Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **6**, November **2015**

A Spatial Extrapolation Method to Derive High-Order Ambisonics Data from Stereo Sources

Jorge Trevino and Yôiti Suzuki

Graduate School of Information Sciences and Research Institute of Electrical Communication
Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai, Japan
jorge@ais.riec.tohoku.ac.jp; yoh@riec.tohoku.ac.jp

Takuma Okamoto

Universal Communication Research Institute
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
okamoto@nict.go.jp

Yukio Iwaya

Faculty of Engineering
Tohoku Gakuin University
1-13-1 Chuuou, Tagajou-shi, Miyagi, Japan
iwaya.yukio@mail.tohoku-gakuin.ac.jp

Junfeng Li

Institute of Acoustics
Chinese Academy of Sciences
No. 21 Beisihuan Xilu, Haidian, Beijing, China
lijunfeng@hccl.ioa.ac.cn

Received July, 2015; revised August, 2015

Abstract. *Recent advances in computer and multimedia technologies make the natural presentation of three-dimensional contents possible. Sound field reproduction techniques, such as High-Order Ambisonics (HOA), are now within the reach of commercially available computation devices. Unfortunately, content production and its coding for distribution still focuses on legacy formats, mainly stereophonic sound. This paper presents a method to extrapolate full-surround, HOA-encoded sound fields from conventional stereo mixes. The proposal assumes its input to be mixed using a conventional panning law. Its formulation resembles traditional matrix-encoded surround methods; however, it does not assume any specific layout for the loudspeakers comprising the reproduction system. The output of the proposed method is the second-order HOA encoding of a sound field in the horizontal plane spanning the full 360-degrees panorama. In this sense, the proposal outperforms traditional matrix-encoding methods by allowing sound sources to be placed freely around the listener. It also has better spatial resolution than the Ambisonics UHJ format, which is limited to first-order horizontal encodings when working with stereo signals. The proposed method is evaluated using simple stereo mixes of a wavelet panned to different positions on the horizontal plane. The extrapolated HOA-encoding exhibits stable sound-source positioning and negligible crosstalk between disjoint sources.*

Keywords: Spatial sound, High-Order Ambisonics, Sound field reproduction, Spatialization, Surround

1. **Introduction.** Fast-paced technological advances in computing and telecommunications have made ultra-realistic presentation systems possible [1, 2]. A basic requirement for these systems is the capability to convey spatial sound accurately. In this context, most spatial sound presentation systems can be classified into three categories: binaural reproduction, multi-channel surround sound and sound field reproduction.

Binaural systems work by reproducing the sound pressure at the listener's ears [3, 4, 5, 6]. They can produce very realistic results; however, their implementation relies on individual measurements of a head-related transfer function (HRTF) [7]. Furthermore, the signal processing stage must be continuously updated to account for changes in the listener's position and head orientation [8].

Multi-channel surround sound relies on a predefined loudspeaker arrangement to present sound from multiple directions [9, 10]. The main advantage of this approach is the simplicity of its reproduction stage which requires no special processing of the audio signals. However, mainstream implementations such as 5.1-channel surround lack the spatial resolution offered by other approaches. In addition, audio contents must be prepared with a specific loudspeaker system in mind; users must use a reproduction system which adheres to the assumed specification.

Sound field reproduction systems seek to re-create the sound pressure over an extended region enclosing the listeners [11, 12, 13]. This approach has several advantages over the others. It does not require individual adjustments for each listener or their positions and orientations as is the case with binaural systems. Sound field reproduction can achieve extremely high spatial resolutions by using the available loudspeakers in a more efficient manner than multi-channel surround systems. On the other hand, sound field reproduction requires complex signal processing and large channel counts which were impractical until recent years. This has limited its historical use to research facilities and technical demonstrations [14, 15].

While modern computing and multi-channel audio systems make sound field reproduction possible for end-users, the lack of contents that take advantage of the technology has hindered its adoption. In this paper we seek to ameliorate this problem by proposing a method to extrapolate the spatial information needed for sound field reproduction from mainstream stereo recordings. In particular, the proposed method uses a sound field encoding format known as High-Order Ambisonics (HOA) [13]. The main reason behind this choice is that the HOA format is system-agnostic; it can be reproduced using almost any loudspeaker array with a suitable decoder [16] and can also be applied in binaural presentation systems [6].

There have been some efforts to make HOA compatible with conventional stereo and multi-channel systems. An encoding method known as the Ambisonics UHJ format can represent HOA data using a small number of audio channels [17, 18]. However, this format has low spatial resolution. It cannot go beyond the first-order of horizontal Ambisonics (the smallest improvement over monaural sound) when limited to stereo signals. The method advanced in this paper, while requiring a more complex decoder than UHJ, has better spatial resolution extending beyond the first Ambisonic order.

The proposed method extracts spatial information from the inter-channel level and phase differences of stereo signals. Differences in level are assumed to encode the left-right coordinates of sound souces, in agreement with conventional stereo mixing. Information regarding the front-back coordinates, on the other hand, should not affect stereophonic reproduction. Therefore, these coordinates are assumed to be encoded as phase differences, since phase is an effective way to hide information in acoustic signals [19, 20]. Looking at the two channels in a stereo signal, rather than data hidden in individual channels not only makes our proposal compatible with existing stereo recordings, it also improves its robustness to changes in the signals [21], for example due to lossy compression. The endurance of information embedded in two independent but similar datasets has been explored in the context of watermarking for 3D video [22, 23].

Section 2 of this manuscript presents an overview of modern stereo sources emphasizing the properties used by the proposed method. In Section 3 we offer a brief introduction to HOA and an existing method to make Ambisonics compatible with stereo systems. The proposed method is outlined in Section 4. This method builds on our previous results [24] with important improvements in the HOA encoding stage. We evaluate the proposed algorithm in Section 5 using stereo mixings of a simple sound signal. Finally, we draw our conclusions for this study in Section 6.

2. **Stereophonic sound.**

Stereophonic sound consists on the use of two independent audio signals to achieve some degree of spatial sound reproduction. The resolution of stereo systems is very limited in comparison to other spatial audio systems. Nevertheless, its simplicity has led to its widespread adoption. It is compatible with most modern methods to distribute audio contents, from optical discs to codecs for network streaming.

A typical stereo reproduction system uses two loudspeakers placed in front of the listener, usually at azimuth angles of $-30°$ and $30°$ measured from the front. The two channels are independent and can be used for different sound sources. However, modern stereo signals also use common components with different amplitudes in order to produce a sound image between the loudspeakers. More complex signals also modulate the phase of each channel to widen the reproduction panorama.

