IEICE TRANSACTIONS

on Information and Systems

VOL. E99-D NO. 1
JANUARY 2016

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

INVITED PAPER *Special Section on Enriched Multimedia—Creation of a New Society through Value-added Multimedia Content—*

Enhancing Stereo Signals with High-Order Ambisonics Spatial Information

Jorge TREVINO[†a)], ***Nonmember***, **Shuichi SAKAMOTO**[†b)], ***Member***, **Junfeng LI**[††c)], ***Nonmember***, ***and*** **Yôiti SUZUKI**[†d)], ***Fellow***

SUMMARY There is a strong push towards the ultra-realistic presentation of multimedia contents made possible by the latest advances in computational and signal processing technologies. Three-dimensional sound presentation is necessary to convey a natural and rich multimedia experience. Promising ways to achieve this include the sound field reproduction technique known as high-order Ambisonics (HOA). While these advanced methods are now within the capabilities of consumer-level processing systems, their adoption is hindered by the lack of contents. Production and coding of the audio components in multimedia focus on traditional formats such as stereophonic sound. Mainstream audio codecs and media such as CDs or DVDs do not support advanced, rich contents such as HOA encodings. To ameliorate this problem and speed up the adoption of spatial sound technologies, this paper proposes a novel way to downmix HOA contents into a stereo signal. The resulting data can be distributed using conventional methods such as audio CDs or as the audio component of an internet video stream. The results can be listened to using legacy stereo reproduction systems. However, they include spatial information encoded as the inter-channel level and phase differences. The proposed method consists of a downmixing filterbank which independently modulate inter-channel differences at each frequency bin. The proposal is evaluated using simple test signals and found to outperform conventional methods such as matrix-encoded surround and the Ambisonics UHJ format in terms of spatial resolution. The proposal can be coupled with a previously presented method to recover HOA signals from stereo recordings. The resulting system allows for the preservation of full-surround spatial information in ultra-realistic contents when they are transferred using a stereo stream. Simulation results show that a compatible decoder can accurately recover up to five HOA channels from a stereo signal (2nd order HOA data in the horizontal plane).

key words: *spatial sound, high-order Ambisonics, spatialization, surround, sound signal encoding*

1. Introduction

Sound plays a critical role in multimedia communications. Realistic sound is necessary to convey the rich perceptual and affective information humans need to understand a perceptual scene. For this reason, the present paper focuses on the enrichment of multimedia contents, understood as their quality enhancement through signal processing, from the point of view of the auditory modality. Previous studies have found that the addition of even small amounts of linguistic information in the form of text can significantly enhance the perception of an auditory scene [1]. The present research, on the other hand, focuses on media enhancements that are fully contained within the auditory modality.

An important function of the auditory system is to provide the listener with spatial information, such as the approximate positions of sound sources around them; this is known as spatial hearing [2]. The realistic presentation of multimedia contents must take this into account and convey spatial information through sound. The importance of this is highlighted by the fact that the auditory modality provides the listener with information covering all directions around them, while vision covers only the front half-space. Furthermore, the auditory modality plays a critical role in determining the affective aspects of multimedia perception, such as the sense-of-presence [3].

The processing and presentation of spatial sound information is now possible [4], [5], thanks to recent advances in computing and telecommunication technologies. There are three mainstream approaches to the problem of spatial sound reproduction: binaural, multi-channel surround and sound field reproduction.

Binaural sound reproduction attempts to control the sound pressure at the listener's ears. There is a wide variety of binaural systems capable of presenting recorded or synthesized sounds using either headphones or loudspeakers [6]–[9]. Binaural techniques can accurately convey spatial sound information; however, they require individual measurements of the head-related transfer function (HRTF) [2]. Binaural systems must be coupled with sophisticated tracking and processing systems if the control points (the position of the listener's ears) are allowed to move. This condition is, however, mandatory for the accurate presentation of spatial sound [10]. Binaural recordings are commercially available, but they represent a niche market.

In a multi-channel surround sound system, a number of loudspeakers are arranged into a predefined configuration and used to present sounds from their respective directions [11], [12]. The reproduction stage of a multi-channel surround system, in general, does not require any special processing of the audio signals. This has made it a popular choice for mainstream spatial sound reproduction. However, commercial systems such as 5.1-channel surround have very limited spatial resolution when compared to other

Manuscript received July 31, 2015.
Manuscript revised September 19, 2015.
Manuscript publicized October 21, 2015.

[†]The authors are with the Research Institute of Electrical Communication and the Graduate School of Information Sciences, Tohoku University, Sendai-shi, 980–8577 Japan.
[††]The author is with the Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190 China.
a) E-mail: jorge@ais.riec.tohoku.ac.jp
b) E-mail: saka@ais.riec.tohoku.ac.jp
c) E-mail: lijunfeng@hccl.ioa.ac.cn
d) E-mail: yoh@riec.tohoku.ac.jp

DOI: 10.1587/transinf.2015MUI0001

technologies. Furthermore, the lack of processing in the reproduction stage requires the audio contents to be mixed in a studio for a specific loudspeaker distribution. Any changes or upgrades to the reproduction system also require the contents to be updated. Nevertheless, multi-channel surround dominates the consumer market for spatial audio contents.

Spatial sound reproduction systems in the third category, sound field reproduction, represent a relatively new method made possible by faster signal processing and multi-channel technologies. They work by re-creating the sound pressure over an extended region surrounding the listeners [13]–[15]; this has several advantages over the other methods. Their focus on an extended region rather than two control points eliminates the required adjustments for each listener and their positions needed in binaural reproduction. The sound field reproduction approach reaches higher spatial resolutions than multi-channel surround systems by using the available loudspeakers more effectively. Until recent years, sound field reproduction was limited to research facilities and technical demonstrations [16], [17]. These technologies are now within the reach of modern consumer-level devices; however, there is only a handful of sound field recordings available to mainstream users. The lack of contents is due to the relative novelty of the method and the absence of a standard way to encode sound field information for distribution using conventional media.

