OBJECT-BASED STEREO UP-MIXER FOR WAVE FIELD SYNTHESIS BASED ON SPATIAL INFORMATION CLUSTERING

Noriyoshi Kamado, Masayuki Hirata, Hiroshi Saruwatari, Kiyohiro Shikano

Nara Institute of Science and Technology

ABSTRACT

To build an acoustic system that can maintain the localization of sound images included in stereo mixed signals, we propose a new object-based up-mixer that performs sound source separation and sound location estimation. First, in a preliminary experiment, we show the effectiveness of sound location estimation using the proposed up-mixer via objective tests. Next, we evaluate the perception accuracy of sound localization by wave field synthesis using the proposed up-mixer via subjective tests. The results show that the proposed up-mixer provides a good localization of sound images included in stereo mixed signals at several listening positions.

Index Terms— Wave field synthesis, Object-based up-mixer, Perception of sound localization, Subjective test

1. INTRODUCTION

Multi-channel sound field reproduction (SFR) methods are promising for improving the quality and interactivity of acoustic telecommunications. The ultimate aim of SFR is to perfectly reproduce the characteristics of natural hearing over the entire spatial and frequency domains. However, the present listening experience provided by SFR does not have the naturalness of a real face-to-face conversation because recent technological developments have focused on only sound quality.

In recent years, many multichannel sound field reproduction systems for reproducing the characteristics of natural hearing over the entire spatial and frequency domains based on wavefront synthesis have been extensively investigated. *Wave field synthesis* (WFS) [1] is one of the most promising SFR methods, which assumes an anechoic reproduction environment and provides a large listening area with high perceptual reproduction quality for multiple listeners.

Despite the advances in SFR methods as typified by WFS, the availability of a multichannel audio recording method for multichannel SFR has been extremely limited until recently. In particular, in the application of WFS to commercially available multichannel sound contents (mostly *stereo*), to maintain the localization of sound images included in multichannel sound contents in the entire region covered by WFS, the localization information for each sound source is required as a cue for the primary sources generated in WFS. This means that a method of analyzing and decomposing the primary sound sources included in the sound contents, which is called the *object-based up-mixer* method, is indispensable for the total system of multichannel SFR [2]. As WFS systems are not yet widely deployed, object-based up-mixing methods have seldom been discussed in the literature. Therefore, the development of a object-based up-mixer for WFS is a problem requiring urgent attention. In this paper, we propose a new object-based up-mixer of conventional multichannel audio contents for WFS.

We have previously proposed a sound field coding method for multichannel audio content to decrease the data size of transmission [3]. To analyze and decompose the primary source information required for WFS, we utilize this coding method to decompose mixed primary source into each primary sources. We propose a new estimation method for the location of the primary sources using the inverse operation of vector-based amplitude panning (VBAP) [4] for an object-based up-mixer. In addition, we evaluate the effectiveness of the proposed up-mixer by carrying out objective and subjective assessments.

This work was supported by the MIC SCOPE, and JST Core Research of Evolution Science and Technology (CREST), Japan.

2. RELATED STUDIES

2.1. Wave field synthesis

In this section, WFS and VBAP are described theoretically and the equations used for filter calculations in SFR are derived in detail. The geometric configuration and parameters in WFS are depicted in Fig. 1, where $S_{Pv}(\omega)$ and $S_{S(v,n)}(\omega)$ denote the spectra of the vth primary and nth secondary sources, respectively, on the x-y horizontal plane. The spectrum of the nth secondary source, which synthesizes the primary spherical wavefront, $S_{S(v,n)}(\omega,\tau)$, is expressed as [5]

$$\begin{aligned} &S_{S(v,n)}(\omega,\tau) \\ &= S_{Pv}(\omega,\tau)\sqrt{\frac{\mathrm{sign}\left(\zeta(\omega,\tau)\right)k}{2\pi j}}\sqrt{\frac{\zeta(\omega,\tau)}{\zeta(\omega,\tau)-1}} \\ &\quad \frac{\exp\left(\mathrm{sign}\left(\zeta(\omega,\tau)\right)jkr_{PS(v,n)}(\omega,\tau)\right)}{\sqrt{r_{PS(v,n)}(\omega,\tau)}}\cos\left(\theta_{PS(v,n)}(\omega,\tau)\right)\Delta x, \end{aligned} \tag{1}$$