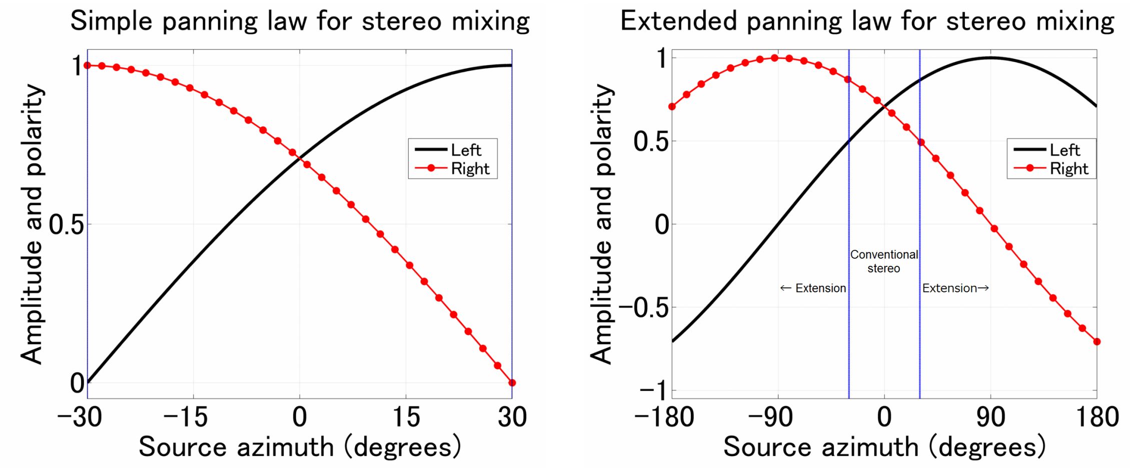
2.1. **Panning laws.**

Stereophonic sound can convey the illusion of sound sources located between the loudspeakers. This is achieved by feeding the same signal with different amplitudes to both channels. A panning law defines the weights required to modulate each channel's amplitude and position the sound image at a desired angle. Figure 1a shows a typical cosine panning law ensuring the total energy of both channels remains the same for all sound image positions. It can be seen that the inter-channel amplitude difference encodes the position along the left-right axis.

Panning laws can be extended beyond the $-30°$ to $30°$ region covered by the loudspeakers. Stereo reproduction systems cannot pan the sound image outside this region; however the additional spatial information is retained in the stereo mix and can be recovered through signal processing methods [25]. Figure 1b shows a cosine panning law spanning the entire 360-degrees panorama around the listener. The vertical lines delimit the region covered by the non-extended panning law. Notice that amplitude gains peak at $\pm 90°$ instead of $\pm 30°$; this leads to both channels having the same polarity for sound images on the front and opposite signs when the image is on the back. The inter-channel phase difference is, thus, encoding the position along the front-back direction.

2.2. **Matrix-encoded surround.**

One application of extended panning laws is to encode multi-channel surround signals as stereo data. Surround sound systems do not require the continuous encoding of all directions; it is enough to find amplitude and phase differences to encode the positions of the loudspeakers in the target reproduction system. This leads to a simple formulation in matrix notation [26] known as matrix-encoded surround. The stereo signals $\mathrm{Left}(\omega)$ and $\mathrm{Right}(\omega)$ are generated by mixing the N-channel surround data



(A)

(B)

FIGURE 1. Two stereo panning laws. Panel (a) shows a typical panning law used to position a sound image between the left and right loudspeakers in a stereo reproduction system. Panel (b) illustrates an extended panning law encoding spatial information for all directions in the horizontal plane.

using a 2-by-N encoding matrix M

$$\begin{bmatrix}\mathrm{Left}(\omega)\\ \mathrm{Right}(\omega)\end{bmatrix} = \mathrm{M}\begin{bmatrix}\mathrm{Ch}_1(\omega)\\ \mathrm{Ch}_2(\omega)\\ \vdots\\ \mathrm{Ch}_N(\omega)\end{bmatrix}. \tag{1}$$

The resulting signals can be reproduced by a conventional stereo system or decoded with the inverse of Eq. (1) to recover an approximation to the original multi-channel data

$$\begin{bmatrix}\mathrm{Ch}_1(\omega)\\ \mathrm{Ch}_2(\omega)\\ \vdots\\ \mathrm{Ch}_N(\omega)\end{bmatrix} \approx \mathrm{M}^{+}\begin{bmatrix}\mathrm{Left}(\omega)\\ \mathrm{Right}(\omega)\end{bmatrix}. \tag{2}$$

Here, M^{+} denotes a pseudo-inverse of M and is referred to as the decoding matrix. The elements of M and M^{+} are, in general, complex numbers. This allows the encoding matrix to modulate both the inter-channel amplitude and phase differences. These matrices do not depend on the frequency ω. An example used to encode four-channel surround is [27]:

$$\mathrm{M} = \begin{bmatrix}0.92 & 0.38 & 0.46\pi i & 0.19\pi i\\ 0.38 & 0.92 & -0.19\pi i & -0.46\pi i\end{bmatrix}. \tag{3}$$

Some commercial products rely on more elaborate methods to encode multi-channel surround as a stereo signal [28]. These methods add feedback loops to modulate the decoding matrix according to the sound source. Nevertheless, they rely on the same notion of encoding left-right position as an amplitude difference and front-back position as a phase difference. The proposal introduced in this paper will also follow this approach. However, it will extrapolate a complete sound field on the horizontal plane rather than a set of multi-channel signals.

3. **High-Order Ambisonics.** Sound field reproduction systems are based on physical models for sound propagation. In particular, they consider the spatial part of the wave equation, known as the Helmholtz equation [29]:

$$\left(\nabla^2 + k^2\right)\psi_k(\vec{r}) = 0, \tag{4}$$

where $\psi_k(\vec{r})$ denotes the sound field, i.e. the sound pressure at any position $\vec{r}$. k stands for the wavenumber, related to the angular frequency ω by the speed of sound c using the formula $k = \omega/c$.

Solving the Helmholtz equation requires specifying a coordinate system and a set of boundary conditions. Different choices lead to different formulas and to several sound field reproduction methods [11, 12, 13]. Among them, a method based on the solution in spherical coordinates, known as High-Order Ambisonics (HOA) stands out by defining a general way to encode sound field information. The HOA encoding makes no assumptions on the way sounds are recorded or reproduced. It can be used to represent spatial sound information for use with almost any sound reproduction system.

The HOA encoding uses the spherical harmonic functions defined as:

$$\mathrm{Y}_{nm}(\theta,\varphi) = \begin{cases} \sqrt{\frac{2n+1}{4\pi}\cdot\frac{(n+m)!}{(n-m)!}}\mathrm{P}_{n,-m}(\sin\varphi)e^{im\theta} & m < 0 \\ \sqrt{\frac{2n+1}{4\pi}}\mathrm{P}_{n,0}(\sin\varphi) & m = 0 \\ (-1)^m\sqrt{\frac{2n+1}{4\pi}\cdot\frac{(n-m)!}{(n+m)!}}\mathrm{P}_{n,m}(\sin\varphi)e^{im\theta} & m > 0 \end{cases} \tag{5}$$

Here, n is referred to as the order and is related to the spatial frequency (and, therefore, the resolution) of the harmonic functions. Index m is called the degree and is associated with the orientation of the harmonic. Functions $\mathrm{P}_{n,m}$ are the Legendre polynomials [29]. Coordinates θ and φ stand, respectively, for the azimuth and elevation angles in the spherical coordinate system.