The present paper seeks to accelerate the adoption of rich multimedia technologies by filling the gap between conventional audio systems and new technologies that can convey enhanced spatial sound characteristics. In particular, we propose a new method to enhance stereo signals with spatial sound information. Our proposal relies on a technology known as high-order Ambisonics (HOA) to encode sound field information into a multi-channel stream [15]. This is then downmixed into a stereo signal by modulating the inter-channel level and phase differences. The results are a stereo mix that can be reproduced by legacy systems. In addition, the original HOA data can be recovered using a previously proposed spatialization algorithm for stereo signals [18], [19].

The proposal focuses on two established technologies: HOA and stereo sound. The main reason behind the first choice is the system-agnostic property of the HOA format. Sound fields encoded using HOA can be reproduced by virtually any spatial audio system by adding a decoding stage [20]. It is also possible to reproduce them over headphones using binaural techniques [9]. The choice of a stereo signal as the output of the proposed algorithm is due to its widespread use, making it fully compatible with current technologies for broadcasting (radio, TV), distribution on physical media (audio CDs, DVDs) and internet streaming (MP3, FLAC, AAC).

An alternative method to represent HOA data using stereo and multi-channel signals is known as the Ambisonics UHJ format [21], [22]. This approach is similar to the matrix-encoding of surround sound used to downmix multi-channel data into a stereo signal and upmix it at the reproduction stage [23]–[26]. The main difference between these methods and the proposal lies in the way inter-channel differences are modulated. Conventional methods use a matrix of gains and delays that are applied equally to all frequency components. The proposal modulates the inter-channel differences using a filterbank stabilized by a non-linear spatial warping. Our results show that the proposal can achieve better spatial resolution than the Ambisonics UHJ format, which is limited to first-order horizontal Ambisonics (the lowest spatial resolution above monaural sound) when applied to generate stereo signals.

Section 2 reviews existing technologies to represent spatial sound using stereo signals. Section 3 summarizes a previously presented method to synthesize HOA data from an extended stereo mix. Section 4 introduces a new algorithm to generate stereo signals from HOA data. The methods of Sect. 3 and Sect. 4 are used together to evaluate the proposed system in Sect. 5. Finally, Sect. 6 summarizes the results and presents our conclusions.

2. Stereo Representation of Spatial Sound

Stereophonic sound is a well-established but limited method to convey spatial sound information. It uses two independent audio signals, a left and a right channel. In comparison with recent technologies, stereophonic systems have poor spatial resolution and cannot present sounds from all directions. Nevertheless, its long history and widespread adoption means that most sound systems can handle stereo signals.

The two-channel signals used in stereo systems can transport spatial information for more sophisticated reproduction methods. Binaural systems use stereo signals where the left and right channels carry the sound pressure data for the left and right ears respectively. In this Section, we review some of the existing methods that use stereo signals to transport the data for multi-channel surround and sound field reproduction systems.

2.1 Stereo Panning

Conventional stereophonic systems consist of two loudspeakers placed in front of the listener at azimuth angles of $-30°$ and $30°$. A common signal presented on both loudspeakers at different levels results in a sound image located somewhere between the loudspeakers. The level differences required to present sound from different directions are characterized by a *panning law*. The design of panning laws has been extensively studied. It is common to use sinusoid or tangent functions to define optimal laws [27], [28].

Conventional stereo signals consider panning laws spanning only the $-30°$ to $30°$ interval, which is the region covered by the loudspeakers. Presenting sound images outside of this interval with only two front loudspeakers requires more sophisticated techniques [8].

Some modern techniques, however, consider the panning law at directions outside of the loudspeaker coverage.

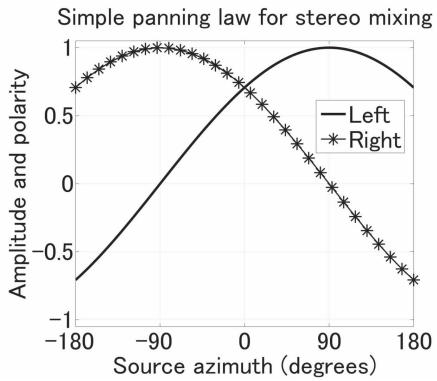

Fig. 1 Extended stereo panning law. Conventional stereophonic systems use similar laws restricted to angles between −30° and 30° from the frontal direction.

In these systems, the stereo signal is used to encode spatial information for reproduction using a more sophisticated system, such as multi-channel surround [23]. Figure 1 shows an extension of the conventional sinusoid panning law [23] covering all directions in the horizontal plane. An important observation is that, systems relying on the full panning law of Fig. 1, will give opposite polarities to the left and right channels when the sound image is located outside of the front half-space. The panning law encodes the left-right position of the desired sound image as an inter-channel amplitude difference, while its front-back position is encoded as an inter-channel phase difference.

2.2 Matrix-Encoded Surround

Panning laws similar to that of Fig. 1 can be used to encode the data for multi-channel surround systems as stereo signals. Each loudspeaker position in a multi-channel surround system can be associated with a pair of inter-channel amplitude and phase differences by a panning law. Since the loudspeaker positions are fixed, the stereo downmix of multi-channel surround sound can be conveniently denoted as a matrix operation [24]:

$$\begin{bmatrix} S_{\mathrm{L}}(\omega) \\ S_{\mathrm{R}}(\omega) \end{bmatrix} = \mathrm{M} \begin{bmatrix} S_1(\omega) \\ S_2(\omega) \\ \vdots \\ S_N(\omega) \end{bmatrix}. \tag{1}$$

In Eq. (1), the stereo signals $S_{\mathrm{L}}(\omega)$ and $S_{\mathrm{R}}(\omega)$ are linear combinations of N multi-channel surround signals. The gains and phase adjustments are summarized in an encoding matrix M of 2-by-N complex elements. A simple example used to encode four-channel surround is [25]:

$$\mathrm{M} = \begin{bmatrix} 0.92 & 0.38 & 0.46\pi i & 0.19\pi i \\ 0.38 & 0.92 & -0.19\pi i & -0.46\pi i \end{bmatrix}. \tag{2}$$

