Archives of Acoustics Vol. 47, No. 1, pp. 57–70 (2022), doi: 10.24425/aoa.2022.140732

Research Paper

Influence of Loudspeaker Configurations and Orientations on Sound Localization

Shu-Nung YAO

Department of Electrical Engineering, National Taipei University
No. 151, University Rd., Sanxia Dist., New Taipei City 237303, Taiwan; e-mail: snyao@gm.ntpu.edu.tw

(*received September 14, 2020; accepted November 14, 2021*)

As the virtual reality (VR) market is growing at a fast pace, numerous users and producers are emerging with the hope to navigate VR towards mainstream adoption. Although most solutions focus on providing high-resolution and high-quality videos, the acoustics in VR is as important as visual cues for maintaining consistency with the natural world. We therefore investigate one of the most important audio solutions for VR applications: ambisonics. Several VR producers such as Google, HTC, and Facebook support the ambisonic audio format. Binaural ambisonics builds a virtual loudspeaker array over a VR headset, providing immersive sound. The configuration of the virtual loudspeaker influences the listening perception, as has been widely discussed in the literature. However, few studies have investigated the influence of the orientation of the virtual loudspeaker array. That is, the same loudspeaker arrays with different orientations can produce different spatial effects. This paper introduces a VR audio technique with optimal design and proposes a dual-mode audio solution. Both an objective measurement and a subjective listening test show that the proposed solution effectively enhances spatial audio quality.

Keywords: audio quality; ambisonics; immersive sound; loudspeaker array; spatial effect; virtual reality.

1. Introduction

As sound localization is one of the key factors affecting immersion and presence to virtual reality (VR) content, headphone-based immersive audio is an important issue for the entertainment audio industry. The senses of sight and hearing can be equally important in some conditions. For first-person shooter video games, sound localization helps players to quickly identify where the targets are. If live music is recorded using a sound-field microphone and every part of a scene in a concert is captured by a 360° camera, the viewer can scroll around both audio and video in the recorded film. Moreover, sound-based interactive entertainment has been designed for blind and visually impaired people (Gaudy *et al.*, 2001; Gardenfors, 2003).

In VR headsets, the binaural system is usually equipped with one of the following three types of binaural system: head-related transfer functions (HRTFs) (Matsumura *et al.*, 2005), motion-tracked binaural (MTB) sound (Algazi, Duda, 2004), and ambsionics (Yao, 2017). HRTF sound uses many microprocessor resources and synthesizes a virtual acoustic space, whereas MTB sound renders only the two-dimensional (2D) auditory space and records the natural acoustic space. The ambisonics-based binaural system for VR (Yao, 2017) outperforms the other two systems. Ambisonics was originally introduced for multichannel surround sound. Because of its narrow sweet spot and hardware cost, the system was not popular when it was introduced to the market. Nowadays, many methods exist to optimize ambisonic sound quality and localization. Yao *et al.* (2015) used a spilt-band decoder to enhance the audio quality. Different loudspeaker arrays produce different frequency spectra, some of which are impaired. The spilt-band decoder selects nearly perfect reconstructed spectra from different loudspeaker arrays and combines their frequency components to produce optimal audio quality. Although a dense loudspeaker array produces more accurate low-frequency cues, it also suffers from severe spectral impairment. A 1/3-octave filter bank was employed to overcome spectrum distortion in a dense loudspeaker array (Yao, 2018). Yao (2018) described the reproduction of a sound field in a dense loudspeaker array equipped with an equalizer for spectrum compensation. The architecture of the proposed equalization ambisonics contains a controller, which is used to assess the spectrum distortion, and then, adjust the equalizer for spectrum compensation. In the binaural system, we can virtually place the loudspeakers as uniformly as possible, but an appropriate HRTF dataset must be chosen to function as a virtual loudspeaker. If the dataset is not selected correctly, HRTF mismatch causes localization blur (Yao, Chen, 2013; Yao *et al.*, 2017). Because binaural synthesis requires calibration, Yao (2017) conducted a listening test to help listeners to determine the appropriate dataset. However, the listener had to listen to several audio pieces during the calibration stage, which is time consuming. Therefore, a neural network system was proposed to achieve efficient HRTF selection (Yao *et al.*, 2017).

In this study, we investigate sound localization in different loudspeaker configurations and propose a dual-mode structure for localization enhancement. Although many papers (Scaini, Arteaga, 2014; Yao *et al.*, 2015; Yao, 2018) have discussed the placement and number of loudspeakers in an array, few studies have examined the listening perceptions of different loudspeaker array orientations. Even if the polygons of two loudspeaker arrays are the same, the orientation of the arrays can influence the perceived sound localization. Therefore, we conducted listening measurements in the same loudspeaker configuration but with different orientations. The binaural simulation was built using two popular databases from the Institute for Research and Coordination in Acoustics/Music (IRCAM) and Center for Image Processing and Integrated Computing (CIPIC) Interface Laboratory. We also propose a dual-mode decoder that obtains the optimum localization cues between two basic loudspeaker arrays.

2. Binaural ambisonics

Ambisonics is used to represent a two- or three-dimensional acoustic pressure field as a function of cylindrical or spherical harmonic components, respectively. The expression of the sound field as a combination of spherical harmonic signals was discussed in Yao *et al.* (2015). The idea that a two-dimensional spectrum can be expressed as a sum of cylindrical harmonic components was considered in Yao (2018). This section provides the fundamentals of ambisonics and describes the binaural implementation.

2.1. Two-dimensional ambisonics

Ambisonics originated from the concept of the Blumlein pair – a stereo recording technique invented by Alan Blumlein (Clark *et al.*, 1958). Two figure-of-eight microphones are positioned at 90° from each other, as shown in Fig. 1a; hence, the polar pattern of the microphone array is the dashed curve in Fig. 1b. Using this setting, the features of the two-dimensional sound field can be captured. Gerzon (1975) introduced the polar pattern of an omni-directional microphone to enhance the sound pressure inside the field. As shown in Fig. 2a, the sound fields recorded by the upper microphone and lower microphone are called the X and Y components, respectively, and that recorded by the omni-directional microphone is the W component. The polar pattern of ambisonics is composed of W, X, and Y, as shown in Fig. 2b. In addition to the microphone recording, ambisonic components can be synthesized using spherical harmonic functions. Suppose that we intend to synthesize sound S in the direction of angle θ. The W, X, and Y components can be produced by

$$W = S \cdot 0.7071, \tag{1}$$

$$X = S \cdot \cos\theta, \tag{2}$$

and

$$Y = S \cdot \sin\theta. \tag{3}$$

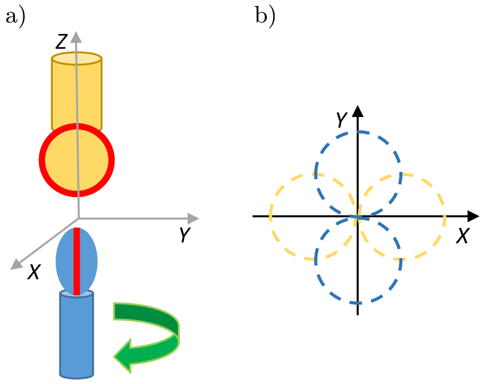


Fig. 1. Figure-of-eight microphones: a) one is perpendicular to another, b) polar pattern of the microphone array.

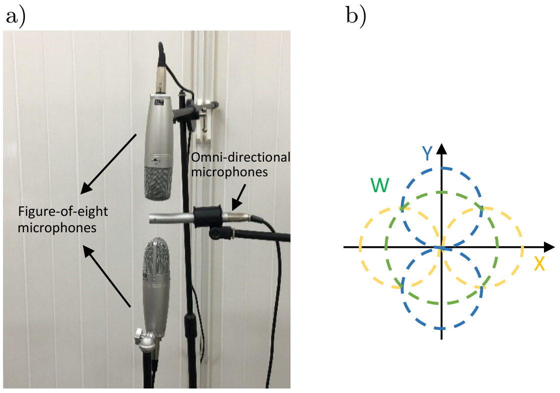


Fig. 2. Sound field microphone array by using: a) two figure-of-eight microphones and an omni-directional microphone, b) polar pattern of the first-order ambisonics.

