGURE 2: Flowchart of the proposed signal processing for generating an individualized B-format signal.

COMPUTER SIMULATIONS

Performance in Binaural Listening

Computer simulations were conducted to evaluate the performance of the proposed method, focusing especially on sound localization on the horizontal plane. Virtual loudspeakers were arranged every 5 deg on a circle, and a set of HRTFs of a KEMAR dummy head for the horizontal plane provided by MIT were used [8]. For the computer simulations, it was assumed that a plane wave is travelling from one of every 10 degrees and that it is captured with higher-order Ambisonics. It was then transformed into an individualized B-format signal using the proposed method. Subsequently, the binaural signal was reproduced from the created B-format signal using (5) again. The interaural level difference (ILD) and interaural phase difference (IPD) of the obtained binaural signal were calculated using

$$\mathrm{ILD}\ (\omega) = 10 \log \frac{|\hat{\boldsymbol{B}}_r(\omega)|}{|\hat{\boldsymbol{B}}_l(\omega)|}, \tag{9}$$

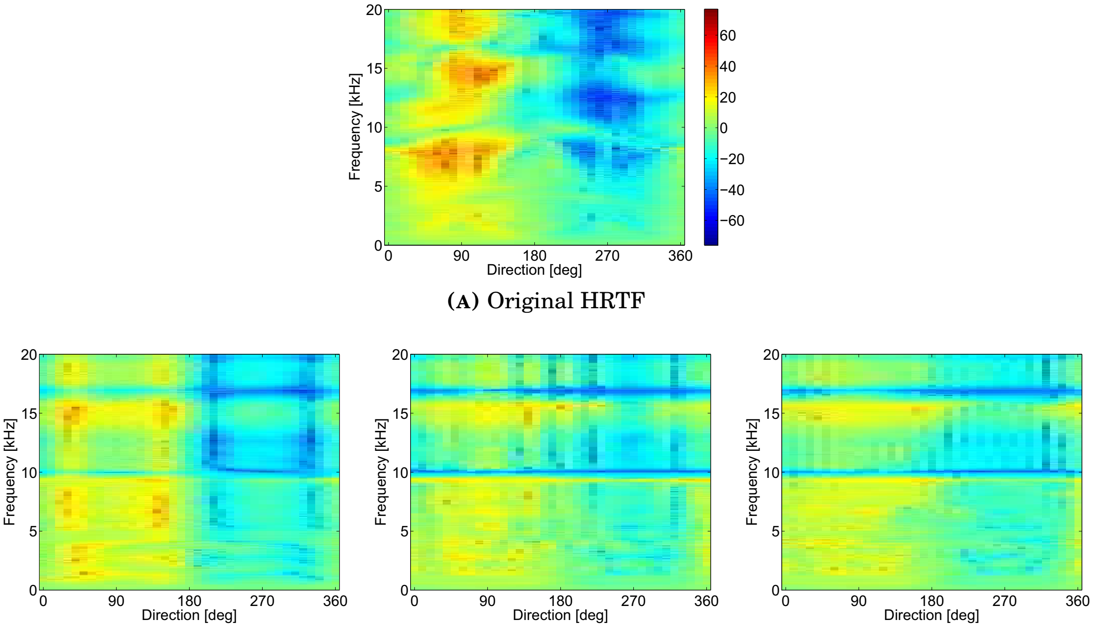
$$\mathrm{IPD}\ (\omega) = \angle\ \frac{\hat{\boldsymbol{B}}_r(\omega)}{\hat{\boldsymbol{B}}_l(\omega)}, \tag{10}$$

where

$$\hat{\boldsymbol{B}}(\omega) = \boldsymbol{H}_b \cdot \boldsymbol{Y}_b^{+} \cdot \hat{\boldsymbol{A}}(\omega). \tag{11}$$

Because neither $\boldsymbol{H}_b$ nor $\boldsymbol{Y}_b$ is a square matrix, $\hat{\boldsymbol{B}}(w)$ becomes different from $\boldsymbol{B}(w)$. Actually, advanced research has been conducted on the transcoding of B-format to a binaural signal such as [9], but the method proposed in [5] was regarded as the conventional method in the following comparisons.

Fig. 3 shows the results for ILDs. The front direction of the listener was defined as 0 deg. The angle increases clockwise. The upper panel shows the result calculated directly using the original HRTFs. The left panel on the lower row shows that for the conventional method. The other two show the proposed method but with different initial Ambisonics orders. The left half of each panel corresponds to directions on the right side of the listener. Although positive ILDs are observed for these directions as expected, it is not clear whether ILD patterns of the conventional and the proposed method resemble those of the original.