Inverting Eq. (1) makes it possible to approximate the original multi-channel data from its stereo downmix:

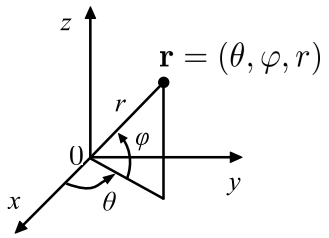

Fig. 2 The spherical coordinate system used in this paper. Angles θ and φ are the azimuth and elevation angles, respectively. The radial coordinate is denoted by r.

$$\begin{bmatrix} S_1(\omega) \\ S_2(\omega) \\ \vdots \\ S_N(\omega) \end{bmatrix} \approx \mathrm{M}^{+} \begin{bmatrix} S_{\mathrm{L}}(\omega) \\ S_{\mathrm{R}}(\omega) \end{bmatrix}. \tag{3}$$

In this equation, the decoding matrix M^{+} is the pseudo-inverse of M. The decoding equation associated with the example in Eq. (2) is:

$$\mathrm{M}^{+} = \begin{bmatrix} 0.33 & 0.21 \\ 0.21 & 0.33 \\ -0.4i & 0.05i \\ -0.05i & 0.4i \end{bmatrix}. \tag{4}$$

It is important to notice that these matrices are constant for all frequencies ω, since the panning law is not frequency dependent.

2.3 Sound Field Reproduction and the Ambisonics UHJ Format

High-order Ambisonics (HOA), a sound field reproduction technology, is a relatively new but promising method to present spatial sound where the sound pressure over an extended region surrounding the listener is controlled by a loudspeaker array. Sound field reproduction methods are both listener-independent and system-agnostic. This means that no individual adjustments are needed, as is the case of binaural presentation, and there are no prescribed positions for the loudspeakers. A decoder is used to calculate the loudspeaker signals that a specific system must use to recreate the target sound field.

The input to a sound field reproduction system consists of a description of the target sound field. A widely used characterization of sound field information is known as the spherical harmonic expansion [15]:

$$\psi_k(\vec{r}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \mathrm{B}_{nm}(k) \mathrm{R}_n(kr) \mathrm{Y}_{nm}(\theta, \varphi). \tag{5}$$

This equation uses the spherical coordinate system shown in Fig. 2. Equation (5) is expressed in the frequency domain through k, the wavenumber, which is related to the angular frequency ω by the speed of sound c using the formula $k = \omega/c$. The radial functions $\mathrm{R}_n(kr)$ are combinations of spherical Bessel and spherical Hankel functions [15], [29]. It is common to ommit them to reduce system complexity

since they are related to the sound source distance [30] and do not affect the angular resolution of the system. The spherical harmonic function of order n and degree m, denoted as $\mathrm{Y}_{nm}(\theta,\varphi)$, is defined in terms of the Legendre polynomials $\mathrm{P}_{n,m}$ by the formula

$$\mathrm{Y}_{nm}(\theta,\varphi)=\begin{cases}\sqrt{\frac{2n+1}{4\pi}\cdot\frac{(n+m)!}{(n-m)!}}\mathrm{P}_{n,-m}(\sin\varphi)e^{im\theta} & m<0\\ \sqrt{\frac{2n+1}{4\pi}}\mathrm{P}_{n,0}(\sin\varphi) & m=0\\ (-1)^m\sqrt{\frac{2n+1}{4\pi}\cdot\frac{(n-m)!}{(n+m)!}}\mathrm{P}_{n,m}(\sin\varphi)e^{im\theta} & m>0\end{cases} \tag{6}$$

The expansion coefficients $\mathrm{B}_{nm}(k)$ are referred to as the HOA encoding of the sound field $\psi_k(\vec{r})$. Practical systems consider these coefficients up to a maximum order $\mathrm{N}_{\max}$.

An explicit formula for the HOA encoding of the sound field due to a plane wave incident from azimuth θ_{inc} and elevation φ_{inc} is given in [15]:

$$\mathrm{B}_{nm} = -4\pi i^n \mathrm{Y}^*_{nm}(\theta_{\mathrm{inc}},\varphi_{\mathrm{inc}}). \tag{7}$$

Here $|\cdot|^*$ denotes the complex conjugate.

Equations (5) and (7) are valid for all directions, including those outside of the horizontal plane. However, this paper will consider only the case where the elevation angle $\varphi = 0$. This is justified since stereophonic systems, as well as the multi-channel surround systems considered in the previous Subsection, are also limited to the horizontal plane. In this case, all of the expansion coefficients $\mathrm{B}_{nm}(k)$ for which $|m| \neq n$ are zero. Therefore, a $\mathrm{N}_{\max}$-order HOA encoding in the horizontal plane consists of $2\mathrm{N}_{\max}+1$ sets of coefficients in the frequency domain, or signals in the time domain.

The HOA format can encode the spatial sound information corresponding to any desired sound field as a multichannel signal. These signals can be further downmixed into a stereo stream using Eq. (1) with a suitable encoding matrix. This reasoning led to a format known as Ambisonics UHJ [21], [22]. In particular, the encoding matrix used to represent first-order HOA data in the horizontal plane as a stereo signal is:

$$M=\begin{bmatrix}0.47-0.086\pi i & 0.93+0.128\pi i & 0.328\\ 0.47+0.086\pi i & 0.93-0.128\pi i & -0.328\end{bmatrix}. \tag{8}$$

The first column generates an omnidirectional component corresponding to $\mathrm{B}_{0,0}$; the second column encodes front-back information $\mathrm{B}_{1,1}$; the third column handles the left-right component $\mathrm{B}_{1,-1}$. The Ambisonics UHJ format can downmix HOA data for higher orders or data outside of the horizontal plane; however, in these cases it produces three or more output signals. Its application to stereo systems is limited to first-order data in the horizontal plane.