The process of obtaining or creating the W, X, and Y components is called ambisonic encoding. After obtaining the ambisonic components, the sound field $p(k \cdot r, \partial)$ can be expressed as the first-order Fourier-Bessel series:

$$p(k \cdot r, \partial) = j_0(k \cdot r) \cdot W \cdot \frac{1}{\sqrt{2}} + i \cdot j_1(k \cdot r) \cdot (X \cdot \cos\partial + Y \cdot \sin\partial), \quad (4)$$

where k is the wave number, r is the distance from the origin, ∂ is an angle; thus, (r, ∂) defines a measurement point and usually refers to the position of a listener's left or right ear, i is $\sqrt{-1}$ and $j_m(x)$ is the m-th order spherical Bessel function defined in Eq. (5). Functions with different orders are shown in Fig. 3:

$$j_m(x) = (-1)^m \cdot x^m \cdot \left(\frac{\mathrm{d}}{x \cdot \mathrm{d}x}\right)^m \frac{\sin x}{x}. \quad (5)$$

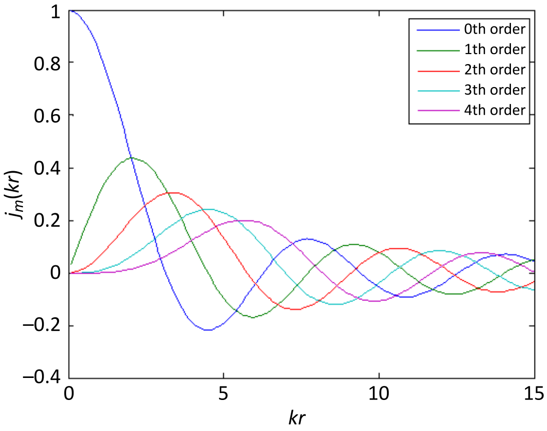


Fig. 3. The spherical Bessel functions from 0th-order to 4th-order.

In ambisonic decoding, the sound field generated by the n-th loudspeaker can also be expressed as a first-order series, $p_n(k \cdot r, \partial)$, as shown in Eq. (6), so the superposition of sound fields caused by N loudspeakers, $\sum_{n=1}^{N} p_n(k \cdot r, \partial)$, is expected to be $p(k \cdot r, \partial)$:

$$p_n(k \cdot r, \partial) = j_0(k \cdot r) \cdot W_n \cdot \frac{1}{\sqrt{2}} + i \cdot j_1(k \cdot r) \cdot (X_n \cdot \cos\partial + Y_n \cdot \sin\partial). \quad (6)$$

When we use a loudspeaker array to reproduce the sound field, the array should be shaped as a regular polygon. Although in real space, it is difficult to place a regular loudspeaker array, Scaini and Arteaga (2014) proposed an algorithm to overcome this problem. Because we use the binaural system to render a virtual acoustic space, the virtual loudspeakers can be uniformly distributed, and if the layout is a regular polygon, the signal o_n feeding the n-th loudspeaker with the location ϑ_n, expressed as in Eq. (7), can theoretically construct the sound field $p(k \cdot r, \partial)$:

$$o_n = \frac{1}{L} \cdot \left[W \cdot \left(\frac{1}{\sqrt{2}}\right) + X \cdot (\cos\vartheta_n) + Y \cdot (\sin\vartheta_n)\right]. \quad (7)$$

In the virtual acoustic space, each loudspeaker is rendered by a set of head-related impulse responses (HRIR) (Yao, Chen, 2013; Matsumura *et al.*, 2005). The space between the listener's ears is not empty, and the sound traveling in that space is expected to experience interruption or reflection. An individual's anthropometric parameters from the pinna, head, and shoulders transform an incoming sound in a direction-dependent way, and the effects are contained in HRIRs. Because the minimum number of required loudspeakers is greater than $(2M+1)$ in M-th order ambisonics (Blauert, Rabenstein, 2012), a quadraphonic loudspeaker array such as that shown in Fig. 4 is the most common first-order two-dimensional ambisonic setting. The four loudspeakers are located clockwise at 45°, 135°, 225°, and 315°. The sounds reaching the listeners' left and right eardrums, e_l and e_r, are shown in Eqs (8) and (9), where h_{Xy} is the HRIR between loudspeaker X (positions 1, 2, 3, and 4, as shown in Fig. 4), and the listener's ear y (l or r), and o_X denotes the audio signal produced by loudspeaker X. "*" symbolizes the convolution operator. Figure 5 shows a block dia-

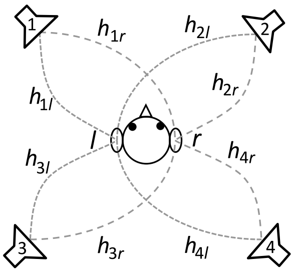


Fig. 4. Quadraphonic loudspeaker array.

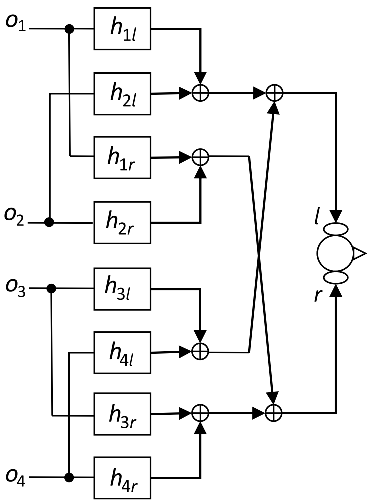


Fig. 5. The architecture for headphones delivering 1st-order 2D ambisonic surround.

gram of headphones rendering the quadraphonic ambisonic audio produced by the square loudspeaker array in Fig. 4. A two-dimensional binaural system is realized by sending convolved signals to the headphones:

$$\begin{aligned} e_l(t) &= o_1(t) * h_{1l}(t) + o_2(t) * h_{2l}(t) \\ &\quad + o_3(t) * h_{3l}(t) + o_4(t) * h_{4l}(t), \end{aligned} \tag{8}$$

$$\begin{aligned} e_r(t) &= o_1(t) * h_{1r}(t) + o_2(t) * h_{2r}(t) \\ &\quad + o_3(t) * h_{3r}(t) + o_4(t) * h_{4r}(t). \end{aligned} \tag{9}$$

2.2. *Three-dimensional ambisonics*

To record a three-dimensional acoustic space, we can use a sound field microphone built by four cardioid microphone capsules orientated left-front, right-front, left-back, and right-back with respect to the recording engineer, as shown in Fig. 6 (where LF, RF, LB, and RB symbolize the sounds recorded by each capsule, respectively). Then, the amplifiers inside a sound field microphone provide the sum and difference functions for each channel. The process can be expressed by the following equations (Rumsey, 2001):

$$W = 0.5 \cdot (\text{LF} + \text{LB} + \text{RF} + \text{RB}), \tag{10}$$

$$X = 0.5 \cdot [(\text{LF} - \text{LB}) + (\text{RF} - \text{RB})], \tag{11}$$

$$Y = 0.5 \cdot [(\text{LF} - \text{RB}) - (\text{RF} - \text{LB})], \tag{12}$$

and

$$Z = 0.5 \cdot [(\text{LF} - \text{LB}) + (\text{RB} - \text{RF})]. \tag{13}$$

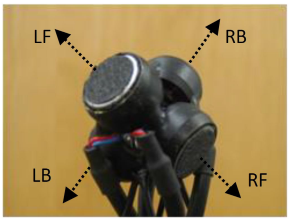


Fig. 6. Ambisonic microphone.

In the three-dimensional ambisonic expression, the Z component contains the information of the elevation of a sound source. The ambisonic microphone has been used for spatial impulse-response measurements (Kleczkowski *et al.*, 2015) and animal sound recordings (Ozga, 2017).

Similar to the two-dimensional ambisonic encoding procedure mentioned earlier, we can also synthesize a sound source positioned anywhere on the surface of a virtual sphere, as shown in Fig. 7. After determining the azimuthal angle θ and elevation φ, the W, X, Y, and Z components can be generated using the following equations:

$$W = S \cdot 0.7071, \tag{14}$$

$$X = S \cdot \cos\theta \cdot \cos\varphi, \tag{15}$$

$$Y = S \cdot \sin\theta \cdot \cos\varphi, \tag{16}$$

and

$$Z = S \cdot \sin\varphi. \tag{17}$$

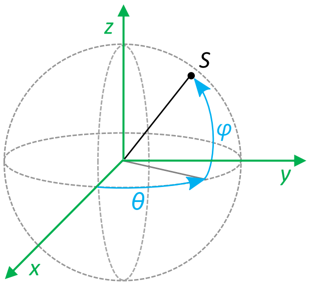


Fig. 7. Coordinates used in ambisonics. S is a sound source.