In HOA, the target sound field is represented as a weighted sum of spherical harmonic functions [13]:

$$\psi_k(\vec{r}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\mathrm{B}_{nm}(k)\mathrm{R}_n(kr)\mathrm{Y}_{nm}(\theta,\varphi). \tag{6}$$

Here, r, denotes the radial coordinate in the spherical coordinate system. The functions $\mathrm{R}_n(kr)$ are products of spherical Bessel and spherical Hankel functions [13, 29]. However, they can be assumed to be constants and absorbed into the expansion coefficients $\mathrm{B}_{nm}(k)$ without any impact in angular resolution. This simplifies the formulation of HOA, but removes the possibility of encoding distance information [30].

The coefficients $\mathrm{B}_{nm}(k)$, truncated to a maximum order $n = \mathrm{N}_{\max}$, are the $\mathrm{N}_{\max}^{\mathrm{th}}$-order Ambisonic encoding of the sound field $\psi(\vec{r})$. They can be calculated using the general formula [29]:

$$\mathrm{B}_{nm}(k) = \frac{1}{\mathrm{R}_n(kr)}\iint_{4\pi}\psi_k(\vec{r})\mathrm{Y}^*_{nm}(\theta,\varphi)\cos\varphi d\varphi d\theta, \tag{7}$$

where $|\cdot|^*$ denotes the complex conjugate. After removing the radial dependency, the following formula gives their explicit values for the field of a plane wave incident from azimuth θ_{inc} and elevation φ_{inc} [13]:

$$\mathrm{B}_{nm} = -4\pi i^n\mathrm{Y}^*_{nm}(\theta_{\mathrm{inc}},\varphi_{\mathrm{inc}}). \tag{8}$$

This equation will be used in the proposed algorithm to encode the directions of incidence of all sound sources present in the stereo mix.

3.1. **Ambisonics UHJ format.** There have been some attempts to represent Ambisonics data using matrix encoders like that of Eq. (1). In particular, one approach known as the Ambisonics UHJ format can encode sound fields in the horizontal plane as a stereo signal. It uses the following encoding matrix [17, 18]:

$$M = \begin{bmatrix} 0.47 - 0.086\pi i & 0.93 + 0.128\pi i & 0.328 \\ 0.47 + 0.086\pi i & 0.93 - 0.128\pi i & -0.328 \end{bmatrix}. \tag{9}$$

The first column encodes the omnidirectional component $\mathrm{B}_{0,0}$, the second one encodes the front-back portion $\mathrm{B}_{1,1}$ and the last column encodes the left-right coefficient $\mathrm{B}_{1,-1}$.

The main limitation of the UHJ format, when used to encode and decode stereo signals, is that it cannot go beyond first-order Ambisonics. Our proposal manages to recover second-order Ambisonic encodings by using a frequency-dependent decoding matrix, as will be shown in the next Section.

4. **Extrapolation of full-surround sound fields from stereo sources.** As outlined in Section 2, stereo signals can carry spatial sound information beyond the limits of conventional stereophonic reproduction. In particular, the use of an extended panning law, perhaps in the form of an encoding matrix for multi-channel surround, codifies the left-right coordinate as an amplitude difference and the front-back coordinate as a phase difference.

The proposed method seeks to recover the spatial information by inverting the panning law. However, a straightforward inversion does not yield stable results. Under certain conditions, there may be abrupt changes in the position of the sound sources. This phenomenon is suppressed by applying a non-linear transformation to the estimated coordinates for the sound image. In addition, a spatial warping stage is used to ensure the proposed method keeps most of the sounds in the front, in-line with conventional stereophonic reproduction. The inferred positions are used to generate a filter bank which can be applied to a monaural downmix of the input to synthesize a second-order Ambisonics encoding. The filter bank can be seen as a frequency-dependent matrix encoder.

4.1. **Panning law inversion: estimating angles of incidence.** The first step in the proposed method attempts to invert the panning law used to generate the stereo signals. If a formula is known for the panning law, direct inversion may be possible; however, this information is seldom available. Nevertheless, all common panning laws relate left-right positions (we will call this the y-coordinate, positive on the left side) to inter-channel amplitude differences. The front-back position (which we will refer to as the x-coordinate, positive towards the front), on the other hand, is encoded as the inter-channel phase difference. Therefore, our proposal starts by calculating a normalized version of these inter-channel differences as follows:

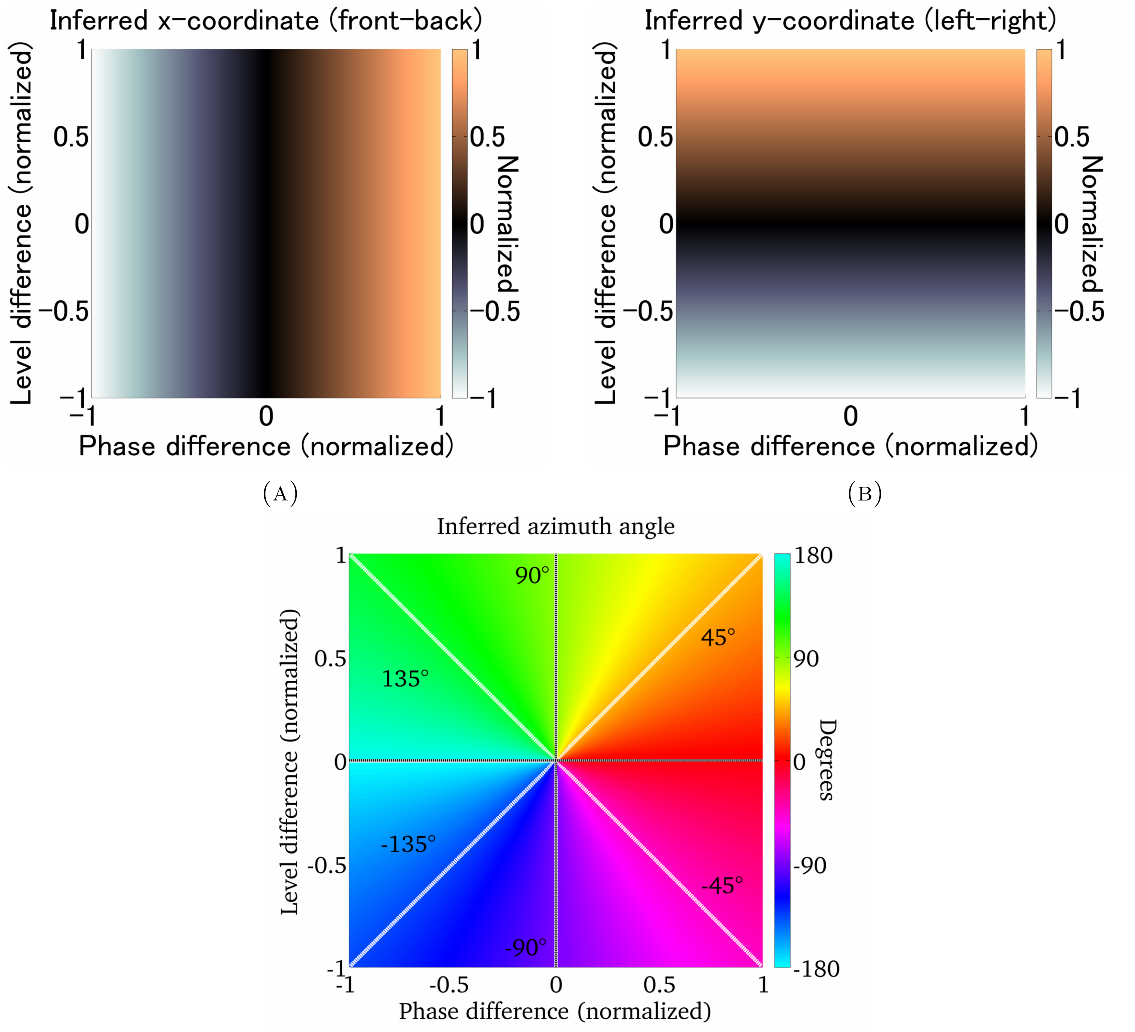
$$A(\omega) = \frac{|\mathrm{Left}(\omega)| - |\mathrm{Right}(\omega)|}{\max\left(|\mathrm{Left}(\omega)|, |\mathrm{Right}(\omega)|\right)}, \tag{10}$$

