Interactive controller for audio object localization based on spatial representative vector operation

Noriyoshi Kamado, Hiroyuki Nawata, Hiroshi Saruwatari and Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma, Nara, 630-0192, Japan

Toshiyuki Nomura
NEC Corporation, Kanagawa, Japan

***Abstract*—In this paper, we propose a new interactive controller for audio object localization based on spatially representative vector operations on a stereo mixed source. First, we developed the interactive controller, which is equipped with a capacitive touchscreen panel so that the listener can intuitively operate audio objects displayed on the touchscreen panel with a touch pen. Next, we assessed the perceptual effects of localization and the sound quality of an audio object after performing individual operations to verify the operation of the interactive controller via a subjective evaluation. The results of the experiments clarify that the interactive controller enables the listener to change the gain and the localization of audio objects without sound degradation if the gain operation is not extreme.**

I. Introduction

In the last decade, with the spread of multichannel audio reproduction systems such as the 5.1 channel surround system, we can easily reproduce sound with a high sense of reality and high quality at home. However, increasing the quantity of signal transmission becomes difficult with increasing number of channels. In addition, the reproduction of music with a high sense of reality has been superseded with the development of sound reproduction systems with a *creative sense of reality* in which a user can operate each sound object individually in the audio signal. As a system for achieving a *creative sense of reality*, Moving Picture Experts Group (MPEG) is currently devising MPEG Spatial Audio Object Coding (MPEG-SAOC) [1], [2], [3], [4]. MPEG-SAOC basically requires separated audio object sources. Such sources are, however, difficult to obtain for common end users because separated sources are not generally supplied via normal music distribution media such as compact discs, in which all sources are mixed. Thus, this codec is difficult to use easily.

In this paper, we first propose a new audio coding framework enabling the localization operation of audio objects (e.g., vocal, guitar, drums) by the temporal quantization of spatial information. An audio object is represented by the centroid vector of spatial clustering algorithm, similarly to k-means [5]. Since the angle of the centroid expresses the direction of the spatial image of the audio object, we can operate the localizaiton of audio objects individually by steering the spatial representative centroid vector of interest.

Next, we introduce an audio object controller developed for interactive sound field reproduction using audio object localization based on spatial representative vector operations. This system allows a listener to operate an initial sound field and build a new virtual sound field by performing sound field operations on a stereo mixed source.

II. System configuration

In this section, we give an overview of the proposed system. Figure 1 shows the configuration of the proposed system. The proposed system consists of three functions: a batch processing encoder, a real-time processing decoder and an interactive audio object localization controller.

First, the stereo mixed signal is inputted into the encoder, and the input signal is analyzed and decomposed to the audio objects contained within, and the encoder outputs are stored as a storage file by a dedicated format container file. The interactive controller reads the storage file and displays the relative positions of the audio objects. Using the touchscreen panel, the listener can control the relative positions via the graphical user interface (GUI). The real-time decoder interprets each listener operation and the operation is reflected immediately in the reproduced sound field. The mathematical principles of the proposed system are described in the following section.

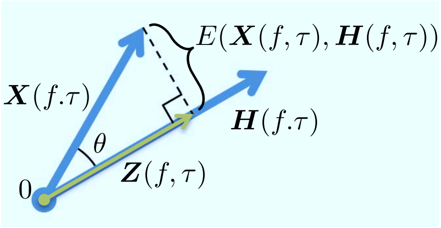

Fig. 2. Relations between the input signal $\boldsymbol{X}$, the decoded signal $\boldsymbol{Z}$, the quantization spatial basis vector $\boldsymbol{H}$ and the quantization error E.

III. Mathematical principles of proposed system

A. Encoder

We assume that the composition of musical instruments does not vary significantly in conventional audio signals. On the basis of this fact, the proposed method quantizes spatial information in the time-frequency domain. Hence, we store spatial localization information in a number of time-invariant basis vectors and extract audio objects efficiently by clustering the mixture of sounds using the basis vectors. Figure 2 shows the quantization error, which is the distance between the M-channel time-series complex vector $\boldsymbol{X}(f,\tau) = [X_1(f,\tau),\cdots,X_M(f,\tau)]^{\mathrm{T}}$, which consists of input signals, and the complex decoded signal vector $\boldsymbol{Z}(f,\tau) = [Z_1(f,\tau),\cdots,Z_M(f,\tau)]^{\mathrm{T}}$. Here we focus on the decoded signals that minimize the quantization error E, which are expressed in terms of the orthogonal projection of $\boldsymbol{X}(f,\tau)$ and supplementary information $\boldsymbol{H}(f,\tau)$, which includes quantized spatial information, as follows:

$$\boldsymbol{H}^{\mathrm{H}}(f,\tau)\boldsymbol{X}(f,\tau) = ||\boldsymbol{H}(f,\tau)||||\boldsymbol{X}(f,\tau)||\cos\theta = ||\boldsymbol{H}(f,\tau)||||\boldsymbol{Z}(f,\tau)||$$
$$\iff \frac{\boldsymbol{H}^{\mathrm{H}}(f,\tau)\boldsymbol{X}(f,\tau)}{||\boldsymbol{H}(f,\tau)||} = ||\boldsymbol{Z}(f,\tau)||, \tag{1}$$

$$\boldsymbol{Z}(f,\tau) = ||\boldsymbol{Z}(f,\tau)||\frac{\boldsymbol{H}(f,\tau)}{||\boldsymbol{H}(f,\tau)||}$$
$$= \frac{\boldsymbol{H}^{\mathrm{H}}(f,\tau)\boldsymbol{X}(f,\tau)}{||\boldsymbol{H}(f,\tau)||}\frac{\boldsymbol{H}(f,\tau)}{||\boldsymbol{H}(f,\tau)||}, \tag{2}$$

where superscript H denotes the complex conjugate transposition of a matrix, $||\cdot||$ denotes the Euclidean norm and $\boldsymbol{H}(f,\tau)$ is described as

$$\boldsymbol{H}(f,\tau) = \boldsymbol{C}_{I(f,\tau)}(f), \tag{3}$$
$$\boldsymbol{C}_n(f) = [C_{(n,1)}(f),\ldots,C_{(n,M)}(f)]^{\mathrm{T}} \quad (n = 1,\ldots,N), \tag{4}$$

where $\boldsymbol{C}_n(f)$ is the nth centroid and the complex basis vector derived from the clustering described below, $I(f,\tau)$ denotes the nth index of the centroid that minimizes the quantization error between $\boldsymbol{X}(f,\tau)$ and $\boldsymbol{Z}(f,\tau)$ at every time-frequency grid in all channels, and N denotes the number of centroids.

The basis vector used in the orthogonal projection is determined by the cosine-distance weighted k-means [5] to minimize the quantization error. We can obtain the resultant single-channel encoded signal $Y(f,\tau)$ using this basis vector.

To minimize the quantization error between input signals and decoded signals, we formulate the magnitude of the compression error. First, we calculate the nth basis vector $\boldsymbol{C}_n(f)$ and the decoded

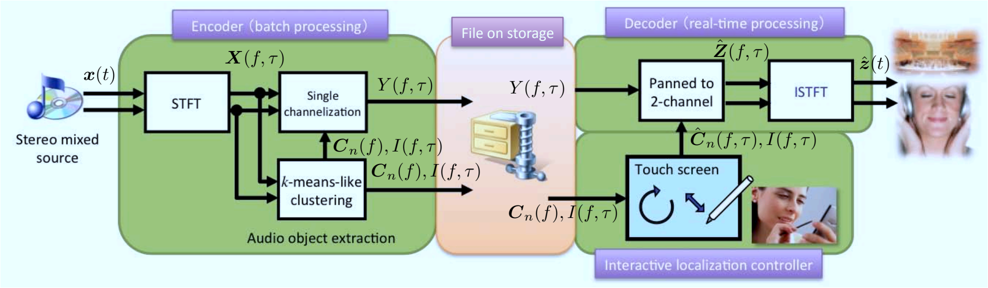
Encoder（batch processing）
File on storage
Decoder（real-time processing）
$\boldsymbol{x}(t)$
Stereo mixed source
STFT
$\boldsymbol{X}(f,\tau)$
Single channelization
$Y(f,\tau)$
$\boldsymbol{C}_n(f), I(f,\tau)$
k-means-like clustering
$\boldsymbol{C}_n(f), I(f,\tau)$
Audio object extraction
$Y(f,\tau)$
$\boldsymbol{C}_n(f), I(f,\tau)$
$\hat{\boldsymbol{Z}}(f,\tau)$
Panned to 2-channel
ISTFT
$\hat{\boldsymbol{z}}(t)$
$\hat{\boldsymbol{C}}_n(f,\tau), I(f,\tau)$
Touch screen
Interactive localization controller

Fig. 1. Configuration of proposed system.