By using ambisonic rotation matrices, a sound field can be rotated about the x-, y-, and z-axis with very low computational cost. This feature is especially useful in a real-time binaural audio system with head tracking. As an example, we take the rotation of a listener's head about the z-axis. Consider a mono sound source S located at angle θ on the horizontal plane. Initially, the B-format signals X and Y are $S \cdot \cos\theta$ and $S \cdot \sin\theta$, respectively. After the listener rotates his head horizontally by angle δ_z, as the x-y plane changes, the transformed X and Y signals become X' and Y' in Eqs (18) and (19); these equations are equivalent to the matrix operation in (20). The 2×2 matrix in Eq. (20) is the first-order ambisonic rotation matrix about the z-axis:

$$X' = S \cdot \cos(\theta + \delta_z) = X \cdot \cos(\delta_z) - Y \cdot \sin(\delta_z), \tag{18}$$

$$Y' = S \cdot \sin(\theta + \delta_z) = Y \cdot \cos(\delta_z) + X \cdot \sin(\delta_z), \tag{19}$$

$$\begin{bmatrix} \cos(\delta_z) & -\sin(\delta_z) \\ \sin(\delta_z & \cos(\delta_z) \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X' \\ Y' \end{bmatrix}. \tag{20}$$

If the rotation is about the x- or y-axis, the rotation matrix can be obtained through a similar procedure. Because W is an omni-directional component, it does not change. The complete first-order ambisonic rotation matrices are shown in Eqs (21), (22), and (23):

$$\mathbf{K}_x(\delta_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\delta_x & -\sin\delta_x \\ 0 & \sin\delta_x & \cos\delta_x \end{bmatrix}, \tag{21}$$

$$\mathbf{K}_y(\delta_y) = \begin{bmatrix} \cos\delta_y & 0 & -\sin\delta_y \\ 0 & 1 & 0 \\ \sin\delta_y & 0 & \cos\delta_y \end{bmatrix}, \tag{22}$$

$$\mathbf{K}_z(\delta_z) = \begin{bmatrix} \cos\delta_z & -\sin\delta_z & 0 \\ \sin\delta_z & \cos\delta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{23}$$

An ambisonic rotation matrix $\mathbf{K}_r(\delta_r)$ can change a column vector of encoded channels $\mathbf{B}$ by rotating by angle δ_r about the r-axis through the following matrix operation:

$$\mathbf{B}' = \mathbf{K}_r(\delta_r) \times \mathbf{B}. \tag{24}$$

Multiplication of the matrices can produce a compound rotation about all three axes:

$$\mathbf{B}' = \mathbf{K}_x(\delta_x) \times \mathbf{K}_y(\delta_y) \times \mathbf{K}_z(\delta_z) \times \mathbf{B}. \tag{25}$$

Because the ambisonic rotation matrices can represent any angles in the three-dimensional space, interpolation algorithms are not required during the auditory space rotation.

Although in first-order ambisonic decoding, a cubic loudspeaker array is normally mounted for three-dimensional space, we use the example of a tetrahedral loudspeaker array for a concise and explicit illustration. The following example shows how the matrix operation works in the first-order ambisonic format with a tetrahedral loudspeaker array, and the higher-order decoder for a regular or irregular loudspeaker setting can be designed in a similar manner. If the four loudspeakers in the array can reproduce the desired sound field, the following matrix equation is true:

$$\begin{bmatrix} W \\ X' \\ Y' \\ Z' \end{bmatrix} = \begin{bmatrix} 0.707 & 0.707 & 0.707 & 0.707 \\ a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ \sin\psi_1 & \sin\psi_2 & \sin\psi_3 & \sin\psi_4 \end{bmatrix} \begin{bmatrix} o_1 \\ o_2 \\ o_3 \\ o_4 \end{bmatrix}, \tag{26}$$

where (ϑ_n, ψ_n), $n = 1, ..., 4$, defines the location of the n-th loudspeaker, and the elements in the middle matrix are the spherical harmonic functions for the loudspeaker locations, and $a_n = \cos\vartheta_n \cdot \cos\psi_n$, $b_n = \sin\vartheta_n \cdot \cos\psi_n$.

The matrix on the left-hand side represents the encoded ambisonic channels, and the matrix on the right-hand side contains the sounds emitted by the four loudspeakers. To obtain the loudspeaker feedings, Eq. (26) is rewritten as

$$\begin{bmatrix} o_1 \\ o_2 \\ o_3 \\ o_4 \end{bmatrix} = \begin{bmatrix} 0.707 & 0.707 & 0.707 & 0.707 \\ a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ \sin\psi_1 & \sin\psi_2 & \sin\psi_3 & \sin\psi_4 \end{bmatrix}^{-1} \begin{bmatrix} W \\ X' \\ Y' \\ Z' \end{bmatrix}. \tag{27}$$

The inverse of the spherical harmonic matrix is the ambisonic decoding matrix in the three-dimensional space.

After designing the loudspeaker feedings, we used a set of HRIRs to render each loudspeaker in a virtual space. Because the number of loudspeakers in the three-dimensional space is normally larger than that used in the two-dimensional space, the system should involve greater computational cost. We applied the concept proposed by McKeag and McGrath (1996) to reduce the computational cost. In the first-order standard design flow shown in Fig. 8, the encoded and rotated signals (W, X', Y', Z') are multiplied by the spherical harmonic coefficients and are then emitted by a virtual loudspeaker array in which each virtual loudspeaker can be thought of as a pair of HRIRs $(h_{1l}, h_{1r}, ..., h_{Ll}, h_{Lr})$. Finally, two adders collect the signals from the left and right HRIRs and send them to the listener's left and right ear, respectively. Because the system is linear, the spherical harmonic coefficients are combined with HRIRs as shown in Fig. 9. A significant advantage of this combination is that the number of finite impulse response (FIR) filters depends only on the ambisonic order and is independent of the number of virtual loudspeakers.

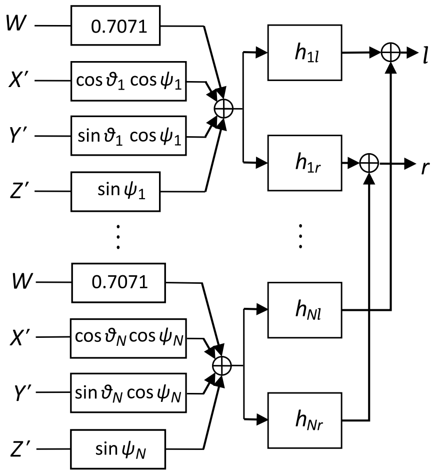


Fig. 8. Ambisonic decoder and virtual loudspeaker array for headphones. The number of loudspeakers is N. Angles inside spherical harmonic coefficients depend on the positions of loudspeakers. l and r are the left and right headphone feeds.

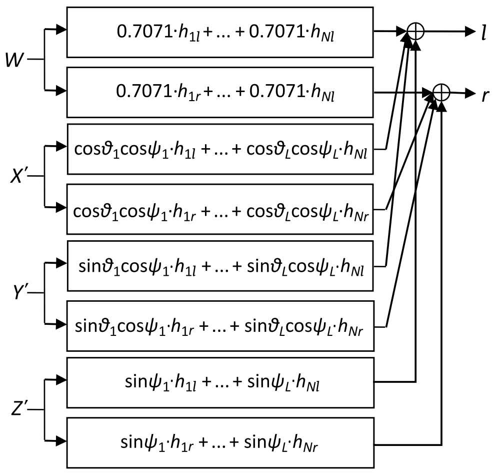

W
0.7071·h1l + ... + 0.7071·hNl
0.7071·h1r + ... + 0.7071·hNl
X'
cosϑ1cosψ1·h1l + ... + cosϑLcosψL·hNl
cosϑ1cosψ1·h1r + ... + cosϑLcosψL·hNr
Y'
sinϑ1cosψ1·h1l + ... + sinϑLcosψL·hNl
sinϑ1cosψ1·h1r + ... + sinϑLcosψL·hNr
Z'
sinψ1·h1l + ... + sinψL·hNl
sinψ1·h1r + ... + sinψL·hNr
l
r

Fig. 9. Optimized binaural 1st-order ambisonic decoder; l and r are the left and right headphone feeds.

When using a computer to synthesize the B-format components, the reverberations in the virtual acoustic space are excluded. To make the sound more realistic, a room model based on a mirror image is applied. The room model contains a direct path and several reflected paths. The sound takes a longer time to travel along the reflected path than the direct path. Furthermore, the reflected path has less energy, and its direction is different from the direct path. Therefore, the encoder equipped with the room model can be represented by the block diagram in Fig. 10. Similar to the optimization in the encoder, we can use FIR filters to simplify the structure. Figure 11 shows an optimized first-order system. Because the transfer functions from input S to each ambisonic channel (W, X, Y, and Z) are computed, the FIR filters (FIR_W, FIR_X, FIR_Y, and FIR_Z) in Fig. 11 contain the features of the delay, decay, and spherical harmonic functions.