$$\zeta(\omega,\tau) = \frac{y_R}{y_{Pv}(\omega,\tau)}, \tag{2}$$

where j is the imaginary unit, k is the wave number (ω/c), c is the sound velocity, ω is the angular frequency, Δx is the interelement interval among the secondary sources, $r_{PS(v,n)}$ is the distance between the vth primary source and nth secondary source, $\theta_{PS(v,n)}$ is the angle between the y-axis and the line connecting the nth secondary and vth primary sources and y_R is the reference listening distance in WFS. From Eq. (1), it is clear that WFS requires information for each primary source before synthesizing the secondary sound

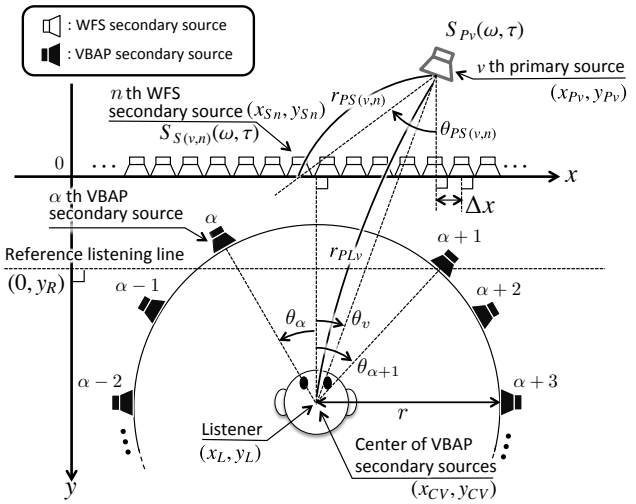

Fig. 1. Relative geometries between listener, vth primary source and secondary sources in VBAP and WFS.

field: the spectra $S_{Pv}(\omega,\tau)$ and the geometries $r_{PS(v,n)}(\omega,\tau)$, $\theta_{PS(v,n)}(\omega,\tau)$ and $y_{Pv}(\omega,\tau)$. Therefore, in the following sections, a method for the simultaneous estimation of this information for each primary source is presented.

2.2. Two-dimensional vector-based amplitude panning

VBAP is a multichannel audio reproduction method that uses amplitude panning. The panning is applied not only to two secondary sources but also to one or two secondary sources adjacent to the primary source out of an arbitrary number of secondary sources distributed on a circle around the listener. In this method, primary sources can appear on a full horizontal circle on the same horizontal plane as the ears of the listener. In Fig. 1, the geometric configuration and parameters in VBAP are also depicted, where θ_α is the angle between the line parallel to the y-axis and intersecting the x-axis at x_L and the line connecting the αth secondary source in VBAP and the point at the center of the listener's head (x_L, y_L). Each loudspeaker in VBAP is located at angle θ_α relative to the listener's position. The αth and $(\alpha+1)$th VBAP loudspeakers are adjacent to the primary source, which is located in the direction $\theta_v(\omega,\tau)$. The index α increases in the clockwise direction. The distance of the VBAP loudspeakers from the center of the listener's head is assumed to be sufficiently large for the wavefront from each loudspeaker at the listener to be approximated by a plane wave.

The panning law for two-dimensional VBAP can be determined as an extension of the *tangent panning law*. To create the vth primary source, the weighting factor of the βth VBAP secondary source $G_{(v,\beta)}(\omega,\tau)$ is obtained from the set of equations

$$G_{(v,\beta)}(\omega,\tau) = \begin{cases} \dfrac{\sin(\theta_{\alpha+1} - \theta_v(\omega,\tau))}{\sin(\theta_{\alpha+1} - \theta_\alpha)} & (\beta = \alpha) \\ \dfrac{\sin(\theta_v(\omega,\tau) - \theta_\alpha)}{\sin(\theta_{\alpha+1} - \theta_\alpha)} & (\beta = \alpha+1) \\ 0 & (otherwise) \end{cases}. \quad (3)$$

3. PROPOSED UP-MIXING METHOD

In this section, we propose a new object-based up-mixing method for WFS along with an estimation method for the location of the primary source using the sound field coding method in our previous work [3].