$$\phi(\omega) = \left(1 + \frac{\arg\left\{\mathrm{Left}(\omega)\right\} - \arg\left\{\mathrm{Right}(\omega)\right\}}{\pi}\right) \bmod 2 - 1. \tag{11}$$

These two equations estimate the inter-channel level difference $A(\omega)$ and phase difference $\phi(\omega)$ and restrict their values to the interval $[-1, 1]$.

The next step is to identify $A(\omega)$ as the y-coordinate and $\phi(\omega)$ as the x-coordinate. This allows us to infer an azimuth angle $\theta(\omega)$ for each frequency:

$$\theta(\omega) = \arctan\left[A(\omega), \phi(\omega)\right]. \tag{12}$$



(A)

(B)

(C)

FIGURE 2. Direct inversion of a typical panning law. Panel (a) shows the inferred front-back coordinate. Panel (b) shows the values for the left-right coordinate. Panel (c) shows the azimuth angle assigned to a sound source from the inter-channel differences. The left-right and front-back axes are shown in black; level contours for some representative angles are shown in white.

Equation (12) uses the four-quadrant arctangent resulting in azimuth angles between $-180°$ and $180°$]. The results of this approximation to inverting the panning law are shown in Fig. 2.

4.2. **Sound field stabilization.** The spatial information obtained by inverting the panning law is inadequate for direct encoding in the HOA format. This becomes clear by noting the case in which a sound image is intended to be presented directly in front of the listener. The panning laws of Figs. 1a and 1b produce the same signal for both stereo channels; that is, there are no inter-channel differences ($A(\omega) = 0$ and $\phi(\omega) = 0$). However, the azimuth angle derived from Eq. (12) is ill-defined for these conditions. The consequence is that sources that should be presented from the front will appear instead at random positions around the listener.

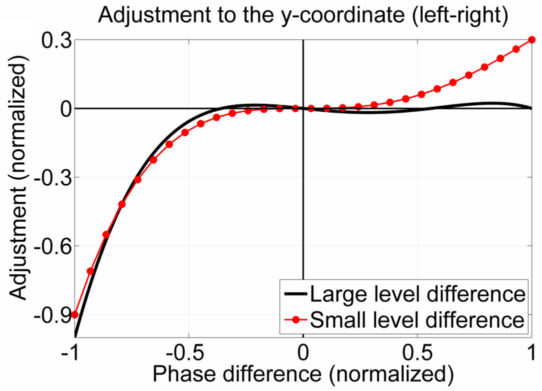


FIGURE 3. Corrections to the inferred y-coordinate introduced in the spatial warping stage. These corrections bring sources located behind the listener closer to each other. If the level difference is small (dotted red curve), the corrections will also make sources in the front semi-plane sparser, thus accounting for the common assumption of front loudspeakers used in typical stereo mixes.

The proposed panning-law inversion also struggles with highly lateral sound images. The extreme cases correspond to sounds that are present in only one of the two stereo channels. For these cases, the amplitude difference $A(\omega) = \pm 1$ and the phase difference is undefined. Sounds would be presented on the correct side, left or right, but would not have a stable position along the x-coordinate.

To account for these cases, we propose the following non-linear transformation of the coordinate system:

$$\hat{x} = x + \tilde{\phi}(1 - x^2) - xy^2. \tag{13}$$

The first correction in Eq. (13), given by the second term on the right-hand side, shifts the ill-defined point at the center of Fig. 2 towards the back by $\tilde{\phi}$ units. Meanwhile, the x^2 factor gradually reduces the shifting as the inferred sound position moves to more stable regions on the front or back. The second power is used to retain front-back symmetry. Choosing an adequate value for $\tilde{\phi}$ is a tradeoff between front-back position accuracy and stability. Empirical tests show that a value of around 0.1 works well with most stereo signals [24].

The second correction, introduced by the $-xy^2$ term, ensures that $x = 0$ for lateral sources with $y = \pm 1$. This guarantees that all sound images that should be presented directly to the left or right will appear at their intended positions despite having no definite inter-channel phase difference. Once again, the second power is used to make the correction gradual while preserving symmetry, in this case left-right symmetry.

4.3. **Spatial warping.** The transformation introduced in Eq. (13) is enough to ensure all sound sources in a stereo mix are assigned stable positions in the horizontal plane. However, its equal treatment of the front and back semi-planes is inconsistent with the front placement of loudspeakers in a typical stereo system. Most sound sources should be expected to be in the front semi-plane. The inconsistency originates from the assumption that the panning law in Fig. 1b was used. The extended panning law places the amplitude peaks at ±90° instead of ±30°, as is the case of the conventional one in Fig. 1a.

The overestimation towards the back of the sound source positions can be corrected by a spatial warping that makes the sources in the front semi-plane sparser. The panning law inversion already considers all directions; therefore, expanding the panorama for the front semi-circle will also require shrinking it for the back. Several spatial transformations can