3. Synthesizing HOA Data from Extended Stereo Signals

Conventional stereo signals encode sound source positions only in front of the listener. However, some stereo signals found in modern multimedia contents use the methods described in the previous Section to encode spatial sound information beyond the classical stereophonic limits. Previously, we proposed a method to extract this spatial information and generate an HOA encoding in the horizontal plane [18], [19]. This Section provides an overview of this method, which forms the basis for a new stereo representation of spatial sound.

3.1 Stable Inversion of the Stereo Panning Law

The techniques described in Sect. 2 are based on the concept of a panning law which represents the left-right positions of sound images as an inter-channel amplitude difference, and their front-back positions as an inter-channel phase difference. Our proposed method attempts to recover these left-right and front-back coordinates by inverting the panning law. This is similar to Eq. (3); however, it considers all azimuth angles.

The first step in our proposal consists of calculating the inter-channel level and phase differences. These inter-channel differences correspond to the inferred sound source position along the front-back (x), and the left-right (y) axes; they can be calculated using the following formulas:

$$x(\omega) = \left(1+\frac{\arg\{S_{\mathrm{L}}(\omega)\}-\arg\{S_{\mathrm{R}}(\omega)\}}{\pi}\right) \bmod 2 - 1. \tag{9}$$

$$y(\omega) = \frac{|S_{\mathrm{L}}(\omega)| - |S_{\mathrm{R}}(\omega)|}{\max(|S_{\mathrm{L}}(\omega)|, |S_{\mathrm{R}}(\omega)|)}, \tag{10}$$

These formulas yield results in the interval $[-1, 1]$.

The inter-channel differences are enough to invert a panning law like that shown in Fig. 1. The result of doing this is shown in Fig. 3. However, this simple inversion does not result in stable positions for all sound images. Under some conditions, small variations in the stereo signals may lead to large changes in the inferred position for the sound image. The reason for this is that the inter-channel

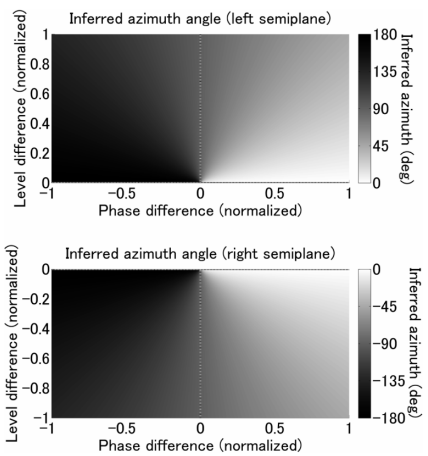

Fig. 3 Simple inversion of a stereo panning law. The azimuth angle for the target sound image is shown as a function of the inter-channel level and phase differences. There is one unstable point corresponding to zero inter-channel differences (sound images directly in front of the listener).

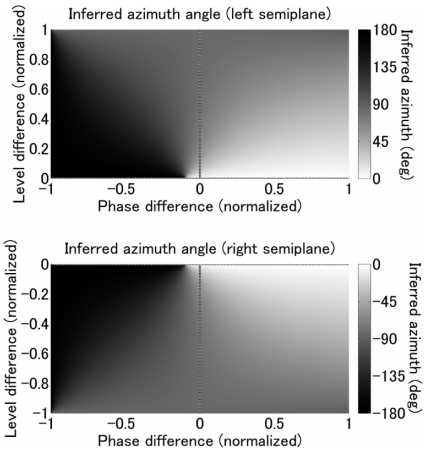

Fig. 4 Stable inversion of a stereo panning law after a non-linear warping of the horizontal plane. Inferred azimuth angles are shown in relation to the inter-channel differences. The unstable point is shifted to the back of the horizontal plane, ensuring the stable presentation of sound images in front of the listener.

differences may not be reliably calculated when they are small or may be undefined for sounds present in only one of the stereo channels. To remedy this, our proposal uses the non-linear warping of the coordinate system introduced in [18], [19].

$$\hat{x} = x + \tilde{\phi}(1 - x^2) - xy^2, \tag{11}$$

$$\hat{y} = y + \left[dx^3y - ex^4y\right] - \left[fx^3y^3 + gx^4y^3\right]. \tag{12}$$

The first correction in Eq. (11) shifts a singularity at the center of Fig. 3, ensuring the stable presentation of sound images at the front. This is achieved by a global shift along the front-back coordinate x of $\tilde{\phi}$ units. The shift is removed for points away from the singularity by the x^2 factor. The parameter $\tilde{\phi}$ must be small; an appropriate value to use with typical stereo signals is 0.1 [18]. The second correction, introduced by the $-xy^2$ term in Eq. (11), solves the problem of lateral sound images with undefined inter-channel phase differences. This correction places the sounds present in only one channel directly to the left of right, as appropriate.

Corrections along the left-right coordinate y are defined by four parameters d, e, f and g. The first two are used to expand the front half-space by $d - e$ while shrinking the back half-space by $d + e$. Our proposal recommends the values of $d + e = 0.9$ and $d - e = 0.3$ (i.e. $d = 0.6$ and $e = 0.3$) for typical stereo contents [18]. The last two parameters remove the corrections for sound sources directly in front of the listener. Therefore, they must satisfy $f + g = d - e$. The difference $g - f$ can be adjusted to further stabilize sound sources directly behind the listener. A recommended value for this is $g - f = 0.1$, and therefore $f = 0.1$ and $g = 0.2$ [18]. The result of applying Eqs. (11) and (12) to the inversion of the panning law are shown in Fig. 4.

3.2 High-Order Ambisonic Encoding

Inverting the panning law yields an inferred azimuth angle, $\theta(\omega) = \arctan[\hat{y}(\omega), \hat{x}(\omega)]$, for each frequency in the stereo signal. This can be used in combination with Eq. (7) to generate the HOA encoding. The corresponding sound field will contain all of the sound images present in the stereo signal as plane waves arriving from the directions encoded in the inter-channel differences.