First, we consider a simple instantaneous linear mixing model in which M microphone input signals are exposed to V primary source output signals. The M-channel time-frequency-series complex vector $\boldsymbol{X}(\omega,\tau) = [X_1(\omega,\tau), \cdots, X_M(\omega,\tau)]^{\mathrm{T}}$ consists of input signals, and superscript T denotes the transposition of a vector or a matrix. Moreover, we define the time-frequency-series complex vector $\boldsymbol{S}(\omega,\tau) = [S_1(\omega,\tau), \cdots, S_V(\omega,\tau)]^{\mathrm{T}}$, which consists of each primary source output, and we define the $M \times V$ mixing matrix $\boldsymbol{H}(\omega,\tau)$, which consists of each spatial transfer function $H_{(m,v)}(\omega,\tau)$ from the vth primary source to the mth microphone. As established by the principle of superposition, the input signal $\boldsymbol{X}(\omega,\tau)$ can be mathematically expressed as follows using the instantaneous linear mixing model:

$$\boldsymbol{X}(\omega,\tau) = \boldsymbol{H}(\omega,\tau)\boldsymbol{S}(\omega,\tau), \quad (4)$$

$$\boldsymbol{H}(\omega,\tau) = \begin{bmatrix} H_{(1,1)}(\omega,\tau) & \cdots & H_{(1,V)}(\omega,\tau) \\ \vdots & \ddots & \vdots \\ H_{(M,1)}(\omega,\tau) & \cdots & H_{(M,V)}(\omega,\tau) \end{bmatrix}. \quad (5)$$

In addition, we assume that the composition of musical instruments does not vary significantly in each musical composition section in conventional audio signals. On the basis of this fact, the proposed method quantizes spatial information in the time-frequency domain, which can be described with Eq. (4) as

$$\boldsymbol{X}(\omega,\tau) \approx \boldsymbol{H}(\omega)\boldsymbol{S}(\omega,\tau). \quad (6)$$

When there are no constraints on Eq. (6), the arbitrariness of $\boldsymbol{H}(\omega)$ and $\boldsymbol{S}(\omega,\tau)$ make it difficult for these terms perform their required role. Therefore, we place a norm constraint on the mixing matrix $\boldsymbol{H}(\omega)$, which is described as

$$\|\boldsymbol{H}_v(\omega)\| = 1, \quad (7)$$

where $\|\cdot\|$ denotes the Frobenius norm. As a result of this constraint, each column vector of $\boldsymbol{H}(\omega)$ represents the spatial transfer information from each primary source to each microphone. Therefore, we refer to the vth column vector $\boldsymbol{H}_v(\omega)$ as the *spatial representative vector* (SRV) hereafter.

From another viewpoint, considering the listener's ability to spatially localize a primary source in a reverberant room with localization dominance owing to the *precedence effect*, it would appear that the listener can only perceive a single primary source in a short bounded time, for example, the frame time of a short-time Fourier transform. Hence, the model in Eq. (6) can be redefined using the vth SRV $\boldsymbol{H}_v(\omega)$ and a function $I(\omega,\tau)$ that describes the index number of the dominant primary source in auditory perception at every time-frequency grid as follows:

$$\boldsymbol{X}(\omega,\tau) \approx \boldsymbol{H}_{I(\omega,\tau)}(\omega)S_{I(\omega,\tau)}(\omega,\tau), \quad (8)$$

where $S_{I(\omega,\tau)}(\omega,\tau)$ denotes the set of spectra of the primary sources that are dominant in auditory perception at every time-frequency grid. From Eq. (8), the problem of primary

source estimation can be rewritten as a problem of the joint estimation of $\boldsymbol{H}(\omega)$, $\boldsymbol{S}(\omega,\tau)$ and $I(\omega,\tau)$ in each musical composition section.