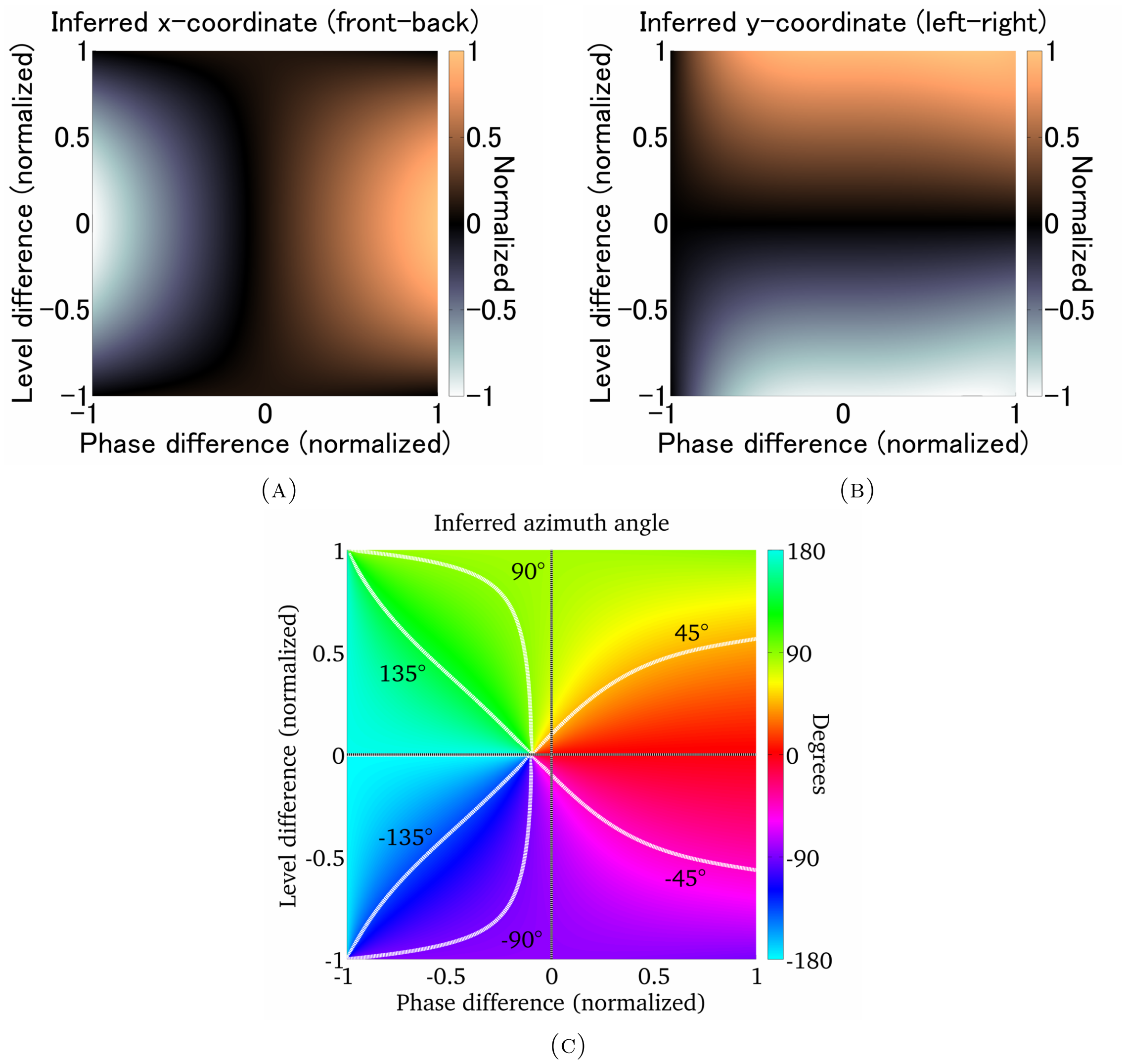


Figure 4. Proposed inversion of a stereo panning law after corrections for sound field stabilization and a spatial warping. Panel (a) shows the inferred values for the front-back coordinate. Panel (b) shows the values for the left-right coordinate. Panel (c) shows the azimuth angles assigned to sources according to their the inter-channel differences. Black lines mark the left-right and front-back axes; white lines are level contours for some representative azimuth angles.

achieve this. In this paper, we propose warping the horizontal plane along the left-right coordinate y while leaving the x coordinate intact. This will result in changes to the inferred azimuth angles, but will not mix the front and back semi-planes.

Following this approach, sound sources in the back semi-plane are forced closed together by reducing the effect of inter-channel level difference on their inferred y-coordinate. The opposite is done for the front semi-plane, where inter-channel level differences should have a stronger impact on the estimated value of y. It must be noted that the correction introduced in Eq. (13) already steers lateral sounds to $\theta = \pm 90°$, where no further correction is possible by changing the y-coordinate. To keep the spatial warping continuous, the expansion of the front semi-plane should focus on sources with small inter-channel level differences and become weaker as this difference becomes larger (as is the case of lateral sources).

The front and back semi-planes can be distinguished by looking at the inter-channel phase difference in Eq. (11). This equation yields positive values for the front and negative ones for the back semi-plane. Therefore, when the inter-channel level difference is small, the corrections to the y-coordinate should have the same sign as the inter-channel phase difference or the uncorrected coordinate x; it should be an odd function. Meanwhile, it should be forced to values close to zero when the inter-channel level difference is large and x is positive. To achieve this, we propose the following spatial warping along y:

$$\hat{y} = y + \left[cx^3y - dx^4y\right] - \left[ex^3y^3 + fx^4y^3\right]. \tag{14}$$

The terms in the first bracket, being linear in y, dictate the transformation for small inter-channel level differences. The terms in the second bracket are relevant only when the uncorrected y is close to ±1 (large inter-channel level differences). The choice of third and fourth powers of x is to ensure that monaural sources will remain unaffected by this spatial warping. Figure 3 shows the value of the correction to y as a function of the inter-channel phase difference for both, small and large inter-channel level differences.

The coefficients c and d in Eq. (14) can be adjusted freely to change the strength of the spatial warping required by a specific panning law. Rather than considering the coefficients independently, it is convenient to notice that the value $c + d$ determines the maximum correction applied to shrink y in the back semi-plane; meanwhile, $c - d$ indicates the largest correction used to expand the front semi-plane. In general, $c + d$ should be close to 1, while $c - d$ should be a fraction of this. Empirical tests show that $c + d = 0.9$ and $c - d = 0.3$ work well with typical stereo sources [24]. This yields the values of $c = 0.6$ and $d = 0.3$.

The terms in the second bracket in Eq. (14) reduce the spatial warping for sound sources on the front and with a large inter-channel level difference. The coefficients e and f should take values that cancel the corrections introduced in the first bracket when x approaches 1. This constrains their values to $e + f = c - d$. On the other hand, the difference $f - e$ can be chosen to moderate the shrinking of the back semi-plane if the inter-channel level difference is small. This value should, nevertheless, be either zero or a small positive fraction. For illustration purposes, this paper will use $f - e = 0.1$, and therefore $e = 0.1$ and $f = 0.2$.

Equations (13) and (14) describe a non-linear coordinate transformation for the inferred x, y and θ shown in Fig. 2. The result of applying this transformation is summarized in Fig. 4. Notice that the corrections of Eq. (13) shifted the singularity away from the center of the plane (which corresponds to a monaural signal). Also, the spatial warping of Eq. (14) ensures that a larger portion of the level-phase plane maps to azimuth angles between 90° and −90°.

4.4. **High-Order Ambisonic encoding of the extrapolated sound field.**

The previous sections describe a way to calculate an azimuth angle $\theta(\omega) = \arctan[\hat{y}(\omega), \hat{x}(\omega)]$ from the inter-channel differences in a stereo source. A single stereo signal can contain multiple sources at different positions, as long as they differ in frequency. The next step is to represent the spatial information along with the original sound information as a high-order Ambisonics (HOA) encoding.

Generating an HOA encoding requires two types of information: time-related and space-related. The time-related information is a monaural signal containing the actual sound. Meanwhile, the space-related information consists of the position from which the sound should be presented. The former is calculated by downmixing the original stereo signal using the following frequency-domain formulas:

$$|\mathrm{M}(\omega)| = \sqrt{|\mathrm{Left}(\omega)|^2 + |\mathrm{Right}(\omega)|^2} \tag{15}$$

$$\mathrm{Ang}\,[\mathrm{M}(\omega)] = \begin{cases} \mathrm{Ang}\,[\mathrm{Right}(\omega)] & \theta < 0 \\ \mathrm{Ang}\,[\mathrm{Left}(\omega)] + \mathrm{Ang}\,[\mathrm{Right}(\omega)] & \theta = 0 \\ \mathrm{Ang}\,[\mathrm{Left}(\omega)] & \theta > 0 \end{cases} \tag{16}$$

These formulas result in the frequency-domain representation of a monaural downmix $\mathrm{M}(\omega)$. They are needed to ensure that sounds presented behind the listener (opposite polarity for the left and right channels) do not cancel, as would occur when calculating the inter-channel average in the time domain.