Equation (7) deals with the spatial information of a plane wave field. The actual sound sources signals, however, must be extracted from the stereo data. This can be done by downmixing it to a monaural signal. Our method assumes the presence of important out-of-phase components; therefore, the downmixing should be carried out in the frequency domain. The monaural downmix $\mathrm{O}(\omega)$ can be calculated using the following formulas [18], [19]:

$$|\mathrm{O}(\omega)| = \sqrt{|S_\mathrm{L}(\omega)|^2 + |S_\mathrm{R}(\omega)|^2} \tag{13}$$

$$\mathrm{Ang}\left[\mathrm{O}(\omega)\right] = \begin{cases} \mathrm{Ang}\left[S_\mathrm{R}(\omega)\right] & \theta < 0 \\ \mathrm{Ang}\left[S_\mathrm{L}(\omega)\right] + \mathrm{Ang}\left[S_\mathrm{R}(\omega)\right] & \theta = 0 \\ \mathrm{Ang}\left[S_\mathrm{L}(\omega)\right] & \theta > 0 \end{cases} \tag{14}$$

The spatial information can be added by passing $\mathrm{O}(\omega)$ through a filterbank calculated from Eq. (7). Considering that the elevation angle φ is zero, the formulas for the spherical harmonics simplify into complex exponentials depending only on the degree m. The expansion coefficients are zero for all terms with $n \neq |m|$. The resulting filters $\mathrm{F}_m(\omega)$ are

$$\mathrm{F}_m(\omega) = \begin{cases} -\sin\left[m\theta(\omega)\right] & m < 0 \\ \cos\left[m\theta(\omega)\right] & m \geq 0 \end{cases} \tag{15}$$

Multiplying these filters by $\mathrm{O}(\omega)$ (equivalent to the convolution in the time domain) results in an HOA encoding inferred from the inter-channel differences of a stereo signal. The case $m = 0$ is just the monaural downmix (the omnidirectional component $\mathrm{B}_{0,0}$); the results for $m = 1$ are the front-back difference (the HOA component $\mathrm{B}_{1,1}$); when $m = -1$ the result is the left-right difference (corresponding to the HOA component $\mathrm{B}_{1,-1}$). The filters can be calculated to any desired order; however, the proposal does not yield significant improvements in spatial resolution above order 2 [18], [19].

4. Stereo Encoding of HOA Data

The method detailed in Sect. 3 can synthesize HOA data from a stereo signal by looking at its inter-channel differences. The proposal, however, assumes that the stereo source was mixed using a panning law similar to the one shown in Fig. 1. Applying the panning law is straightforward if the individual sound sources are available and their positions are known. However, this data is not directly available if the target sound has been already encoded in the HOA format. Furthermore, a significant advantage of HOA is that it allows for the direct recording of spatial sound using microphone arrays [31]. A method to downmix sound field recordings and HOA data to stereo while preserving

their spatial information is needed.

The simplest way to generate a stereo signal from HOA data is to discard all coefficients except the omnidirectional component $B_{0,0}$ and the left-right one $B_{1,-1}$. A stereo signal can be calculated by taking the sum and the difference between these two HOA channels. This approach, while simple to implement, preserves only a minimum of the spatial sound information found in the original HOA data.

Another alternative is to look at the sound field characterized by the HOA data and simulate its recording using virtual directional microphones. This approach uses a technique known as beamforming [32]. It has been successfully used to extract sound sources and their directions from first-order Ambisonics data [33]. The method can be extended to higher orders; however the intermediate microphone simulation stage introduces additional parameters and sources of inaccuracy in the system.

In this Section, we propose a new method to represent HOA data as a stereo signal. The proposal follows the procedure of Sect. 3 in reverse order and results in a stereo signal that can be decoded back into the HOA format by our previously proposed method [18], [19].

4.1 Recovering Azimuth Angles from HOA Data

As previously stated, our proposal considers only sound sources in the horizontal plane. Therefore, the input to our system consists of $2N_{max} + 1$ channels containing the HOA encoding of a sound field.

The omnidirectional component $B_{0,0}$ carries no spatial information; however, it is a common part of all channels in the HOA data. Ideally, the relationship between this and the other channels should be described by Eq. (15). This can be summarized in matrix notation:

$$\begin{bmatrix} B_{0,0}(\omega) \\ B_{1,-1}(\omega) \\ B_{1,1}(\omega) \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ -\sin[\theta(\omega)] \\ \cos[\theta(\omega)] \\ \vdots \end{bmatrix} B_{0,0}(\omega). \tag{16}$$

At a given frequency, Eq. (16) reduces to the multiplication of a vector by a scalar. This can be easily inverted; however, since the HOA coefficients are sound signals a more stable approach is to calculate the deconvolution of each channel by $B_{0,0}$. This can be done by computing an inverse filter, through a Wiener filter or using more sophisticated, non-linear algorithms if required by the source signals [34]. Another important consideration is the range of the sine and cosine functions; the results of the deconvolution must be wrapped inside the interval $[-1, 1]$. The inverse of Eq. (16) is then

$$\begin{bmatrix} 1 \\ \tilde{\theta}_{-1}(\omega) \\ \tilde{\theta}_{1}(\omega) \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ -\arcsin\left[\mathrm{frac}\left\{B_{1,-1}(\omega) / B_{0,0}(\omega)\right\}\right] \\ \arccos\left[\mathrm{frac}\left\{B_{1,1}(\omega) / B_{0,0}(\omega)\right\}\right] \\ \vdots \end{bmatrix}. \tag{17}$$

Where $\cdot/\cdot$ denotes the deconvolution of two signals and frac$\{\cdot\}$ stands for the fractional part of a number. The inferred azimuth angles $\tilde{\theta}_n(\omega)$ are expected to be similar since they correspond to the same sound field. Our proposal considers the mean value as the inferred azimuth angle

$$\theta(\omega) = \frac{1}{2N_{max}} \sum_{\substack{n=-N_{max} \\ n\neq 0}}^{N_{max}} \tilde{\theta}_n(\omega). \tag{18}$$

4.2 Generating the Stereo Signal

As previously stated, stereo signals can be generated by applying a panning law to a common monaural signal. Obtaining this common signal is straightforward since it corresponds to the omnidirectional component $B_{0,0}$. The panning law consists of a weight and a phase shift set by a target azimuth angle.