To solve this problem, we consider the joint sequential optimization of $\boldsymbol{H}(\omega)$ and $I(\omega,\tau)$ by minimization of the sin distance $E\left(\boldsymbol{X}(\omega,\tau), \boldsymbol{H}_{I(\omega,\tau)}(\omega)\right)$ between $\boldsymbol{X}(\omega,\tau)$ and $\boldsymbol{S}(\omega,\tau)$ as a first step. The sin distance $E\left(\boldsymbol{X}(\omega,\tau), \boldsymbol{H}_{I(\omega,\tau)}(\omega)\right)$ is defined as follows:

$$E\left(\boldsymbol{X}(\omega,\tau), \boldsymbol{H}_{I(\omega,\tau)}(\omega)\right) = \|\boldsymbol{X}(\omega,\tau)\| \sqrt{1 - \left(\frac{\boldsymbol{X}^{\mathrm{H}}(\omega,\tau)\boldsymbol{H}_{I(\omega,\tau)}(\omega)}{\|\boldsymbol{X}(\omega,\tau)\| \, \|\boldsymbol{H}_{I(\omega,\tau)}(\omega)\|}\right)^2}, \quad (9)$$

where superscript H denotes the complex conjugate transposition of a matrix. Owing to the constraint of Eq. (7), the iterative formulas of $\boldsymbol{H}(\omega)$ and $I(\omega,\tau)$ used updating are given by the following set of equations using k-means clustering in a previously proposed sound field coding method [3]:

$$I^{[k]}(\omega,\tau) = \underset{v}{\operatorname{argmin}} \left(E\left(\boldsymbol{X}(\omega,\tau), \boldsymbol{H}_v^{[k]}(\omega)\right)\right)^2, \quad (10)$$

$$\Theta_v^{[k]} = \{\tau : \boldsymbol{I}^{[k]}(\omega,\tau) = v\}, \quad (11)$$

$$\boldsymbol{H}_v^{[k+1]}(\omega) = \boldsymbol{u}_1^{[k]}(\omega), \quad (12)$$

where k is the iteration number and $\Theta_v^{[k]}$ denotes the vth class of the cluster of the kth iteration. Each frame number τ of the input signal $\boldsymbol{X}(\omega,\tau)$ is assigned to the vth class $\Theta_v^{[k]}$ using the distance function $E\left(\boldsymbol{X}(\omega,\tau), \boldsymbol{H}_{I(\omega,\tau)}(\omega)\right)$. In Eq. (10), $\operatorname{argmin}_v(\cdot)$ denotes the set of values v of the argument that minimize the function $\cdot$ and $\{\cdot\}$ denotes the class that corresponds to a set of $\cdot$. In addition, $\boldsymbol{u}_1^{[k]}(\omega)$ can be derived as the first left-singular vector of matrix $\boldsymbol{T}_v^{[k]}(\omega)$. $\boldsymbol{T}_v^{[k]}(\omega)$ can be described in terms of the input signal $\boldsymbol{X}(\omega,\tau)$ and class $\Theta_v^{[k]}(\omega)$ as

$$\boldsymbol{T}_v^{[k]}(\omega) = \left[\{\boldsymbol{X}(\omega,\tau) : \tau \in \Theta_v^{[k]}(\omega)\}\right]. \quad (13)$$

The single-channel encoded signal is given by

$$\tilde{S}_{I(\omega,\tau)}(\omega,\tau) = \boldsymbol{H}_{I(\omega,\tau)}^{\mathrm{H}}(\omega)\boldsymbol{X}(\omega,\tau). \quad (14)$$

To consider the meaning of each term of Eq. (14), we can interpret the encoded signal $\tilde{S}_{I(\omega,\tau)}(\omega,\tau)$ as the estimated spectra of the primary sources. Therefore, the observed signal of the vth estimated primary source included in the input signal $\boldsymbol{X}(\omega,\tau)$ at the microphones is expressed by applying $\boldsymbol{H}_{I(\omega,\tau)}(\omega)$ to the single channel signal $\tilde{S}_{I(\omega,\tau)}(\omega,\tau)$ under the constraint $I(\omega,\tau) = v$. Therefore, the observed time-frequency signal $\tilde{\boldsymbol{S}}_v(\omega,\tau)$ from the vth primary source is given by

$$\tilde{\boldsymbol{S}}_v(\omega,\tau) = \begin{cases} \tilde{S}_{I(\omega,\tau)}(\omega,\tau)\boldsymbol{H}_{I(\omega,\tau)}(\omega) & (I(\omega,\tau) = v) \\ 0 & (otherwise) \end{cases}. \quad (15)$$

In applying the estimated information for the primary source to WFS, the localization information for each sound source is required as a cue for the primary sources generated in WFS. However, whereas the estimated spectra of the primary sources $\tilde{S}_{I(\omega,\tau)}(\omega,\tau)$ can be used as a substitute for the term $S_{Pv}(\omega,\tau)$ in Eq. (1), the estimated spatial information for the primary sources $\boldsymbol{H}_{I(\omega,\tau)}(\omega)$ cannot be used directly because it is not the location information, $\theta_{PS(v,n)}(\omega,\tau)$ and $r_{PS(v,n)}(\omega,\tau)$, for the primary sources in WFS.