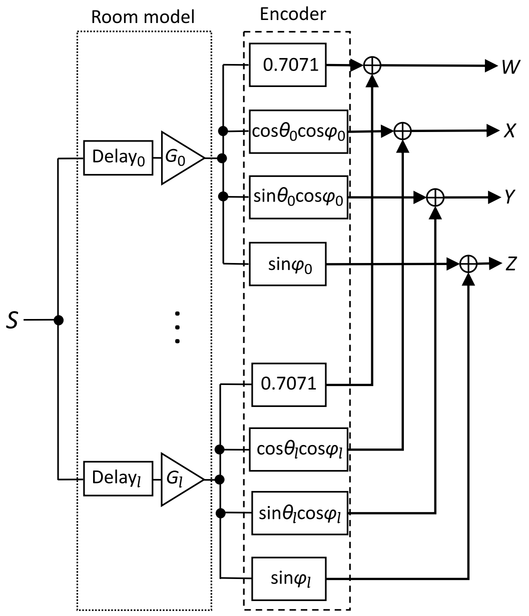


Fig. 10. Original room model and the 1-order ambisonic encoder. The number of mirror images is I. The angles inside the trigonometric functions depend on the positions of a real source and mirror sources.

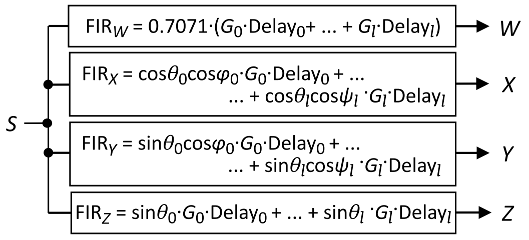

S
FIR_W = 0.7071·(G0·Delay0+ ... + GI·DelayI)
FIR_X = cosθ0cosφ0·G0·Delay0 + ... + cosθIcosψI ·GI·DelayI
FIR_Y = sinθ0cosφ0·G0·Delay0 + ... + sinθIcosψI ·GI·DelayI
FIR_Z = sinθ0·G0·Delay0 + ... + sinθI ·GI·DelayI
W
X
Y
Z

Fig. 11. Optimized first-order encoding filters.

3. Practical issues and proposed solution

Although ambisonics can reconstruct the sound field in the center point of the loudspeaker array, in a practical situation, the filtering effect occurs at the listener's ear position. Suppose that the radius of a listener's head is 0.1 m and the velocity of sound is 340 m/s. We consider a circular 500-loudspeaker array, with each loudspeaker playing an impulse signal, as shown in Fig. 12. Considering the time domain response and the frequency domain response 0.1 m away from the origins, as shown in Figs 13a and 13b, respectively, we can clearly determine that the spectrum produced by such a high-density loudspeaker array is no longer flat. Further, Figs 14a and 14b show the responses in a three-dimensional situation. There are 1250 loudspeakers in a sphere, at the same positions as those in a CIPIC setting. In the interaural–polar coordinate, elevations are sampled in 5.625° steps from −45° to 230.625°, and in the azimuthal direction at

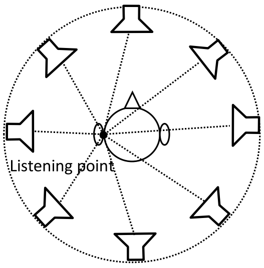


Fig. 12. Off-center listening position.

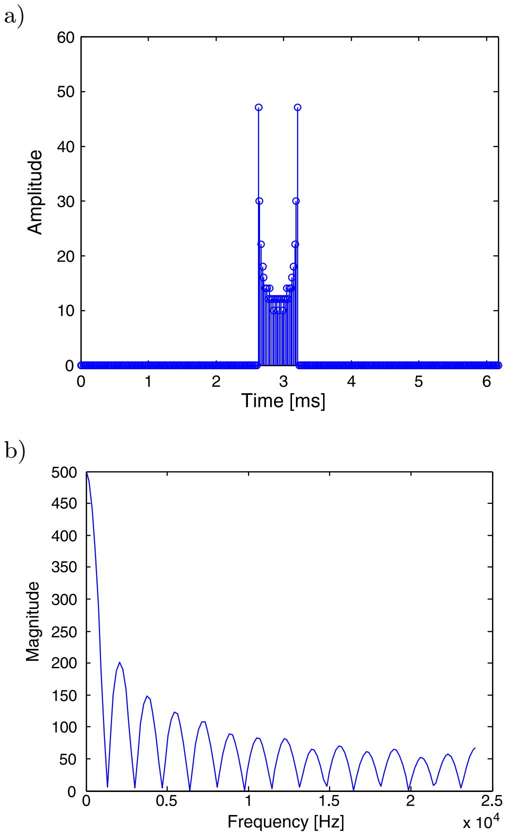


Fig. 13. The responses at the ear position at a 500-loudspeaker array in the 2-D space: a) time amplitude response, b) frequency magnitude response.

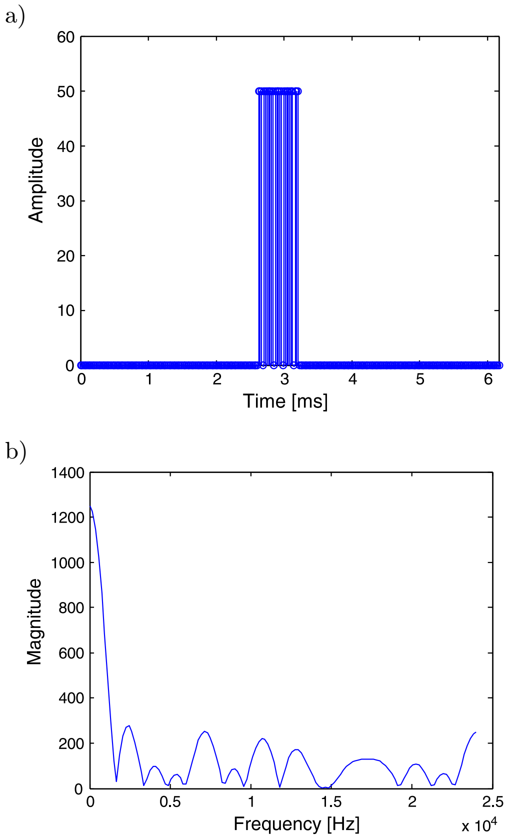


Fig. 14. The responses at the ear position at a 1250-loudspeaker array in the 3-D space: a) time amplitude response, b) frequency magnitude response.

±80°, ±65°, ±55°, and then from −45° to 45° in steps of 5°. In these two figures, we observe the comb- and low-pass filtering.

According to the above simulation, although ambisonics can theoretically construct the sound field in the center of the loudspeaker array, the listener's ear positions, as shown in Fig. 12, are not in the center. This is why, Blauert and Rabenstein (2012) emphasized the disturbance in the ambisonic sound field caused by a listener's head. Therefore, we further use HRIRs to simulate the practical situation and look into the ITD and IID cues.

The ITD estimation used in this study is based on the interaural cross-correlation function between a subject's ears (D'Orazio *et al.*, 2009; Sato, 2014) and the IID cues are calculated from the sound energy differences (Satongar *et al.*, 2013; Gaik, 1993). We virtually develop a square array and 24-loudspeaker array as shown in Fig. 15. The example HRTF dataset is IRC_1026_C_HRIR from the IRCAM database. We first place an impulse at 270°, in the direction of the listener's right ear. The absolute errors of the ambisonics-generated sound source are calculated using Eqs (28) and (29). The HRTF-generated source is the reference signal and the ambisonics-generated sound source is the signal under test. In ITD analysis, the impulse response is 700 Hz low-pass filtered because of the limitation of ITD cues (Yao, 2018). According to Eq. (28), the absolute ITD error of the square array is 0.2494 ms and that of the 24-loudspeaker array is 0.1587 ms. The larger error in the square array indicates that the sound source in the lateral region leads to poor localizations, which agrees with Collins (2013). In IID analysis, the impulse response is 700 Hz high-pass filtered. According to Eq. (29), the absolute IID error of the square array is 7.079 dB and that of the 24-loudspeaker array is 8.203 dB. The sound source in the lateral region should lead to poor localization; however, as we previously noted, comb-filtering in a dense loudspeaker array causes spectral distortion, degrading the IID accuracy:

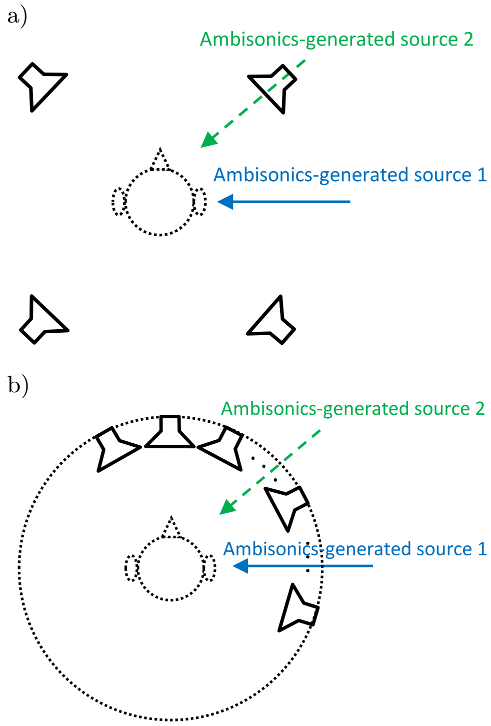

a)
Ambisonics-generated source 2
Ambisonics-generated source 1
b)
Ambisonics-generated source 2
Ambisonics-generated source 1

Fig. 15. a) Square array and (b) 24-loudspeaker array.

$$E_{\text{ITD}}(\theta) = \left|\text{ITD}_{\text{HRIR}}(\theta) - \text{ITD}_{\text{Ambisonics}}(\theta)\right|, \quad (28)$$

$$E_{\text{ILD}}(\theta) = \left|\text{ILD}_{\text{HRIR}}(\theta) - \text{ILD}_{\text{Ambisonics}}(\theta)\right|. \quad (29)$$

We then change the source position from 270° to 315°; thus, it is positioned in the direction of a loudspeaker in both arrays. The absolute ITD and IID errors are 0.0227 ms and 4.499 dB in the square array and 0.0454 ms and 6.603 dB in the 24-loudspeaker array, respectively. The result supports the conclusions of Collins (2013) and Yao (2018). That is, placing a sound source in the direction of a loudspeaker can produce accurate localization. We place the source at positions ranging from 0° to 345° in steps of 15° to obtain the full picture of localization cues in both loudspeaker arrays. The ambisonics-generated

ITD and IID cues are presented in Figs 16 and 17. In Fig. 16, the mean absolute ITD and IID errors in the square array are 0.1483 ms and 4.872 dB, respectively. In Fig. 17, the mean absolute ITD and IID errors in the 24-loudspeaker array are 0.1247 ms and 5.317 dB, respectively. By examining the ITD cues in the lateral regions in the square array, at approximately 0°, 90°, 180°, and 270°, we observe that the ambisonics-generated ITDs are very different from the HRIR-generated ITDs, as shown in Fig. 16a. Specifically, there are large ITD errors in the lateral regions in the square array. When we extend the system to the 24-loudspeaker array, the ITD errors decrease, as can be seen in Fig. 17a. However, a side effect of this process is poor IID cues in the 24-loudspeaker array, as can be seen from the comparison between Fig. 16b and Fig. 17b.

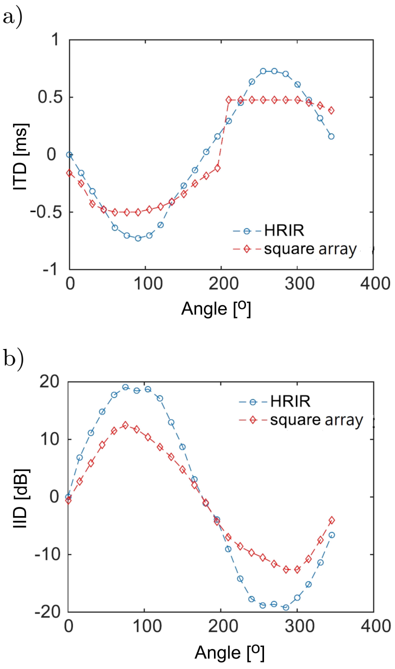


Fig. 16. a) ITD and b) IID in a square array constructed by IRC_1026_C_HRIR in the IRCAM database.

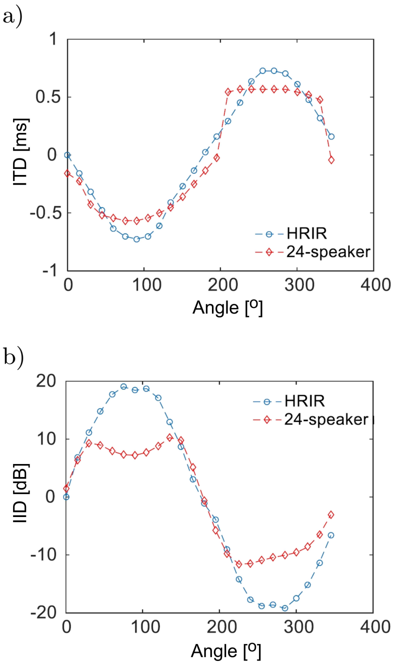


Fig. 17. a) ITD and b) IID in a 24-loudspeaker array constructed by IRC_1026_C_HRIR in the IRCAM database.

Because the CIPIC database contains more HRTF datasets than the IRCAM database, we use the CIPIC database to demonstrate low-pass filtering in a dense loudspeaker array. Two types of ambisonic loudspeaker array were constructed virtually: a 50-loudspeaker array and 4-loudspeaker array. We placed an impulse at 0° and examined the magnitude responses at the listener's left and right ears. When we compared the ambisonics-generated spectra with the HRTFs, lowpass filtering was observed in the dense ambisonic loudspeaker array, as shown in Figs 18a and 18b. Although

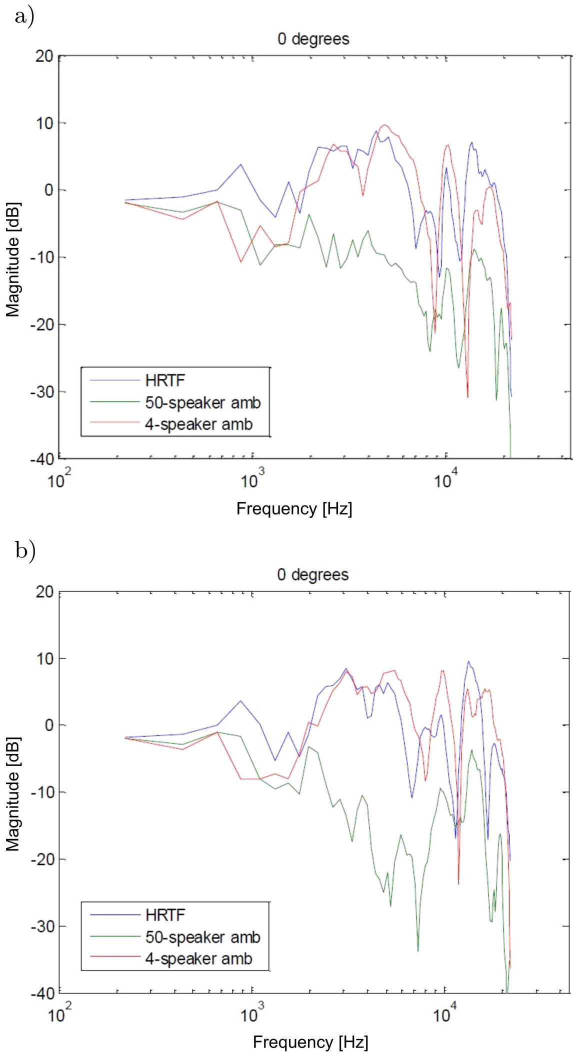


Fig. 18. Comparison between HRTF responses and ambisonic responses at the listener's left (a) and right (b) ears. The HRTF dataset was obtained from the CIPIC database (HRTF number: 156).

Blauert and Rabenstein (2012) reported poor localization in lateral directions between loudspeakers, we could hardly prevent the side effect of adding more loudspeakers than the minimum requirement; that is, a dense loudspeaker array makes the lateral region small, but it also results in a low-pass filtering frequency spectrum.

The number of loudspeakers in an ambisonic array influences the frequency spectrum. When using fewer loudspeakers, the localization cues in the lateral region can be poor. When using a dense loudspeaker array, the phase shift leads to spectral impairment, which can cause incorrect IID cues. As a result, Yao *et al.* (2015) proposed a split-band decoder to extract the nearly perfect reconstructed frequency components from two loudspeaker configurations. In Yao (2018), an intelligent equalizer was employed to automatically compensate the degraded high-frequency spectrum. However, both methods have disadvantages. In the former case, most frequency components are from a small loudspeaker array in the split-band decoder; therefore, the split-band decoder is good for high-pitch sound. In the latter case, most frequency components are from a large loudspeaker array in the equalization decoding; therefore, equalization decoding is good for low-pitch sound. Moreover, the two papers did not discuss the same loudspeaker array with different rotation angles.