Equation (18) provides an azimuth angle for each frequency. It is possible to apply the panning law in Fig. 1 frequency-by-frequency at these angles to generate a stereo signal. The results of this approach are accurate along the left-right axis; however, the simple panning law encodes the front-back axis as either positive or negative polarity. The HOA data is not limited to these two values and may contain sounds that should be presented from any position along the front-back axis. To account for this, we propose a new method to calculate the inter-channel amplitude and phase differences from the inferred azimuth angles in Eq. (18). The goal is to ensure that the inter-channel differences can be used by the method in Sect. 3 to recover the original HOA data.

The first step is to calculate the front-back ($\hat{x}$) and left-right ($\hat{y}$) coordinates that correspond to each azimuth angle:

$$\begin{aligned} \hat{x} &= \cos[\theta(\omega)], \\ \hat{y} &= \sin[\theta(\omega)]. \end{aligned} \tag{19}$$

These correspond to the values calculated in Eqs. (11) and (12). It is now necessary to perform the inverse of the spatial warping introduced by the method proposed in Sect. 3. Inverting a system of polynomial equations like Eqs. (11) and (12) is a difficult problem and a solution is not guaranteed. Numerical methods can yield some results; however, a better approach is to consider the geometric meaning behind each of the corrections introduced by the warping equations.

The first correction is the global shift by $\tilde{\phi}$ along the x coordinate. This can be easily reversed by changing the sign of the parameter $\tilde{\phi}$. The second correction stabilizes lateral sources; this is not a concern when generating the stereo signals as long as the amplitude differences are calculated correctly. Therefore, the normalized inter-channel phase difference x can be calculated as

$$x = \hat{x} - \tilde{\phi}(1 - \hat{x}^2). \tag{20}$$

Once the value for x is established, Eq. (12) becomes a simple polynomial equation which can be solved by

factorization. This yields the normalized inter-channel amplitude difference y. Finally, these inter-channel differences can be applied to the omnidirectional component to generate a stereo signal as follows:

$$\begin{bmatrix} S_{\mathrm{L}}(\omega) \\ S_{\mathrm{R}}(\omega) \end{bmatrix} = \begin{bmatrix} \frac{1+y}{2} e^{\frac{x}{2}} \\ \frac{1-y}{2} e^{-\frac{x}{2}} \end{bmatrix} \mathrm{B}_{0,0}(\omega). \tag{21}$$

Some matrix-encoded surround methods take special care to avoid left and right channel signals of opposite polarities [23]. This is not the case in the proposed downmixing method. The proposal intentionally retains phase differences even if the resulting left and right channel signals have opposite polarities to improve front-back positioning of the sound sources. However, this adversely affects reproduction over typical stereo systems.

5. Evaluation

The method proposed in Sect. 4 can generate stereo signals from HOA encodings of sound fields. On the other hand, the method outlined in Sect. 3 does the opposite and synthesizes HOA data from stereo signals. Together, these methods form a system that can downmix HOA data for its distribution using conventional systems. On the receiving end, the stereo stream can be decoded back into an approximation of the original HOA data.

To evaluate the performance of our proposal, we consider the simple signal shown in Fig. 5. It consists of a 1 kHz pure tone multiplied by a 10 ms Hanning window. This signal was encoded using second-order HOA. The spatial information corresponds to a plane-wave field incident from different directions in the horizontal plane taken at intervals of 1°. The resulting HOA coefficients for a frequency of 1 kHz are shown in Fig. 6.

The HOA data for each of the 360 directions in the horizontal plane was downmixed to stereo using the method proposed in Sect. 4. Figure 7 shows four of the resulting stereo signals. These are consistent with the results expected from a panning law like that of Fig. 1. There are no inter-channel differences at the front, highly lateral signals appear at 90° and 270°, signals of opposite polarity represent a sound source behind the listener. Small errors are visible in the signals for the left and right directions. These are not significant since their contribution to the inter-channel level difference is, at most, −26 dB (approx. 0.05 in a normalized linear scale).

The stereo signals obtained with the proposed method were then processed using the algorithm outlined in Sect. 3. The reconstructed HOA signals for 1 kHz are shown in Fig. 8. The wider bands, compared to those of Fig. 6, indicate a slight loss in spatial resultion. Nevertheless, the

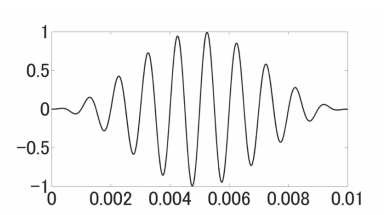

Fig. 5 The test signal used for evaluation: a 1 kHz pure tone multiplied by a 10 ms Hanning window.

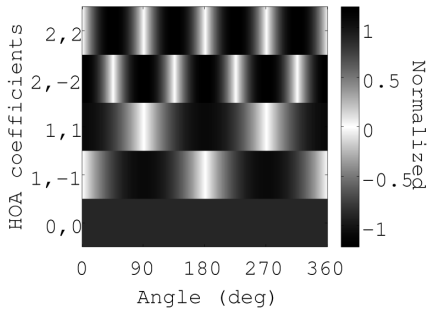

Fig. 6 The second-order HOA encoding coefficients for plane waves in the horizontal plane.

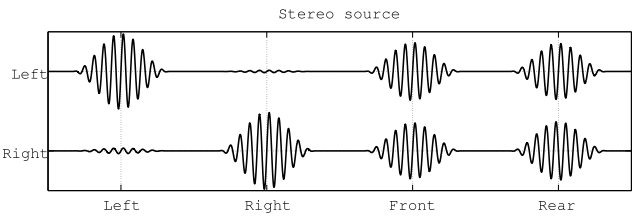

Fig. 7 The stereo signals generated by the proposed method for four representative directions.