Generally speaking, conventional multichannel audio content assumes a listening environment in which the waves radiated from secondary sources are planar and the position of each secondary source complies with the standard ITU-R recommendation [6]. Hence, it is considered that the localization information for primary sources included in the input signal $\boldsymbol{X}(\omega,\tau)$ is generated by the panning method under a listening environment with ITU-R-recommended secondary source positions. This implies that if we can estimate the weighting factors in the panning law for the primary sources, the directional information for the primary source location can be estimated by the inverse operation of a conventional panning method such as VBAP. On the basis of these assumptions, we focused on the inverse operation of VBAP. From Eq. (3), the inverse operation of VBAP for estimating the direction of primary source $\theta_v(\omega,\tau)$ is obtained as follows:

$$\theta_v(\omega,\tau) = \begin{cases} \cos^{-1}\left(G_\alpha(\omega,\tau)\cos\theta_\alpha + G_{\alpha+1}(\omega,\tau)\cos\theta_{\alpha+1}\right) \\ \sin^{-1}\left(-G_\alpha(\omega,\tau)\sin\theta_\alpha + G_{\alpha+1}(\omega,\tau)\sin\theta_{\alpha+1}\right) \end{cases}. \quad (16)$$

Meanwhile, from Eq. (15), the mth-channel signal observed from the vth primary source is given by

$$\tilde{S}_{(m,v)}(\omega,\tau) = \begin{cases} \tilde{S}_{I(\omega,\tau)}(\omega,\tau)H_{(m,I(\omega,\tau))}(\omega) & (I(\omega,\tau) = v) \\ 0 & (otherwise) \end{cases}. \quad (17)$$

From Eqs. (7), (14) and (15), the estimated mth channel weighting factor $\tilde{G}_{(m,v)}(\omega,\tau)$ of the panning function can be written in terms of the vth SRV as

$$\tilde{G}_{(m,v)}(\omega,\tau) = \begin{cases} \dfrac{|\tilde{S}_{(m,v)}(\omega,\tau)|}{\|\tilde{\boldsymbol{S}}_v(\omega,\tau)\|} = \|H_{(m,I(\omega,\tau))}(\omega)\| & (I(\omega,\tau) = v) \\ 0 & (otherwise) \end{cases}. \quad (18)$$

The estimated direction of the vth primary source $\tilde{\theta}_v(\omega,\tau)$ can be derived by applying the estimated weighting factor $\tilde{G}_{(m,v)}(\omega,\tau)$ to Eq. (17). The radius of the circle consisting of the secondary sources in VBAP r_{PLv} can be assigned on arbitrary value provided it is sufficiently large for the wavefront from each secondary source at the listener to be approximated by a plane wave. From Fig. 1, the estimated direction of the vth primary source $\tilde{\theta}_v(\omega,\tau)$ and the radius of the horizontal circle in VBAP r_{PLv}, the location of the primary source $(\tilde{x}_{Pv}(\omega,\tau), \tilde{y}_{Pv}(\omega,\tau))$ is given by

$$\begin{cases} \tilde{x}_{Pv}(\omega,\tau) &= x_L + r_{PLv}\cos\tilde{\theta}_v(\omega,\tau) \\ \tilde{y}_{Pv}(\omega,\tau) &= y_L + r_{PLv}\sin\tilde{\theta}_v(\omega,\tau) \end{cases}. \quad (19)$$

Finally, from Fig. 1 and (19), the location information for the vth primary source, $r_{PSv}(\omega,\tau)$ and $\theta_{PS(n,v)}(\omega,\tau)$, which are required by the up-mixer for WFS to reproduce the primary sound field, can be derived as

$$\tilde{r}_{PS(v,n)}(\omega,\tau) = \sqrt{(x_{Sn} - \tilde{x}_{Pv}(\omega,\tau))^2 + (y_{Sn} - \tilde{y}_{Pv}(\omega,\tau))^2}, \quad (20)$$

$$\tilde{\theta}_{PS(v,n)}(\omega,\tau) = \arctan\left(\frac{x_{Sn} - \tilde{x}_{Pv}(\omega,\tau)}{y_{Sn} - \tilde{y}_{Pv}(\omega,\tau)}\right). \quad (21)$$