The space-related information is limited in our proposal to the azimuth angle of incidence $\theta(\omega)$. The lack of distance information means that the resulting sound field can be represented using the encoding for plane waves, shown in Eq. (8). Furthermore, elevation angles are zero for all sources; this reduces the number of non-zero coefficients and simplifies the encoding equation. Full-sphere HOA encodings of order $\mathrm{N}_{\max}$ require $(\mathrm{N}_{\max}+1)^2$ coefficients, each related to a spherical harmonic function. Horizontal HOA encodings, on the other hand, need only $2\mathrm{N}_{\max}+1$ coefficients, associated to simple sine and cosine functions. The encoding equation used in this proposal is:

$$\mathrm{B}_n(\omega) = \begin{cases} -\mathrm{M}(\omega)\sin\left[n\theta(\omega)\right] & n < 0 \\ \mathrm{M}(\omega)\cos\left[n\theta(\omega)\right] & n \geq 0 \end{cases} \tag{17}$$

The first coefficient is, therefore, the monaural downmix of Eqs. (15) and (16): $\mathrm{B}_0(\omega) = \mathrm{M}(\omega)$. First order coefficients $\mathrm{B}_{-1}(\omega)$ and $\mathrm{B}_1(\omega)$ encode the left-right and front-back directions, respectively. Unlike the matrix approach of UHJ encoding, shown in Eq. (9), our proposal derives these coefficients by filtering the monaural signal $\mathrm{M}(\omega)$. This is equivalent to applying a frequency-dependent mixing matrix to the stereo signal. There is no reason to limit the proposed method to first-order Ambisonics, as is the case of the UHJ format. Computing higher orders can improve sound source separation and spatial resolution.

In summary, our proposal computes a high-order Ambisonics encoding by applying a filterbank (space-related information) to a monaural downmix (time-related information). The filterbank is formed by taking the sine and cosine of integer multiples of the inferred azimuth angle $\theta(\omega)$. Empirical tests, shown in the next Section, show that our proposal can produce meaningful data for second-order Ambisonic encodings from a stereo source.

5. **Evaluation.** Our proposal is now evaluated using a test signal synthesized according to the extended panning law of Fig. 1b. The resulting HOA encoding is analyzed using a simple beamforming method in the circular harmonics domain. This allows us to extract the sounds that will be presented from different directions around the listener once the HOA data is decoded and reproduced. The resulting sounds are compared to the original signal at each of the test azimuth angles.

5.1. **Testing conditions.** Our evaluation uses a simple wavelet as test signal. This is generated by applying a Hann window to a pure tone. The window has a length of 8000 samples and contains two oscillations for the pure tone. Four repetitions of the test signal are distributed in a 1-second long stereo recording. The left and right channels are generated by applying the extended panning law of Fig. 1b for the following angles: 90°, −90°, 0°, and 180°. These were chosen as they represent the most problematic cases in panning law inversion (lateral, monaural, and inverted polarity sources). The resulting stereo signal is shown in Fig. 5.

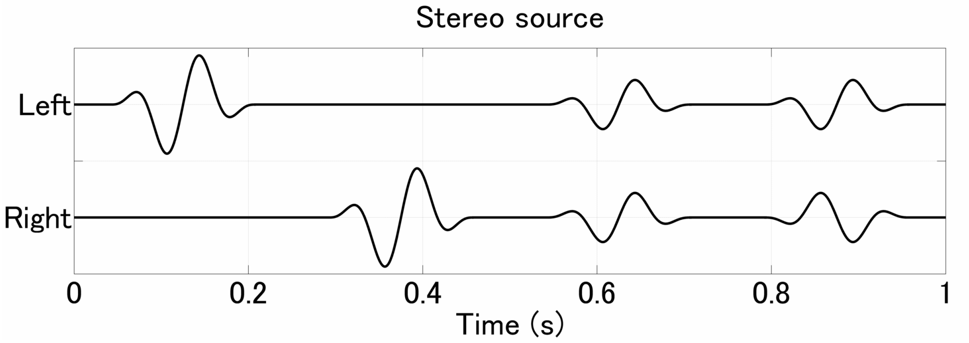


FIGURE 5. Stereo test signal used to evaluate the proposed method. It consists of a pure tone modulated by a Hann window. Four iterations of the signal are used, each panned to a different azimuth angle.

5.2. **Full-surround reproduction results.** The proposed method is applied to generate a HOA encoding from the test signal above. Rather than analyzing the encoded data itself, we opt to look at the results that would be expected after it is decoded and reproduced. To this end, we apply a simple beamforming technique [31]. The beamforming approach allows us to separate the sound sources in the HOA encoding according to their azimuth angles.

Figure 6a shows the results of applying the beamforming method to the first-order Ambisonics encoding generated by our proposal. Sounds are separated for each of the four directions at which the test signal was panned. The results show that the test signal is recovered and its amplitude is highest when the angle of the panning law matches the one in the beamformer. However, the separated signals still show significant values when the test sound was panned away from the beamforming direction. This is not an artifact introduced by our proposal, but rather a limitation due to the poor spatial resolution of first-order Ambisonics.

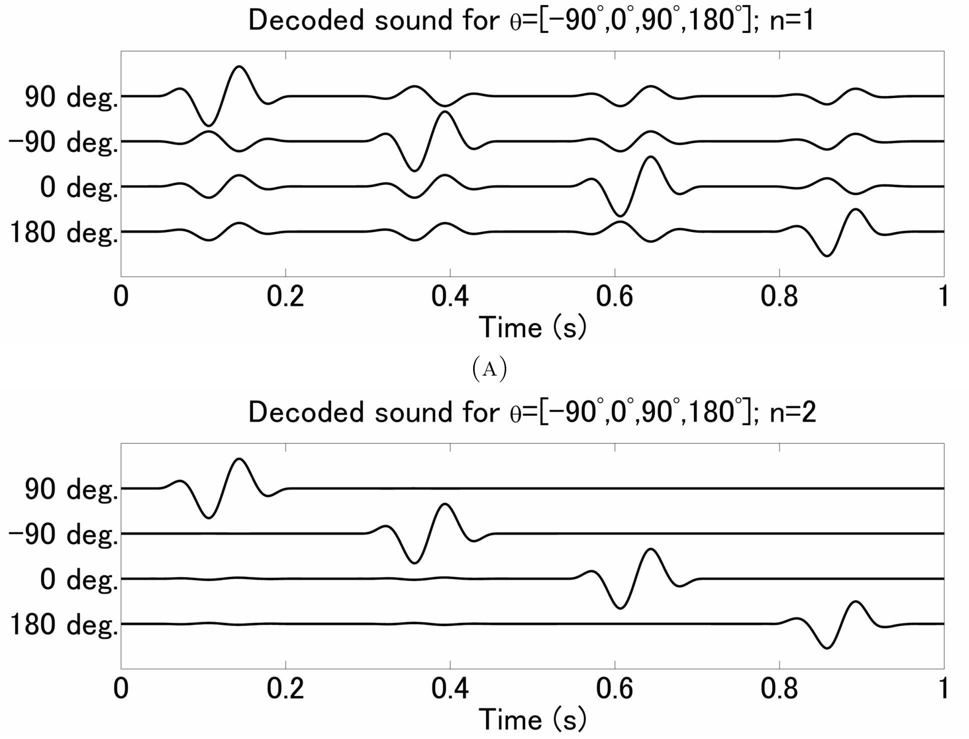
In contrast, Fig. 6b shows the results obtained when the proposed method is used to synthesize a second-order Ambisonics encoding. The higher spatial resolution of this encoding leads to better spatial separation between the test sounds. The amount of leakage between directions is negligible and further increasing the order does not improve the results any further.