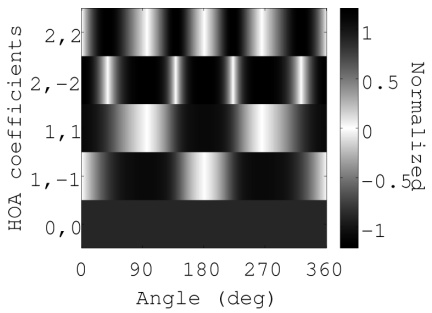

Fig. 8 The second-order Ambisonics spatial data recovered by the proposed method.

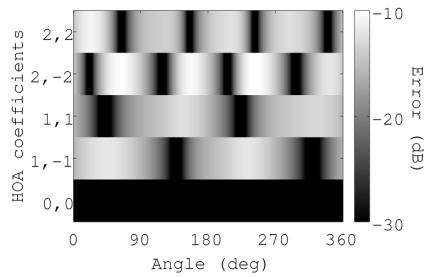

Fig. 9 Difference between the original HOA data and the results after downmixing to stereo and later recovering the HOA data using the proposed methods.

proposal yields consistent results for all azimuth angles. The difference between the recovered HOA encoding and the original one appears in Fig. 9. The largest deviation from the target data occur at azimuth angles close to 90° and 270°. The maximum deviation reaches a level of −10.3 dB with respect to the peak of the target data. In a normalized linear scale, this corresponds to a deviation of approximately 0.31. From these results, we conclude that the methods described in Sect. 3 and 4 can be applied to generate stereo signals directly from HOA data. Furthermore, it is possible to recover a good approximation of the original HOA data from the inter-channel differences in the resulting stereo signals.

6. Conclusions

We proposed a method to represent spatial sound information encoded in the HOA format using stereo signals. The resulting signals are consistent with traditional panning laws. Furthermore, they preserve the spatial information available in the original HOA data encoded as inter-channel level and phase differences. This allows us to recover an approximation of the original HOA data using a previously proposed technique. Simulation results show that the proposed method retains adequate spatial resolution when applied to second-order HOA encodings. In this way, the proposal outperforms other techniques such as the Ambisonics UHJ format, which is limited to first-order HOA when downmixing to stereo signals.

Acknowledgments

This study was partly supported by Grant-in-Aid of JSPS for Scientific Research (no. A24240016) to SY and the A3 Foresight Program for "Ultra-realistic acoustic interactive communication on next-generation Internet."

References

[1] K. Abe, K. Ozawa, Y. Suzuki, and T. Sone, "Comparison of the effects of verbal versus visual information about sound sources on the perception of environmental sounds," Acta Acustica united with Acustica, vol.92, no.1, pp.51–60, 2006.

[2] J. Blauert, Spatial hearing: The psychophysics of human sound localization, Revised ed., The MIT Press, Cambridge MA USA, 1997.

[3] K. Ozawa, S. Tsukahara, Y. Kinoshita, and M. Morise, "Instantaneous Evaluation of the Sense of Presence in Audio-Visual Content," IEICE Trans. Inf. & Syst., vol.E98-D, no.1, pp.49–57, 2015.

[4] Y. Suzuki, J. Trevino, T. Okamoto, Z. Cui, S. Sakamoto, and Y. Iwaya, "High Definition 3D auditory displays and microphone arrays for the use with future 3D TV," Proc. of 3DSA 2013, paper no.132, June 2013.

[5] J. Trevino, T. Okamoto, C. Salvador, Y. Iwaya, Z. Cui, S. Sakamoto, and Y. Suzuki, "High-order Ambisonics auditory displays for the scalable presentation of immersive 3D audio-visual contents," Proc. ICAT 2013, paper no.D5, Dec. 2013.

[6] J. Kawaura, Y. Suzuki, F. Asano, and T. Sone, "Sound localization in headphone reproduction by simulating transfer functions from the sound source to the external ear," J. Acoust. Soc. Jpn. (J), vol.45, pp.756–766, 1989 (in Japanese), English translation: J. Acoust. Soc. Jpn. (E), vol.12, pp.203–216, 1991.

[7] S. Sakamoto, S. Hongo, and Y. Suzuki, "3D sound-space sensing system based on numerous symmetrically arranged microphones," IEICE Trans. Fundamentals, vol.E97-A, no.9, pp.1893–1901, Sept. 2014.

[8] C. Han, T. Okamoto, Y. Iwaya, and Y. Suzuki, "Loudspeaker distributions suitable for crosstalk cancellers robust to head rotation," Acoust. Sci. Technol., vol.33, no.4, pp.266–269, 2012.

[9] M. Noisternig, A. Sontacchi, T. Musii, and R. Holdrich, "A 3D Ambisonic based binaural sound reproduction system," Proc. Audio Eng. Soc. 24th Int. Conf. on Multichannel Audio, paper no.1, June 2003.

[10] Y. Iwaya, Y. Suzuki, and D. Kimura, "Effects of head movement on front-back error in sound localization," Acoust. Sci. Technol., vol.24, no.5, pp.322–324, 2003.

[11] E. Torick, "Highlights in the History of Multichannel Sound," J. Audio Eng. Soc., vol.46, no.1, pp.27–31, 1998.

[12] K. Hamasaki, T. Nishiguchi, R. Okumura, Y. Nakayama, and A. Ando, "A 22.2 Multichannel Sound System for Ultra-High-Definition TV (UHDTV)," SMPTE Motion Imaging J., vol.117, no.3, pp.40–49, 2008.

[13] A. Berkhout, "A holographic approach to acoustic control," J. Audio Eng. Soc., vol.36, no.12, pp.977–995, Dec. 1988.

[14] S. Ise, "A principle of sound field control based on the Kirchhoff-Helmholtz integral equation and the theory of inverse systems," Acta Acoust. united Ac., vol.85, pp.78–87, 1999.