In addition, from Eqs. (15) and (18), the estimated spectrum of the vth primary source $\tilde{S}_{Pv}(\omega, \tau)$ is defined as

$$\tilde{S}_{Pv}(\omega, \tau) = \begin{cases} \tilde{S}_{I(\omega,\tau)}(\omega, \tau) & (I(\omega, \tau) = v) \\ 0 & (otherwise) \end{cases}. \quad (22)$$

4. EVALUATION EXPERIMENTS AND RESULTS

4.1. Experimental conditions of objective assessment

In this section, we evaluate the effectiveness of the proposed method via objective and subjective assessments. In these experiments, we use three monaural audio signals recorded by a professional musician in a low-reverberant sound isolated room, where the instruments are a piano, a woman's vocals, and a guitar. The monaural musical signals used in the assessments are mixed by VBAP into a stereo signal by located at $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_V]$ in the real space, where V is the number of primary sources, which is three in this assessment. The position of each secondary source complies with the standard ITU-R recommendation [6] for the inverse operation of the VBAP given by Eq. (16). The sampling rate is 48 kHz and the number of quantization bits is 16bits.

First, we verify the estimation accuracies of the directions of the primary sources using the proposed method through an objective assessment. The given directions of the instruments from the listener are $\boldsymbol{\theta} = [\theta_{\mathrm{pf}}, \theta_{\mathrm{vo}}, \theta_{\mathrm{gt}}]$ (where pf, vo and gt denote the directions of the piano, woman's vocals and guitar, respectively), their verification direction ranges are $[-25^\circ \le \theta_{\mathrm{pf}} \le -5^\circ, 0^\circ, 15^\circ]$, $[-15^\circ, -10^\circ \le \theta_{\mathrm{vo}} \le 10^\circ, 15^\circ]$ and $[-15^\circ, 0^\circ, 5^\circ \le \theta_{\mathrm{gt}} \le 25^\circ]$, respectively, and the granularity interval of each verification direction is set to 10°. Verification can only be performed one direction in each trial of direction estimation. The number of SRVs is three, which is also the number of instruments included in the input signal, and the SRVs are initialized to the three directions $\boldsymbol{H}_v^{[0]} = [(\sqrt{1-(1/2)^2}, 1/2), \quad (1/\sqrt{2}, 1/\sqrt{2}), \quad (1/2, \sqrt{1-(1/2)^2})]^T$ **[Cond. 1]** or $\boldsymbol{H}_v^{[0]} = [\,(1, 0), (1/\sqrt{2}, 1/\sqrt{2}), (0, 1)]^T$ **[Cond. 2]**.

4.2. Results of objective assessment

[Cond. 1] Figures 2(a) and 2(c) show the results of the objective assessment for condition 1. Figure 2(a) shows the average of the estimated direction results for $\tilde{\theta}_v(\omega, \tau)$ and Fig. 2(c) shows that of the estimation direction error E_{est}. We adopt the estimation direction error of the vth primary source as the evaluation score, defined by

$$E_{\mathrm{est}} = \underset{\omega,\tau}{\mathrm{avg}} \left| \tilde{\theta}_v(\omega, \tau) - \theta_v \right|, \quad (23)$$

where $\mathrm{avg}_{\omega,\tau}$ denotes the averaging function of ω and τ. The horizontal axis of Fig. 2(a) indicates the source direction θ_v and the vertical axis indicates the estimated direction $\tilde{\theta}_v(\omega, \tau)$. The solid diagonal lines in Fig. 2 denote the correct direction in this experiment. In Fig. 2, the white, gray and black circles represent the results for the piano, vocals and guitar, respectively. The vertical axis of Fig. 2(c) indicates E_{est}.

From Figs. 2(a) and 2(c), the estimation accuracies for the piano and guitar, which are located on both sides of the diagonal line, are higher than those for the vocals.