signal vector $\boldsymbol{Z}(f,\tau)$. The quantization error $E(\boldsymbol{X}(f,\tau), \boldsymbol{H}(f,\tau))$ can be expressed in terms of the cosine-distance as

$$\begin{aligned} E(\boldsymbol{X}(f,\tau), \boldsymbol{H}(f,\tau)) &= \|\boldsymbol{X}(f,\tau)\| \sin(\boldsymbol{X}(f,\tau), \boldsymbol{H}(f,\tau)) \\ &= \|\boldsymbol{X}(f,\tau)\| \sqrt{1 - \cos^2(\boldsymbol{X}(f,\tau), \boldsymbol{H}(f,\tau))}, \end{aligned} \tag{5}$$

where $\cos(\boldsymbol{X}(f,\tau), \boldsymbol{H}(f,\tau))$ is the cosine-distance between $\boldsymbol{X}(f,\tau)$ and $\boldsymbol{H}(f,\tau)$, as

$$\cos(\boldsymbol{X}(f,\tau), \boldsymbol{H}(f,\tau)) = \frac{|\boldsymbol{X}^{\mathrm{H}}(f,\tau)\boldsymbol{H}(f,\tau)|}{\|\boldsymbol{X}(f,\tau)\|\|\boldsymbol{H}(f,\tau)\|}. \tag{6}$$

We optimize the basis vector $\boldsymbol{H}(f,\tau)$ so that the total error of the signals, $E_{\mathrm{total},n}(f)$, is minimized. The total error is given by

$$E_{\mathrm{total},n}(f) = \sum_{\tau\in\Theta_n} \Big(E(\boldsymbol{X}(f,\tau), \boldsymbol{H}(f,\tau))\Big)^2, \tag{7}$$

where Θ_n is the nth class of the cluster. Optimization of the basis vector $\boldsymbol{H}(f,\tau)$ minimizing the error is equivalent to the k-means clustering problem for the cosine-distance as follows.

[STEP 1] Prototype centroids are generated as

$$\boldsymbol{C}_n^{[k]}(f) = \left[C_{(n,1)}^{[k]}(f), \ldots, C_{(n,M)}^{[k]}(f)\right]^{\mathrm{T}} \quad (n = 1, \ldots, N), \tag{8}$$

where k is the number of iterations used to update the centroid.

[STEP 2] Each input signal $\boldsymbol{X}(f,\tau)$ is assigned to the nth class Θ_n based on the error between the input signal $\boldsymbol{X}(f,\tau)$ and the centroid vector $\boldsymbol{C}_n(f)$ as follows:

$$I^{[k]}(f,\tau) = \underset{n}{\mathrm{argmin}}\ E(\boldsymbol{X}(f,\tau), \boldsymbol{C}_n^{[k]}(f))^2, \tag{9}$$

$$\Theta_n = \{\tau : I^{[k]}(f,\tau) = n\}, \tag{10}$$

where $\mathrm{argmin}_n \cdot$ denotes the minimization function, $\{\tau\}$ denotes the class that corresponds to a set of τ, and $I^{[k]}(f,\tau)$ is the index function of the kth iteration. The class to which $\boldsymbol{X}(f,\tau)$ belongs is determined by $I^{[k]}(f,\tau)$.

[STEP 3] The optimal basis vector is extracted to minimize the error, as

$$\begin{aligned} \boldsymbol{C}_n^{[k+1]}(f) &= \underset{\boldsymbol{C}_n^{[k]}(f)}{\mathrm{argmin}}\ E_{\mathrm{total},n}(f) \\ &= \underset{\boldsymbol{C}_n^{[k]}(f)}{\mathrm{argmin}} \sum_{\tau\in\Theta_n} \left(E(\boldsymbol{X}(f,\tau), \boldsymbol{C}_n^{[k]}(f))\right)^2 \\ &= \underset{\boldsymbol{C}_n^{[k]}(f)}{\mathrm{argmin}} \sum_{\tau\in\Theta_n} \|\boldsymbol{X}(f,\tau)\|^2 \Big(1 - \cos^2(\boldsymbol{X}(f,\tau), \boldsymbol{C}_n^{[k]}(f))\Big) \\ &= \underset{\boldsymbol{C}_n^{[k]}(f)}{\mathrm{argmin}} \sum_{\tau\in\Theta_n} \|\boldsymbol{X}(f,\tau)\|^2 \left(1 - \frac{|\boldsymbol{X}^{\mathrm{H}}(f,\tau)\boldsymbol{C}_n^{[k]}(f)|^2}{\|\boldsymbol{X}(f,\tau)\|^2 \|\boldsymbol{C}_n^{[k]}(f)\|^2}\right) \\ &= \underset{\boldsymbol{C}_n^{[k]}(f)}{\mathrm{argmin}} \sum_{\tau\in\Theta_n} -|\boldsymbol{X}^{\mathrm{H}}(f,\tau)\boldsymbol{C}_n^{[k]}(f)|^2 \\ &= \underset{\boldsymbol{C}_n^{[k]}(f)}{\mathrm{argmax}} \boldsymbol{C}_n^{[k]^{\mathrm{H}}}(f) \Big(\sum_{\tau\in\Theta_n} \boldsymbol{X}(f,\tau)\boldsymbol{X}^{\mathrm{H}}(f,\tau)\Big) \boldsymbol{C}_n^{[k]}(f). \end{aligned} \tag{11}$$

Owing to the constraint $\|\boldsymbol{C}_n(f)\| = 1$, the maximization problem on the right-hand side of (11) is equivalent to finding the maximum eigenvalue of $\sum_{\tau\in\Theta_n} \boldsymbol{X}(f,\tau)\boldsymbol{X}^{\mathrm{H}}(f,\tau)$. Therefore, the basis vector $\boldsymbol{H}_i(f)$ is derived as the maximum eigenvector of $\sum_{t\in\Theta_n} \boldsymbol{X}(f,\tau)\boldsymbol{X}^{\mathrm{H}}(f,\tau)$.