These results show that the proposed method is capable of recovering significant spatial information to synthesize a second-order Ambisonics encoding from a stereo signal.

5.3. **Performance at different frequencies.** The proposed method calculates an azimuth angle for each frequency bin. One concern is that complex sound sources may have their frequency components dispersed in space. To ensure this is not the case, we applied our method to sweep test signals panned according to Fig. 1b at several representative angles.

The test signals used sweep all frequencies from low to high linearly within their length of 8192 samples. The azimuth angles obtained by our proposal from the panned sweep signals are plotted in Fig. 7.

The results of our simulations show that the proposed method gives consistent results at low frequencies, with the largest deviations from the target angle being smaller than 5 degrees for up to 3 kHz. At frequencies above 10 kHz, however, the proposed method can yield errors of up to 35 degrees. Nevertheless, the accuracy of 2nd-order HOA at these high frequencies is already low. Therefore, this does not affect the applicability of the



(A)

(B)

FIGURE 6. Results of beamforming on the HOA data synthesized by the proposed method. Beamforming is carried out at the four positions where the test signal was originally panned. Panel (a) shows the results obtained when the proposal is used to generate a first-order encoding. Panel (b) corresponds to the results obtained by synthesizing a second-order encoding.

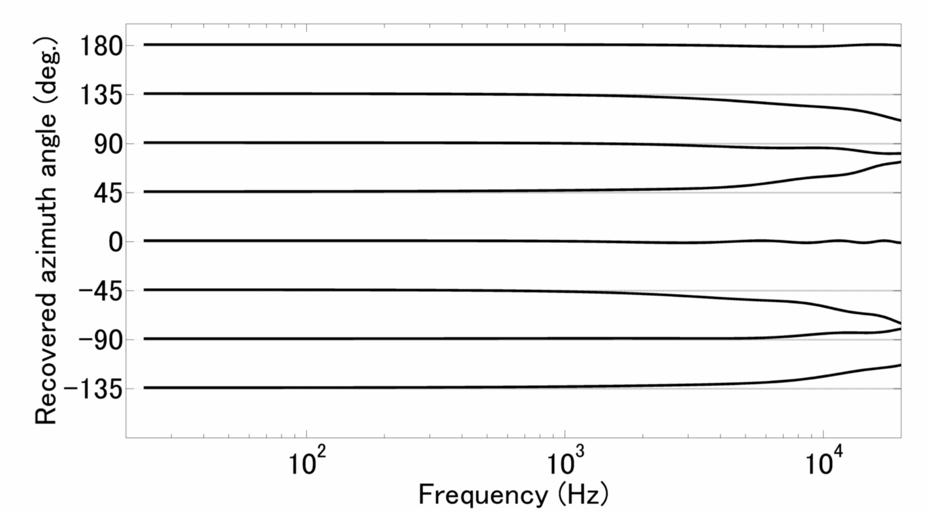


FIGURE 7. Results of applying the proposed method to sweep test signals panned at -135, -90, -45, 45, 90, 135 and 180 degrees. The curves show the recovered azimuth angle as a function of frequency.

proposed method. Our results show that the proposal can be used with complex sounds like speech or, to some extent, musical instruments.

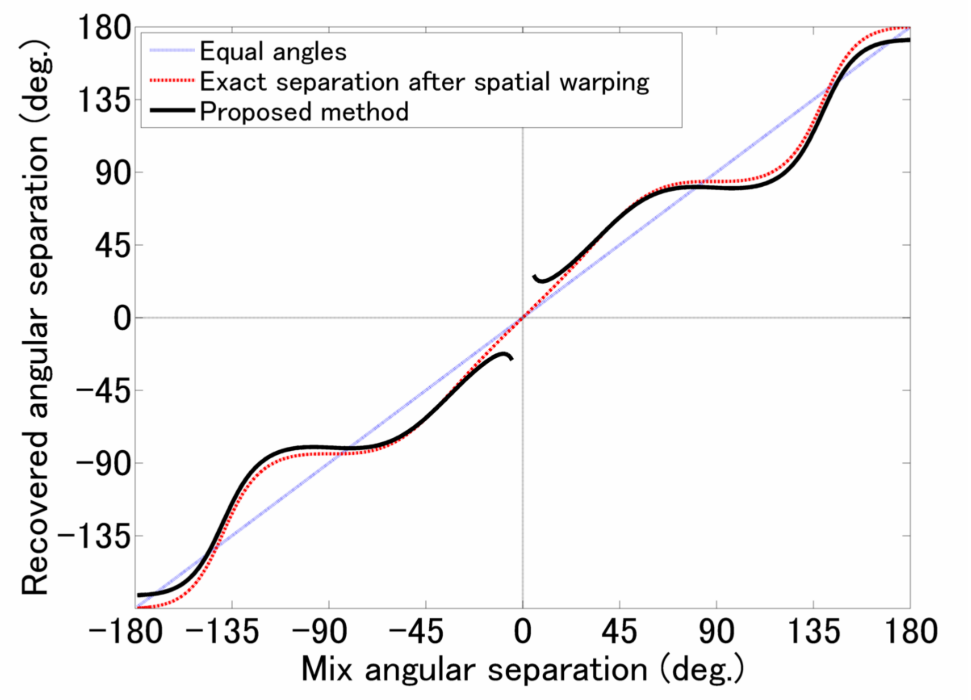


Figure 8. Angular separation between two mixed sound sources positioned in space using the proposed method. One (lower frequency) source was placed at the front with a second one panned to different angles in the horizontal plane.

5.4. **Evaluation for mixed sound signals.** The evaluation results above consider that at any time there is only one sound source represented in the stereo data. Typical contents, however, are the mixture of multiple sound sources. To ensure our proposal can be applied to mixed signals, we performed a simulation using two simultaneous wavelets similar to those of Fig. 5. The frequency of the oscillations in the two test wavelets must be different, otherwise their mixture will be equivalent to the panning of an individual sound. The wavelets used in this test comprise either two or three oscillations.