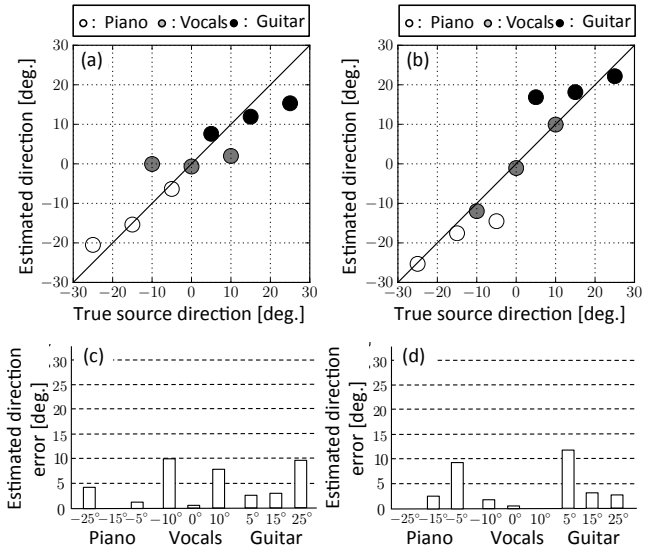

Fig. 2. Results of objective experiments for proposed method. (a) and (c) show the results for condition 1. (b) and (d) show results for condition 2. (a) and (b) show the average estimated direction. (c) and (d) show the average estimated direction error.

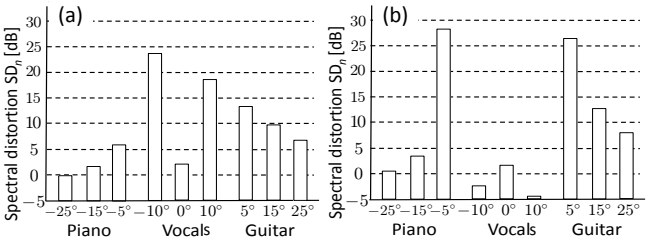

Fig. 3. Spectral distortion of estimated signals obtained by proposed method. (a) shows the results for condition 1 and (b) shows the results for condition 2.

[Cond. 2] Figures 2(b) and 2(d) show the results of the objective assessment for condition 2. Figure 2(b) shows the average of the estimated direction results for $\tilde{\theta}_v(\omega, \tau)$ and Fig. 2(d) shows that of the estimation direction error E_{est}. Whereas the estimation error of the vocal signal is smallest according to Fig. 2(d), we cannot observe the same behavior of the estimation error in Fig. 2(c). Therefore, the results show that the initial setting of the SRV affects the estimated direction $\tilde{\theta}_v(\omega, \tau)$ in the proposed method.

Figure 3 illustrates the results for the spectral distortion (SD) of the estimated primary source. SD for the estimated primary source is defined as

$$\mathrm{SD} = 10 \log_{10} \left(\frac{\sum_{\omega,\tau} (|\tilde{S}_{Pv}(\omega, \tau)| - |S_{Pv}(\omega, \tau)|)^2}{\sum_{\omega,\tau} |S_{Pv}(\omega, \tau)|^2} \right) \ [\mathrm{dB}]. \quad (24)$$

Figure 3(a) shows the results for condition 1 and Fig. 3(b) shows the results for condition 2. The vertical axis in Fig. 3 shows SD. Figures 2 and 3 show the effectiveness of the estimation accuracy of the direction of the primary source using the proposed up-mixer.

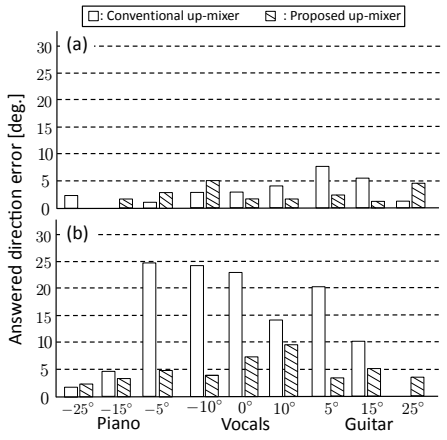

Fig. 4. Average error of answered directions of subjective assessments. (a) shows the results at sweet spot $(x_L, y_L) = (x_{CV}, y_{CV})$ m and (b) shows the results out of sweet spot $(x_L, y_L) = (x_{CV} + 0.5, y_{CV})$ m.