We rotated the square loudspeaker array by 45°; the rotated loudspeaker array is hereafter called a cross array. Figure 4 shows a square array with four loudspeakers placed at 45°, 135°, 225°, and 315°. Figure 19 displays a cross array with four loudspeakers located at 0°, 90°, 180°, and 270°. As can be seen in Fig. 16, the localization performance in a square array was the poorest when the sound image was placed in the direction of the listener's left or right ear. We therefore moved the source from 0° to 345° to obtain the full picture of localization cues in the cross array, as shown in Fig. 20, and examined the localization cues in the direction of the listener's left or right ear. Figure 20 shows that the mean absolute ITD and IID errors in the cross array are 0.1002 ms and 7.512 dB, respectively. The cross array compensates for the poor localization when the sound source is near the listener's ear. We also performed IID and ITD analysis for the combination of a square array and cross array; that is, a regular octagonal array. As shown in Fig. 21, the mean absolute ITD and IID errors in the octagonal array are 0.1276 ms and 6.626 dB, respectively. A com-

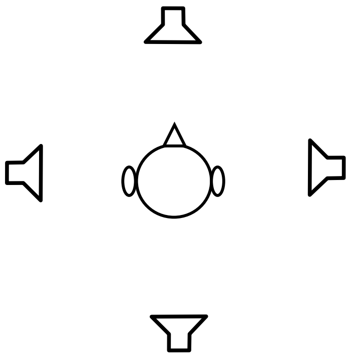


Fig. 19. Configuration of cross array.

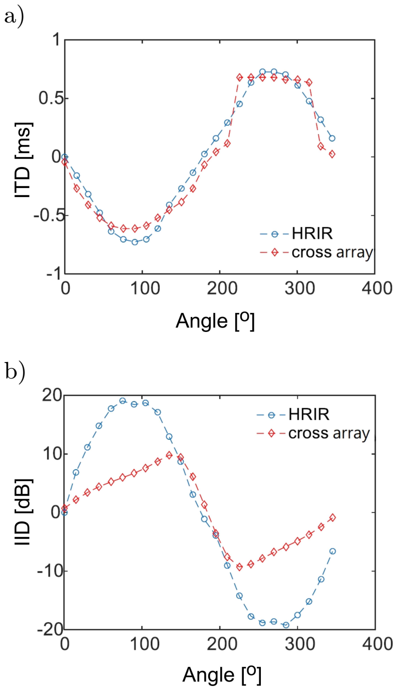


Fig. 20. a) ITD and b) IID in a cross array constructed by IRC_1026_C_HRIR in the IRCAM database.

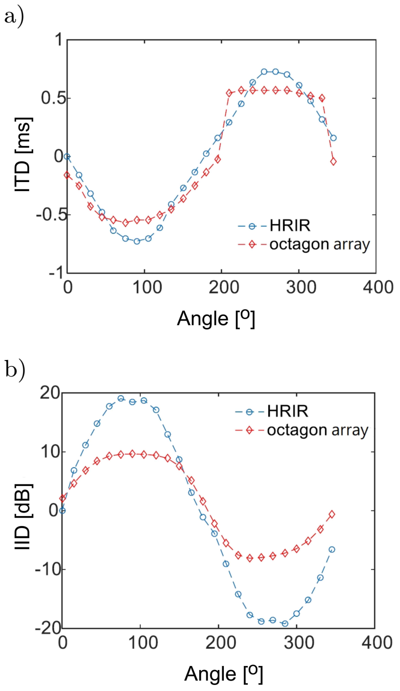


Fig. 21. a) ITD and b) IID in an octagon array constructed by IRC_1026_C_HRIR in the IRCAM database.

parison between the cross array and the square array shows that the same loudspeaker array with different rotations leads to very different results, which has not been discussed in the previous literature. Because the cross array obtains the best ITD cues and the square array obtains the best IID cues, we propose a dual-mode decoder.

Technically, by using a linear composition, the number of loudspeakers is independent of the computational cost. However, a dense loudspeaker array led to spectrum impairment (Yao *et al.*, 2015; Yao, 2018). We divide the frequency range into two parts. The frequencies below the localization transition frequency of 700 Hz are the low-frequency band, whereas those above 700 Hz are the high-frequency band. We use the cross loudspeaker array to produce the ITD cues, because the localization of the low-frequency components relies on these cues. The high-frequency components are decoded by the square array. The dual-mode decoder is shown in Fig. 22. The division is approached by driver filters in software audio crossovers. Yao (2014) designed driver filters for two- or three-way loudspeaker systems and revealed their potential for compensating nonlinear and time-variant distortions in sound. The crossover-filtered signals are convolved with the corresponding FIR encoders and decoders. Finally, the decoded signals are combined and sent to the left and right channels.

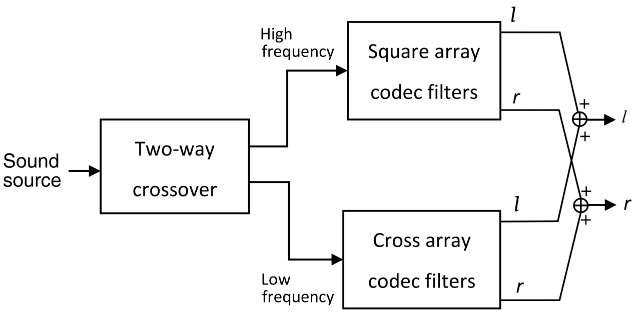
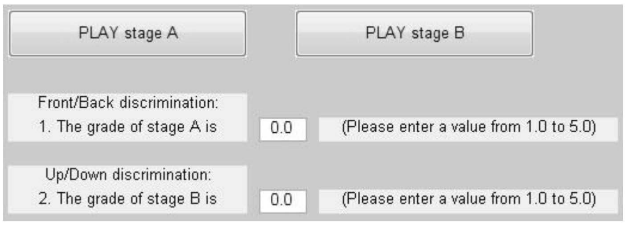


Fig. 22. The proposed dual-mode decoder.

We use HRTF datasets to render loudspeakers over headphones for VR applications. However, in a practical situation, the HRTFs differ from person to person, so we conducted customization. In Yao *et al.* (2017), 30 subjects were recruited to build an artificial intelligence system for HRTF selection. When the anthropometric parameters are known, the algorithm can select the most suitable HRTF dataset for an individual. However, the IRCAM database does not release anthropometric parameters and the CIPIC database does not provide HRIR datasets at 90° and 270° to allow us to construct the cross array. The method described in Yao (2017) was applied. We randomly selected ten HRIR datasets from the IRCAM database. In the first stage of the calibration, listeners listened to a mono sound convolved with a pair of HRIRs coming from 0° and 180°. The listeners were asked to evaluate the level of front–back confusion. In the second stage of the calibration, two separate sound sources were placed in the median plane at different elevations. The mono sound was convolved with a pair of HRIRs at the high-level position and then at the low-level position. The listeners were asked to report how well they could discriminate the sources at the high and low elevations.

The user interface for HRTF calibration is shown in Fig. 23. Each HRTF dataset has two listening scores: one for front–back discrimination and the other for up–down discrimination. The scores are recorded according to a continuous five-grade scale, as shown in Fig. 24. The average of two scores is calculated and the HRTF dataset with the highest average score is the listener's default dataset. In addition to HRTF fitness, Wersényi (2009) indicated that slight head movements are important to avoid in-the-head localization. He designed an experiment by randomly moving the virtual sound source to emulate head movements, so the subjects were not equipped with sensors. In our subjective listening test, a simple gyroscope was applied to detect head movements for sound externalization. The virtual acoustic space could be adjusted for vigorous movement.


PLAY stage A PLAY stage B

Front/Back discrimination:
1. The grade of stage A is 0.0 (Please enter a value from 1.0 to 5.0)

Up/Down discrimination:
2. The grade of stage B is 0.0 (Please enter a value from 1.0 to 5.0)

Fig. 23. User interface for HRTF selection.