The two wavelets were panned individually. The lower frequency one was positioned at the front; the higher frequency source was panned to different angles covering the whole 360 degrees panorama in increments on 1 degree. Our proposal was applied to recover the azimuth angles for these two sources. The resulting angular differences are shown in Fig. 8 with relation to the angle used to generate the stereo signal. The angles used to mix the signal and the recovered ones are not expected to be equal. The spatial warping introduced by Eq. (14) must be taken into account. This is shown by the red dashed curve in Fig. 8. Our proposal, shown as the continuous black curve, closely matches this ideal angular separation. The separation could not be calculated reliably when the two sound sources were very close (angular separations below 5 degrees), as shown by the broken curve in Fig. 8.

These results show that the proposed method can recover the angular separation between mixed sound sources in a panned stereo signal. This, in addition to its performance at different frequencies, shown in the previous subsection, ensure that our proposal can be applied to mainstream stereo contents with multiple, complex sound sources.

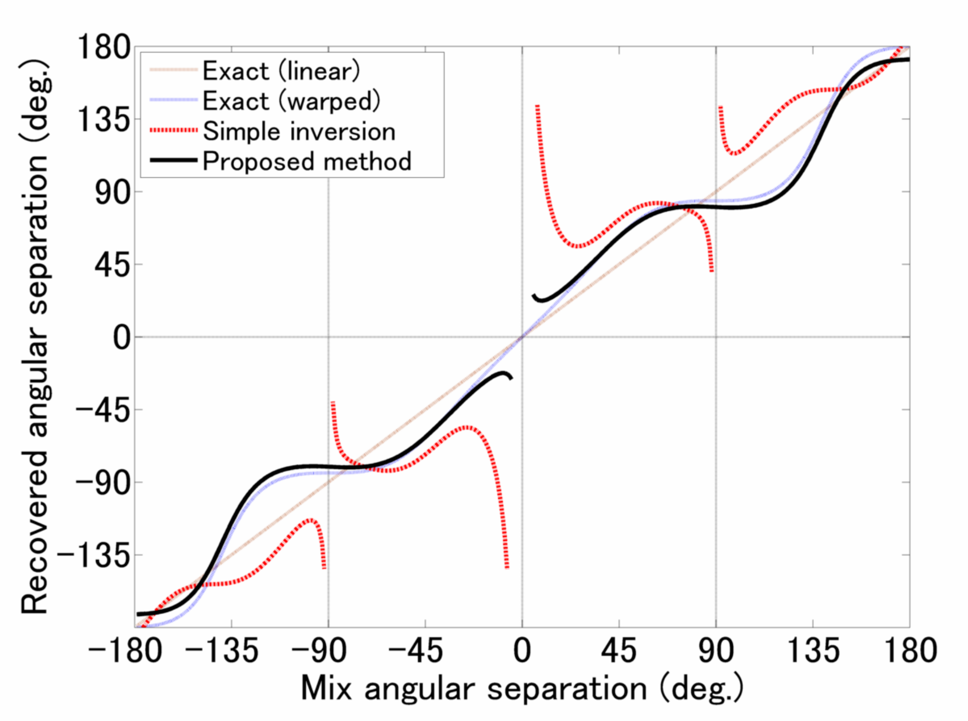


Figure 9. Angular separation between two mixed sound sources positioned in space using the simple inversion of the panning law and the proposed method. The simple inversion shows large deviations from its correspoding exact (linear) curve. Meanwhile, the proposed method closely follows the exact results expected in the warped coordinate system.

5.5. **Effects of the non-linear transformations.** Finally, we compare the performance of our proposal with the simple inversion shown in Fig. 2. The main purpose of the non-linear transformations in Eqs. (13) and (14) is to ensure stability. In a simulation environment, where only round-off errors are expected, single source spatialization is possible without the corrections introduced by our proposal. However, in real-world scenarios, or in the case of multiple mixed sources, simple inversion of the panning law produces inconsistent results.

To verify this, we repeated the evaluation of Subsection 5.4 using the simple inversion of the panning law. The results of this, compared with those of our proposal, are shown in Fig. 9. The simple inversion does not include any spatial warping; therefore, it would ideally match the straight line labeled as ”Exact (linear)” in the figure. The simple inversion is observed to yield poor results when two sound sources are present in the stereo data. Furthermore, it struggles when a sound source is at either 90 or -90 degrees. The large errors are due to the presence of a source at the front, where slight effects from the second source can drastically impact the result of the inversion. The discontinuities to the left and right come from the undefined phase difference of lateral sources, as discussed in Section 4.2. On the other hand, our proposal can reliably estimate the azimuth angle for both sources, closely following the exact curve expected for the warped coordinates it assumes.

6. **Conclusions.**

A method to extract spatial information from conventional stereo sources was presented. The proposed method works by inverting a panning law and introducing non-linear corrections to the inferred sound source positions. The results of our proposal are encoded into high-order Ambisonics data, allowing for reproduction using a variety of loudspeaker arrays or headphones. Results using a simple test signal show that the proposal can extract meaningful information to synthesize a second-order Ambisonics stream. Therefore, its spatial resolution is higher than other methods, such as the Ambisonics UHJ format. Our proposal yields consistent results at all frequencies and, unlike a simple panning law inversion, can reliably render multiple, simultaneous sound sources at their expected positions.

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] 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,* no. 132, 4-page manuscript, June 2013.

[2] 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. of ICAT 2013,* paper no. D5, 2-page manuscript, Dec. 2013.

[3] 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.

[4] 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, Sept. 2014.

[5] 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.

[6] 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, 5-page manuscript, June 2003.

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

[8] 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.

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

[10] 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.

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

[12] 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.

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

[14] 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, 6-page manuscript, March 2012.

[15] 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, 4-page manuscript, May 2010.

[16] J. Trevino, T. Okamoto, Y. Iwaya and Y. Suzuki, Sound field reproduction using Ambisonics and irregular loudspeaker arrays, *IEICE Trans. Fund. Electron. Comm. Comput. Sci.,* vol. 97-A, no. 9, pp. 1832–1839, Sept. 2014.

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

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

[19] N. M. Ngo, M. Unoki, R. Miyauchi and Y. Suzuki, Data Hiding Scheme for Amplitude Modulation Radio Broadcasting Systems, *J. Inf. Hiding and Multimedia Signal Process.,* vol. 5, no. 3, pp. 324–341, July 2014.

[20] K. Sonoda, R. Nishimura and Y. Suzuki, Blind detection of watermarks embedded by periodical phase shifts, *Acoust. Sci. and Tech.,* no. 25, vol. 1, pp. 103–105, 2004.

[21] H. C. Huang and F. C. Chang, Error Resilience for Compressed Sensing with Multiple-Channel Transmission, *J. Inf. Hiding and Multimedia Signal Process.,* vol. 6, no. 5, pp. 847–856, Sep. 2015.

[22] Y. H. Chen and H. C. Huang, A Wavelet-Based Genetic Watermarking Scheme for Stereoscopic Images *Proc. IEEE ICCE-Taiwan 2014,* pp. 113–114, 2014.

[23] Y. H. Chen and H.C. Huang, A Wavelet-Based Image Watermarking Scheme for Stereoscopic Video Frames, *Proc. Intl. Conf. on Intell. Inf. Hiding and Multimedia Signal Process.,* pp. 25–28, 2013.

[24] 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.

[25] 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.

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

[27] 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.

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

[29] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography,* London UK, Academic Press, 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] 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.