4.3. Experimental conditions of subjective assessment

Next, we also evaluate the sound quality and the ability of the up-mixed primary source localization both at sweet spot and outside the sweet spot in the subjective assessment through the two up-mixer: the conventional channel-based stereo up-mixer [2] and the proposed up-mixer. The experiment was conducted via 34 ch linear array loudspeakers BOSE M-2 for reproduction in a room 3.9 m × 3.9 m with a reverberation time of 300 ms. The distance between the elements of the secondary sources Δx are set to 0.085 m and the center of the x-coordinate of the secondary sources are same coordinate as that of the virtual secondary sources of VBAP x_{CV}. The distance from the listener to primary source r_{Pn} set to 6.0 m. The ear level of the sitting position on the chair for the subject is set on an reproduced horizontal plane $z_L = 1.22$ m and the x-coordinate of the sitting position of the subject x_L is set to the center position of the secondary sources x_{CV} and the that of the y-coordinate y_L is set to $y_{CV} = 2.0$ m. The test subjects are 9 adult males and females with normal audibility. The temperature of measurement room is kept at 20 °C by an air-conditioner.

First, conventional channel-based up-mixer are filtered to the stereo-mixed signal with 10 s. length and present the sources at the reproduced field. Next, proposed up-mixer are filtered to the same signal as previous one. Subjects listen a test sound, and answer where do they perceive the primary sound source image that arises from both up-mixed sounds.

Next, we use the MOS [7] to evaluate the sound quality in typical case of the above experimental results through the comparison of conventional and proposed up-mixer; the following five grades are asked: {5: excellent, 4: good, 3: fair, 2: poor, 1: very poor}. The other experimental conditions, calculation conditions and evaluation criteria are the same as those in **[Cond 1.]** of Section 4.1.

4.4. Results of subjective assessment

To show that the proposed up-mixer does not degrade the reproduction at the sweet-spot, we compared the answered direction error of the proposed up-mixer with that of the conventional up-mixer at sweet-spot in Fig. 4(a) and outside the sweet-spot in Fig. 4(b).

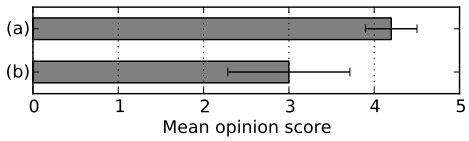

Fig. 5. Subjective evaluation results using MOS to evaluate the sound quality of proposed up-mixer. Error bar shows 95% confidence intervals. (a) shows results for the sound quality at sweet spot $(x_L, y_L) = (x_{CV}, y_{CV})$ m. (b) shows results for the sound quality out of sweet spot $(x_L, y_L) = (x_{CV} + 0.5, y_{CV})$ m.

Figure 5 shows the subjective evaluation results using MOS to evaluate the sound quality of proposed method. These results clarify that the proposed up-mixer provides a good localization of sound images included in stereo mixed signals at several listening positions without excessive sound degradation.

5. CONCLUSION

To build an acoustic system that can maintain the localization of sound images included in stereo mixed signals, we proposed a new object-based up-mixer that performs sound source separation and sound location estimation for WFS using the inverse operation of VBAP. First, in a preliminary experiment, we show the effectiveness of sound location estimation using the proposed up-mixer via objective tests. Next, we evaluate the perception accuracy of sound localization by WFS using the proposed up-mixer via subjective tests. The results show that the proposed up-mixer provides a good localization of sound images included in stereo mixed signals at several listening position without excessive sound degradation.

6. REFERENCES

[1] A. J. Berkhout, D. de Vries, P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, 1993.

[2] M. C. Serrano, *Application of Sound Source Separation Methods to Advanced Spatial Audio Systems*, Ph.D. thesis, Technical University of Valencia, 2009.

[3] H. Saruwatari K. Shikano T. Nomura S. Miyabe, K. Masatoki, "Temporal quantization of spatial information using directional clustering for multichannel audio coding," in *Proc. WASPAA 09*, pp. 261–264. 2009.

[4] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[5] E. N. G. Verheijen, *Sound Reproduction by Wave Field Synthesis*, Ph.D. thesis, Delft University of Technology, 1997.

[6] Rec. ITU-R BS.775-1, *Multichannel Stereophonic Sound System with and without Accompanying Pictures*, 1994.

[7] *Telephone transmission quality, Methods for subjective determination of transmission quality*, ITU-T Recommendation P.800 Annex B, 1996.