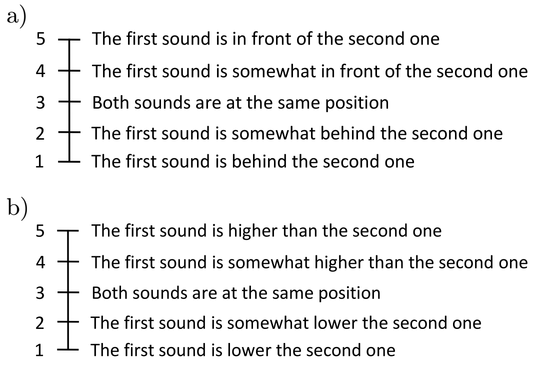

a)
5 The first sound is in front of the second one
4 The first sound is somewhat in front of the second one
3 Both sounds are at the same position
2 The first sound is somewhat behind the second one
1 The first sound is behind the second one
b)
5 The first sound is higher than the second one
4 The first sound is somewhat higher than the second one
3 Both sounds are at the same position
2 The first sound is somewhat lower the second one
1 The first sound is lower the second one

Fig. 24. Five-grade scale for (a) front-back and (b) up-down localization rating.

4. Experimental results

To assess the audio quality of the proposed ambisonic decoder, objective measurements and subjective evaluations were conducted. The loudspeakers in the square array were placed at 45°, –45°, 135°, and

225°. The loudspeakers in the cross array were placed at 0°, 90°, −90°, and 180°. The octagonal array contained all the positions used in the square and the cross arrays.

The objective listening tests were conducted using 51 HRIR datasets from the IRCAM database (IRCAM, 2002). The subjective listening tests were performed by six male and four female participants. In the objective listening tests, five ambisonic decoders were tested: the square array decoder, the cross array decoder, the octagonal array decoder, the in-phase decoder (Monro, 2000), and the proposed dual-mode decoder. In the subjective listening test, to avoid listener fatigue, only three ambisonic decoders were used. The selection included the decoder with the most accurate ITD cues, the decoder with the most accurate IID cues, and the proposed dual-mode decoder.

In both tests, we placed a source horizontally at different angles around the listener. In the objective test, the sampling angles were the same as those of the HRIRs in the IRCAM database, from 0° to 345° in steps of 15°. Because we used ITD and IID analysis, the test signal was an impulse. In the subjective test, the sampling angles were 0°, ±45°, and ±90°. Because the subjective test was designed based on human perception, the test signals were audio pieces that contained double bass for low-frequency perception and trumpet for high-frequency perception.

4.1. Objective measurement

In the previous section, we described the ITD and IID analysis using an example HRTF dataset from the database. In this section, all HRTF datasets in the IRCAM database were involved, and the mean absolute errors were calculated. We included the in-phase decoder (Monro, 2000) that was proposed to overcome the limitations of off-center listening positions by reducing the directional components.

According to Eq. (28), we calculated the mean absolute ITD errors by averaging all $\mathrm{E}_{\mathrm{ITD}}(\theta)$ values, and the corresponding 95% confidence interval is shown in Fig. 25. The mean absolute ITD errors of the square decoder, the cross decoder, the octagonal decoder, the in-phase decoder, and the proposed decoder were 0.1329 ms, 0.0891 ms, 0.103 ms, 0.1877 ms, and 0.0644 ms, respectively. Figure 25 indicates that the ITDs of the in-phase decoder were the least accurate, and the mean values suggest that the proposed dual-mode decoder exhibits the best ITD performance. Because the proposed dual-mode decoder used the cross array to decode low-frequency sound, its performance was expected to be similar to that of the cross array. However, the ITD cues show a significant improvement between the cross array decoder and the proposed decoder. We therefore believe that the low-frequency driver filter in the two-way crossover proposed by Yao (2014) can enhance low-frequency sound.

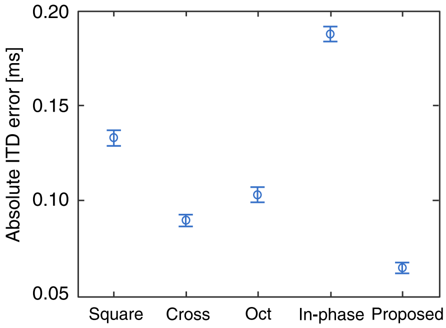


Fig. 25. Absolute ITD errors of five different virtual loudspeaker arrays built by IRCAM database.

The IID cues from the frequency band above 700 Hz were also estimated. By averaging the absolute errors in Eq. (29) from all angles, the mean absolute IID errors of the square array, the cross array decoder, the octagonal array decoder, the in-phase decoder, and the proposed dual-mode decoder were calculated as 6.067 dB, 6.717 dB, 6.589 dB, 5.993 dB, and 5.990 dB, correspondingly. The mean values and 95% confidence intervals were shown in Fig. 26. The cross and octagonal arrays exhibited poor IID accuracies. The in-phase decoder somewhat enhanced the IID accuracy in the square array decoder. It is interesting that the square array together with the two-way crossover can provide competitive IID performance. That is, the absolute IID error in the proposed dual-mode decoder is similar to that in the in-phase decoder.

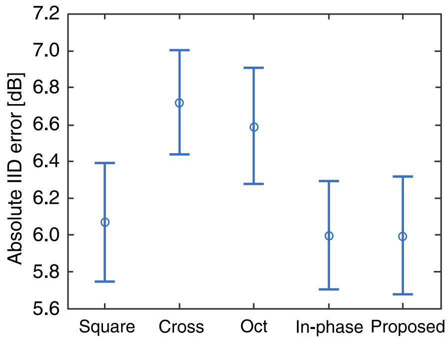
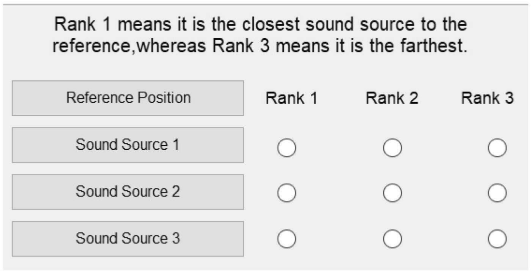


Fig. 26. Absolute IID errors of five different virtual loudspeaker arrays built by IRCAM database.

4.2. Subjective listening test

In the subjective listening test, we first conducted HRTF calibration as described earlier; consequently, the participants could determine the appropriate dataset. To avoid listener fatigue, three loudspeaker arrays were used in the objective listening test: a cross array, an in-phase array, and the proposed dual-mode decoder. This is because a cross array and an in-phase array showed the second-best ITD and IID cues in the objective measurement, respectively. The ambisonic decoders generated the sound sources at −90°, −45°, 0°, 45°, and 90° and were compared with the reference sound sources. The references were the sounds convolved with the corresponding HRTF dataset. The sound sources contained high-frequency trumpet music and low-frequency double-bass music. Taking the case of the trumpet at 0° as an example, we used the three ambisonic decoders to generate the trumpet sound at 0°. The trumpet sound was also convolved with the HRIR dataset at 0°, which was used as reference. As shown in the GUI in Fig. 27, the listeners were asked to evaluate the distance between the ambisonic-generated sources and the reference. Three rankings were used: rank 1 represents the closest distance between the sound source and the reference and rank 3 corresponds to the farthest. The decoder generating the closest sound source obtains 2 points, whereas that producing the farthest one scores 0 points. Although the participants were equipped with a head tracker, head movement was discouraged because the relative angle of the sound source would change when the participants rotated their heads. However, to operate in the practical sound field, the head tracker adjusted the virtual acoustic space when the participants unconsciously moved their heads.



Rank 1 means it is the closest sound source to the reference,whereas Rank 3 means it is the farthest.

Reference Position	Rank 1	Rank 2	Rank 3
Sound Source 1	○	○	○
Sound Source 2	○	○	○
Sound Source 3	○	○	○

Fig. 27. GUI for comparison between decoders.

The results for the tests using double bass as the sound source are shown in Fig. 28, which shows that the proposed dual-mode decoder constantly achieved good listening scores. The proposed dual-mode decoder obtained the best mean scores when the sound sources were at side positions. The cross array obtained the best mean score for the frontal sound source at 0°, but the mean score did not obviously outperform that of the proposed decoder. In the lateral regions like −45° and 45°, the localization of the cross array was as poor as expected. Generally, when using a low-frequency sound source, the proposed decoder tended to achieve the most accurate localization cues and the in-phase decoder obtained the worst, which matches the ITD cues shown in Fig. 25. The results of the tests using trumpet as the sound source are shown in Fig. 29. The cross array also achieved the best listening scores when the sound source was placed directly in front of the listener. However, the cross array was more likely to exhibit the worst performance in the other directions, which matches the objective result shown in Fig. 26. Overall, the proposed decoder achieved the most robust listening score. Because of the symmetry, the listening scores at −90° are similar to those at 90°, and the listening scores at −45° are similar to those at 45°. The cross array could not obtain good localization, even if the sound source was placed in the direction of the left or right loudspeaker. A possible reason is that sound originating from one direction will be reproduced by many loudspeakers, even if one loudspeaker corresponds exactly to the desired direction.