(A) Original HRTF

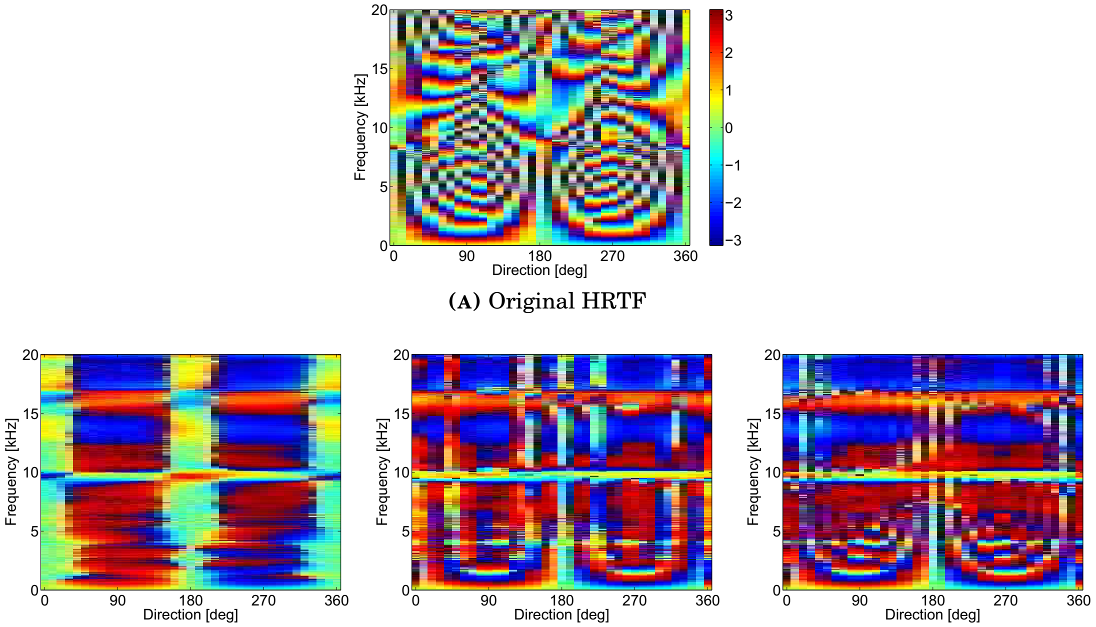
(B) Conventional

(C) Proposed (4th order)

(D) Proposed (9th order)

FIGURE 3: Interaural level differences: (A) calculated from the original HRTF; (B) conventional method; (C) and (D), the proposed method assuming that the original sound scene is captured, respectively with fourth-order and ninth-order Ambisonics.

Similarly, Fig. 4 shows IPDs. It is notable that the proposed method restores IPDs more correctly than the conventional one, especially at a low-frequency region of which the upper bound is approximately 2 kHz to 4 kHz, depending on the initial Ambisonics orders.



(A) Original HRTF

(B) Conventional

(C) Proposed (4th order)

(D) Proposed (9th order)

FIGURE 4: Interaural phase differences: (A) calculated from the original HRTF; (B) conventional method; (C) and (D), the proposed method assuming that the original sound scene is captured, respectively with fourth-order and ninth-order Ambisonics.

To evaluate the similarity of ILD or IPD patterns objectively, cross-correlations between the original HRTFs and those obtained using the conventional or the proposed method were calculated. Fig. 5 presents results represented as a function of frequency, where the upper panel shows results of ILD and the lower panel shows those of IPD. The results clarify that, although no remarkable difference exists in ILD patterns between the conventional and the proposed method, in IPD the proposed method broadens the frequency range where IPDs are restored correctly. Roughly speaking, its upper bound goes up to approximately 1 kHz for the conventional method, 2 kHz for the proposed method with initially fourth- order Ambisonics, and 4 kHz for the initially ninth-order Ambisonics. Considering that IPDs in these low-frequency regions play an important role in human sound localization, this feature of the proposed method is regarded as an advantage.

Performance in Loudspeaker Listening

The proposed method tunes B-format signals to each listener, but it invariably affects the sound in loudspeaker listening. To ascertain how the proposed method affects the sound field in loudspeaker listening, sound pressure patterns for a sinusoidal input are simulated. For this simulation, it was assumed that four loudspeakers were placed in every corner of the square and that the input signal was coming from one of the corners. Both input and loudspeaker signals were assumed as a plane wave. Fig. 6 presents results during one period of time, T, for sinusoidal input of approximately 500 Hz. The simulated area was a square of 2 m by 2 m, situated at the center of the loudspeaker array. The upper panels show those of the original sound field. The middle panels show the conventional B-format, and the lower panels show the proposed method. It appears clear that the wave front deteriorates more in the proposed method than in the conventional one. However, it seems that the sound pressure at the center point is