[15] M.A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," J. Audio Eng. Soc., vol.53, no.11, pp.1004–1025, 2005.

[16] M. Noisternig, T. Carpentier, and O. Warusfel, "ESPRO 2.0 - Implementation of a surrounding 350-loudspeaker array for 3D sound field reproduction," Proc. 4th Int. Symp. on Ambisonics and Spherical Acoust., paper no.13, March 2012.

[17] T. Okamoto, D. Cabrera, M. Noisternig, B. Katz, Y. Iwaya, and Y. Suzuki, "Improving sound field reproduction in a small room based on high-order Ambisonics with a 157-loudspeaker array," Proc. 2nd Int. Symp. on Ambisonics and Spherical Acoust., paper no.5, May 2010.

[18] J. Trevino, T. Okamoto, Y. Iwaya, J. Li, and Y. Suzuki, "Extrapolation of horizontal Ambisonics data from mainstream stereo sources," Proc. IIH-MSP 2013, pp.302–305, Oct. 2013.

[19] J. Trevino, T. Okamoto, Y. Iwaya, J. Li, and Y. Suzuki, "A spatial extrapolation method to derive high-order ambisonics data from stereo sources," J. Inf. Hiding and Multimedia Sig. Proc., vol.6, no.6, pp.1100–1116, Nov. 2015.

[20] J. Trevino, T. Okamoto, Y. Iwaya, and Y. Suzuki, "Sound field reproduction using Ambisonics and irregular loudspeaker arrays," IEICE Trans. Fundamentals, vol.E97-A, no.9, pp.1832–1839, Sept. 2014.

[21] M.A. Gerzon, "Compatible 2-channel encoding of surround sound," Electron. Lett., vol.11, no.25, pp.615–617, Dec. 1975.

[22] M.A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," J. Audio Eng. Soc., vol.33, no.11, pp.859–871, Nov. 1985.

[23] K. Gundry, "A New Active Matrix Decoder for Surround Sound," 19th Audio Eng. Soc. Int. Conf. on Surround Sound, paper no.1905, 9-page manuscript, June 2001.

[24] D. Griesinger, "Multichannel Matrix Surround Decoders for Two-Eared Listeners," Proc. 101th Audio Eng. Soc. Conv., preprint no.4402, Nov. 1996.

[25] E.G. Trendell, "The Choice of a Matrix for Quadraphonic Reproduction from Disk Records," Proc. 47th Audio Eng. Soc. Conv., paper no.E-7, March 1974.

[26] Dolby Surround Pro Logic II Decoder: Principles of Operation, Dolby Laboratories Technical Paper, 2000.

[27] D.M. Leakey, "Some measurements on the effects of interchannel intensity and time differences in two channel sound system," J. Acoust. Soc. Am., vol.31, no.7, pp.977–986, 1959.

[28] M. Poletti, "The design of encoding functions for stereophonic and polyphonic sound systems," J. Audio Eng. Soc., vol.44, no.11, pp.948–963, 1996.

[29] E.G. Williams, Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography, Academic Press, London UK, 1999.
[30] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonics format," 23rd Int. Conf. Audio Eng. Soc. Sig. Proc. in Audio Recording and Reproduction, 15-page manuscript, May 2003.
[31] S. Bertet, J. Daniel, and S. Moreau, "3D Sound Field Recording with Higher Order Ambisonics-Objective Measurements and Validation of Spherical Microphone," Proc. 120 Audio Eng. Soc. Conv., paper no.6857, 2006.
[32] E. Tiana-Roig, F. Jacobsen, and E.F. Grande, "Beamforming with a circular microphone array for localization of environmental noise sources," J. Acoust. Soc. Am., vol.128, no.6, pp.3535–3542, Dec. 2010.
[33] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," J. Audio Eng. Soc., vol.55, no.6, pp.503–516, 2007.
[34] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, "Convolution and Deconvolution Using the FFT," Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd Ed., Cambridge University Press, 1992.

Jorge Trevino graduated from the Monterrey Institute of Technology and Higher Education in 2005. He received the degree of M.Sc. in 2011 and a Ph.D. in information sciences in 2014, both from the Graduate School of Information Sciences of Tohoku University. He is currently an assistant professor in the Research Institute of Electrical Communication of Tohoku University. His research interests include sound field recording and reproduction, array signal processing and spatial audio.

Shuichi Sakamoto received his B.S., M.Sc. and Ph.D. degrees from Tohoku University, in 1995, 1997, and 2004, respectively. He is currently an associate professor at the Research Institute of Electrical Communication, Tohoku University. He was a Visiting Researcher at McGill University, Montreal, Canada during 2007–2008. His research interests include human multi-sensory information processing including hearing, speech perception, and development of high-definition 3D audio recording systems. He is a member of ASJ, IEICE, VRSJ, and others.

Junfeng Li received the Ph.D. degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in March 2006. From April 2006, he was a postdoctoral research fellow at Research Institute of Electrical Communication (RIEC), Tohoku University. From April 2007 to July 2010, he was an Assistant Professor in School of Information Science, JAIST. Since August 2010, he has been a Professor in Institute of Acoustics, Chinese Academy of Sciences. His research interests include psychoacoustics, speech signal processing and 3D audio technology. Dr. Li received the best student award in Engineering Acoustics First Prize from the Acoustical Society of America in 2006, and the Best Paper Award from JCA2007 in 2007, and the Itakura Award from the Acoustical Society of Japan in 2012. Dr. Li is now serving as the Subject Editor for Speech Communication and the Editor for IEICE Trans. on Fundamentals of Electronics, Communication and Computer Sciences.

Yôiti Suzuki graduated from Tohoku University in 1976 and received his Ph.D. degree in electrical and communication engineering in 1981. He is currently a professor at the Research Institute of Electrical Communication, Tohoku University. His research interests include psychoacoustics, multimodal perception, high-definition 3D auditory displays and digital signal processing of acoustic signals. He received the Awaya Kiyoshi Award and Sato Prize from the Acoustical Society of Japan as well as FIT Funai Best Paper award.