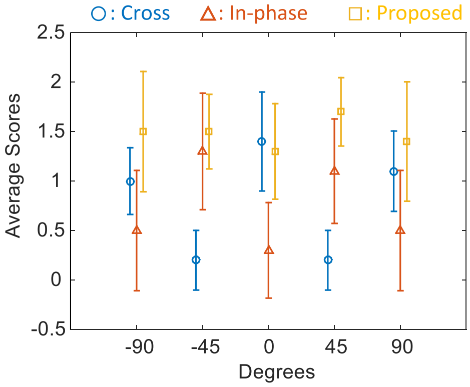


Fig. 28. Subjective listening scores by using a double bass as a sound source.

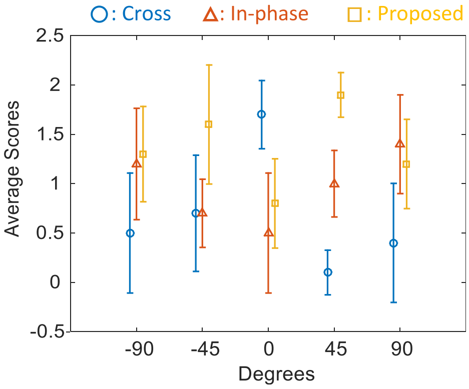


Fig. 29. Subjective listening scores by using a trumpet as a sound source.

5. Conclusion

Multi-channel surround sound can be applied not only to home theaters but also to VR. Notably, the shape of loudspeaker array affects the audio quality and localization, whereas the relationship between audio localization and the orientation of a loudspeaker array is rarely studied. We use binaural recording to show that different orientations can produce different localizations in specific sound directions. We propose a dual-mode decoder combining the merits of two orientations. We utilize two-way crossover to separate the treble and bass, and evaluate the ITD and IID accuracies of three binaural decoders using both subjective and objective listening tests. The proposed decoder shows accurate high- and low-frequency localizations and the crossover filter can compensate the audio distortion.

The proposed method contributes to binaural sound localization for VR applications. We use first-order ambisonics as an example, because the main producers, such as YouTube and Oculus, currently support only the first-order format. Higher-order ambisonics should be further investigated. Although it is difficult to design a higher-order ambisonic microphone, ambisonic signals can be artificially synthesized. When the order is higher, more ambisonic channels are required; furthermore, the optimization technique described in this paper can efficiently reduce computational complexity.

Acknowledgment

This work was supported in part by the Ministry of Science and Technology in Taiwan under Grant MOST 109-2221-E-305-011-MY2.

References

1. Algazi V.R., Duda R.O., Thompson D.M. (2004), Motion-tracked binaural sound, *Journal of the Audio Engineering Society*, **52**(11): 1142–1156, http://www.aes.org/e-lib/browse.cfm?elib=12644.

2. Blauert J., Rabenstein R. (2012), Providing surround sound with loudspeakers: a synopsis of current methods, *Archives of Acoustics*, **37**(1): 55–62, doi: 10.2478/v10168-012-0002-y.

3. Clark H.A.M., Dutton G.F., Vanderlyn P.B. (1958), The 'stereosonic' recording and reproducing system: a two-channel systems for domestic tape records, *Journal of the Audio Engineering Society*, **6**(2): 102–117, doi: 10.1049/pi-b-1.1957.0180.

4. Collins T. (2013), Binaural ambisonic decoding with enhanced lateral localization, *Proceedings of Audio Engineering Society 134th Convention*, http://www.aes.org/e-lib/browse.cfm?elib=16779.

5. D'Orazio D., Guidorzi P., Garai M. (2009), A Matlab Toolbox for the analysis of Ando's factors, *Proceedings of Audio Engineering Society 126th Convention*, http://www.aes.org/e-lib/browse.cfm?elib=14994.

6. Gaik W. (1993), Combined evaluation of interaural time and intensity differences: psychoacoustic results and computer modelling, *The Journal of the Acoustical Society of America*, **94**(1): 98–110, doi: 10.1121/1.406947.

7. Gardenfors D. (2003), Designing sound-based computer games, *Digital Creativity*, **14**(2): 111–114, doi: 10.1076/digc.14.2.111.27863.

8. Gaudy T., Natkin S., Archambault D. (2009), Pyvox 2: An audio game accessible to visually impaired people playable without visual nor verbal instructions, *Transactions on Edutainment II*, **5660**: 176–186, doi: 10.1007/978-3-642-03270-7_12.

9. Gerzon M.A. (1975), The design of precisely coincident microphone arrays for stereo and surround sound, *Proceedings of Audio Engineering Society 50th Convention*, London, http://www.aes.org/e-lib/browse.cfm?elib=2466.

10. IRCAM (2002), *Listen HRTF database*, http://recherche.ircam.fr/equipes/salles/listen/.

11. Kleczkowski P., Krol A., Malecki P. (2015), Reproduction of phantom sources improves with separation of direct and reflected sounds, *Archives of Acoustics*, **40**(4): 575–584, doi: 10.1515/aoa-2015-0057.

12. Matsumura T., Iwanaga N., Kobayashi W., Onoye T., Shirakawa I. (2005), Embedded 3D sound movement system based on feature extraction of head-related transfer function, *IEEE Transactions* on *Consumer Electronics*, **51**(1): 262–267, doi: 10.1109/TCE.2005.1405730.

13. McKeag A., McGrath D. (1996), Sound field format to binaural decoder with head-tracking, *Proceedings of 6th Australian Regional Audio Engineering Society Convention*, http://www.aes.org/e-lib/browse.cfm?elib=7477.

14. Monro G. (2000), In-phase corrections for ambisonics, *Proceedings of International Computer Music Conference*, http://hdl.handle.net/2027/spo.bbp2372.2000.194.

15. Ozga A. (2017), Scientific ideas included in the concepts of bioacoustics, acoustic ecology, ecoacoustics, soundscape ecology and vibroacoustics, *Archives of Acoustics*, **42**(3): 415–421, doi: 10.1515/aoa-2017-0043.

16. Rumsey F. (2001), *Spatial Audio*, Jordan Hill, Oxford: Focal Press.

17. Sato S. (2014), MATLAB program for calculating the parameters of the autocorrelation and interaural cross-correlation functions based on Ando's auditory-brain model, *Proceedings of Audio Engineering Society 137th Convention*, http://www.aes.org/e-lib/browse.cfm?elib=17504.

18. Satongar D., Dunn C., Lam Y., Li F. (2013), Localisation performance of higher-order Ambisonics for off-centre listening, *White Paper*, WHP254, https://www.bbc.co.uk/rd/publications/whitepaper254.

19. Scaini D., Arteaga D. (2014), Decoding of higher order ambisonics to irregular periphonic loudspeaker arrays, *Proceedings of Audio Engineering Society 55th*

Convention, http://www.aes.org/e-lib/browse.cfm?elib=17364.

20. Wersényi G. (2009), Effect of emulated head-tracking for reducing localization errors in virtual audio simulation, *IEEE Transactions on Audio, Speech, and Language Processing*, **17**(2): 247–252, doi: 10.1109/TASL.2008.2006720.

21. Yao S.-N. (2014), Driver filter design for software-implemented loudspeaker crossovers, *Archives of Acoustics*, **39**(4): 591–597, doi: 10.2478/aoa-2014-0063.

22. Yao S.-N. (2017), Headphone-based immersive audio for virtual reality headsets, *IEEE Transactions on Consumer Electronics*, **63**(3): 300–308, doi: 10.1109/TCE.2017.014951.

23. Yao S.-N. (2018), Equalization in ambisonics, *Applied Acoustics*, **139**: 129–139, doi: 10.1016/j.apacoust.2018.04.027.

24. Yao S.-N., Chen L.J. (2013), HRTF adjustments with audio quality assessments, *Archives of Acoustics*, **38**(1): 55–62, doi: 10.2478/aoa-2013-0007.

25. Yao S.-N., Collins T., Jančovič P. (2015), Timbral and spatial fidelity improvement in ambisonics, *Applied Acoustics*, **93**: 1–8, doi: 10.1016/j.apacoust.2015.01.005.

26. Yao S.-N., Collins T., Liang C. (2017), Head-related transfer function selection using neural networks, *Archives of Acoustics*, **42**(3): 365–373, doi: 10.1515/aoa-2017-0038.