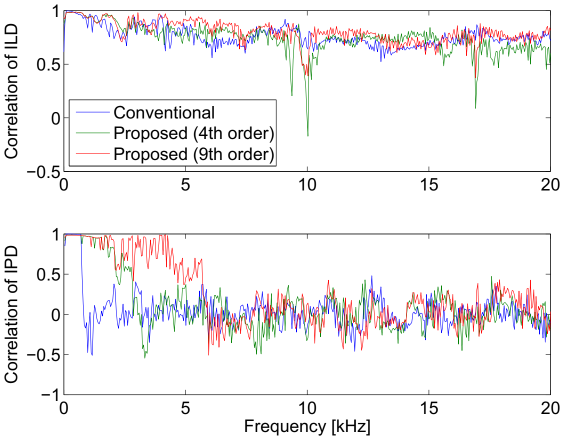
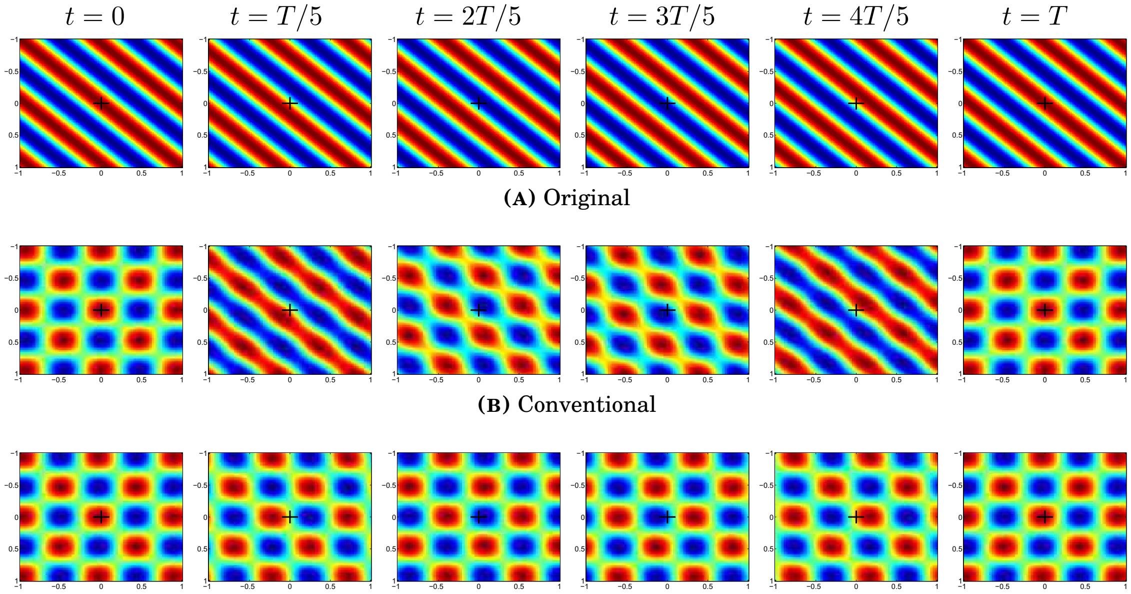


FIGURE 5: Cross-correlations between the patterns of the original HRTF and that obtained using the conventional or proposed method for ILD (upper panel) and IPD (lower panel).

reproduced correctly with both methods.



(A) Original

(B) Conventional

(C) Proposed (9th order)

FIGURE 6: Sound pressure patterns reproduced in loudspeaker listening. Each column shows those at a certain time instance during one period of time. Sound is coming from the direction of the upper right corner, and the symbol "+" indicates the center point.

Table 1 shows signal-to-noise ratios (SNRs) of sound pressure patterns where the signal is the sinusoidal input wave and the noise is the error in the sound pressure patterns reproduced by the conventional or the proposed method. The SNRs were averaged over both one period of time and the whole simulated area shown in Fig. 6. Although the conventional method achieves 6 dB to 8 dB in SNR, the proposed method is -2 dB to 3 dB. Therefore, the proposed method is inferior to the conventional method in terms of reproduction of a sound field by 5 dB to 8 dB. This is regarded as a shortcoming of the proposed method.

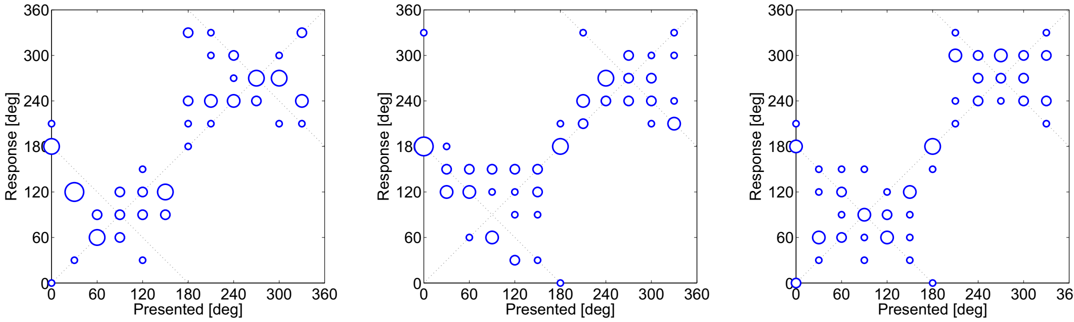
TABLE 1: Mean signal-to-noise ratio, represented in decibels, of sound pressure patterns where the signal is the sinusoidal input and the noise is the distortion in the sound pressure pattern reproduced by the conventional or the proposed method.

Method	250 Hz	500 Hz	1 000 Hz	2 000 Hz	4 000 Hz
Conventional	6.89	7.60	7.28	7.06	6.93
Proposed (4th order)	−1.85	2.36	1.22	1.58	−0.18
Proposed (9th order)	−1.84	2.36	1.21	1.57	0.01

HEARING TESTS

Listening tests were also conducted to ascertain whether restoration of binaural information by the proposed method brings about a good effect on sound localization in headphone listening. Three male and three female listeners took part in the tests. Three consecutive noise bursts of 150 ms each with intervals of 150 ms were presented to a listener through a pair of headphones (HDA-200; Sennheiser Electronic GmbH and Co. KG). Rising and falling points of each burst were smoothed by a raised cosine function of 10 ms. Sound stimuli were generated for every 30 deg on the horizontal plane. Listeners were asked to select the sound direction from 12 possible directions and mark them on an answer sheet within seven seconds. For each listener, all 12 directions and 4 methods were presented in random order. Therefore, every listener performed 48 trials in a single session. The same set of HRTFs as that used in the computer simulations was used. To let listeners be accustomed to the sound made using non-individualized HRTFs, they once listened to sound stimuli made by original HRTFs of all 12 directions with visual feedback of the presenting directions.

The results accumulated over all the listeners are presented in Fig. 7, omitting one result for the proposed method with fourth-order Ambisonics because it resembled that obtained with ninth-order Ambisonics. Localization performance was not so good in Fig. 7(A) where the original HRTFs were used, despite the training sessions in which listeners experienced those signals. This bad localization performance can be attributed to non-individualized HRTFs and the insufficiency of training. In comparison between Figs. 7(B) and 7(C), although the conventional method shows a tendency by which many listeners perceived sound stimuli as if they were coming from the rear, this tendency weakens in the proposed method. As a result, the scatter plot of responses for the proposed method more closely resembles that for the original HRTFs than the conventional method does.



(A) Original

(B) Conventional

(C) Proposed (9th order)

FIGURE 7: Results of sound localization tests where the abscissa is the presented direction and the ordinate is the response of listeners. Sizes of circles represent the number of responses.

Conclusion

A method to produce an individualized B-format signal for headphone listening was discussed, assuming that recording is conducted with higher-order Ambisonics. Computer simulations showed that the proposed method has the capability of restoring binaural information better than the conventional method, at a sacrifice of performance in reproducing sound field using loudspeakers. Results obtained through listening tests demonstrated that this capability of the proposed method can improve sound localization in headphone listening. In the future, not only sound localization but also sound quality should be investigated. Moreover, practical techniques of HRTF individualization are strongly desired because the performance of the proposed method depends on how well the HRTFs fit the listener.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 23500162.

References

[1] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space", J. Acoust. Soc. Am. **100**, 3785–3793 (1996).

[2] W. L. Martens, A. Guru, and D. Lee, "Effects of individualised headphone response equalization on front/back hemifield discrimination for virtual sources displayed on the horizontal plane", in *International Congress on Acoustics (ICA)* (2010).

[3] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida, "Mechanism for generating peaks and notches of head-related transfer functions in the median plane", J. Acoust. Soc. Am. **132**, 3832–3841 (2012).

[4] M. Gerzon, "Ambisonics. part two: Studio techniques", Studio Sound **17**, 24–30 (1975).

[5] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich, "A 3D Ambisonic based binaural sound reproduction system", in *AES 24th International Conference on Multichannel Audio* (2003).

[6] M. Noisternig, T. Musil, A. Sontacchi, and R. Höldrich, "3D binaural sound reproduction using a virtual Ambisonic approach", in *International Symposium on Virtual Environments, Human-Computer Interfaces, and Measurement Systems*, 174–178 (2003).

[7] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic Press) (1999).

[8] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone", Technical Report 280, MIT Media Lab (1994).

[9] S. Berge and N. Barrett, "A new method for B-format to binaural transcoding", in *AES International Conference* (2010).