[STEP 4] If the centroid vector does not change from that obtained by the previous iteration, the optimal vectors are determined to be the basis vector $\boldsymbol{C}_n(f)$. If the centroid vector changes, the algorithm returns to **[STEP 2]** with $k = k + 1$.

From (2), (4) and the constraint $\|\boldsymbol{C}_n(f)\| = 1$, the single-channel encoded signal $Y(f,\tau)$ is obtained as follows:

$$\begin{aligned} \boldsymbol{Z}(f,\tau) =& \frac{\boldsymbol{H}^{\mathrm{H}}(f,\tau)\boldsymbol{X}(f,\tau)}{\|\boldsymbol{H}(f,\tau)\|} \frac{\boldsymbol{H}(f,\tau)}{\|\boldsymbol{H}(f,\tau)\|} \\ =& \left\{\boldsymbol{C}_{I(f,\tau)}^{\mathrm{H}}(f)\boldsymbol{X}(f,\tau)\right\} \boldsymbol{C}_{I(f,\tau)}(f) \\ =& Y(f,\tau)\boldsymbol{C}_{I(f,\tau)}(f), \end{aligned} \tag{12}$$

$$Y(f,\tau) = \boldsymbol{C}_{I(f,\tau)}^{\mathrm{H}}(f)\boldsymbol{X}(f,\tau). \tag{13}$$

B. Decoder

In the decoding process, multichannel decoded signals $\boldsymbol{Z}(f,\tau)$ are produced by the single-channel compressed signal $Y(f,\tau)$ and the basis vector $\boldsymbol{C}_{I(f,\tau)}(f)$ as follows:

$$\boldsymbol{Z}(f,\tau) = Y(f,\tau)\boldsymbol{C}_{I(f,\tau)}(f), \tag{14}$$

and the nth clustered audio object signal is described as follows:

$$\boldsymbol{Z}_n(f,\tau) = \begin{cases} Y(f,\tau)\boldsymbol{C}_{I(f,\tau)}(f) & (I(f,\tau) = n), \\ 0 & (\mathit{otherwise}). \end{cases}$$

In (14), the quantization vector $\boldsymbol{C}_{I(f,\tau)}(f)$ represents the clustered audio object and its angle expresses the spatial image of the audio object. Therefore, we refer to the quantization vector as the *spatial representative vector* (SRV) hereafter. Finally, the decoder converts $\boldsymbol{Z}(f,\tau)$ into the time-domain expression $\boldsymbol{z}(t)$ by an inverse short-time Fourier transformation (ISTFT) and outputs it to a pair of headphones or earphones. The time-domain signal $\boldsymbol{z}(t)$ is described as

$$\boldsymbol{z}(t) = \mathrm{ISTFT}(\boldsymbol{Z}(f,\tau)). \tag{15}$$

C. Localization operation of audio objects

In this section, we describe signal processing for the localization operation of audio objects by the rotation and expansion/contraction of the SRV. The SRV is operated in the following manner:

$$\hat{\boldsymbol{C}}_n(f,\tau) = G_n(f,\tau)\boldsymbol{R}_n(f,\tau)\boldsymbol{C}_n(f), \tag{16}$$

where $\hat{\boldsymbol{C}}_n(f,\tau)$ is the operated SRV, and $G_n(f,\tau)$ and $\boldsymbol{R}_n(f,\tau)$ are filters that control the perceptual distance (gain) and the azimuth of the perceptual sound image of the nth audio object, respectively. $R_n(f,\tau)$ is obtained as the following rotation matrix:

$$\boldsymbol{R}_n(f,\tau) = \begin{bmatrix} \cos(\vartheta_n(f,\tau)) & -\sin(\vartheta_n(f,\tau)) \\ \sin(\vartheta_n(f,\tau)) & \cos(\vartheta_n(f,\tau)) \end{bmatrix}, \tag{17}$$

where $\vartheta_n(f,\tau)$ denotes the angle of rotation for the nth SRV. By filtering the SRV with $G_n(f,\tau)$ and $\boldsymbol{R}_n(f,\tau)$, we can control the localization of the corresponding audio object of interest on the same

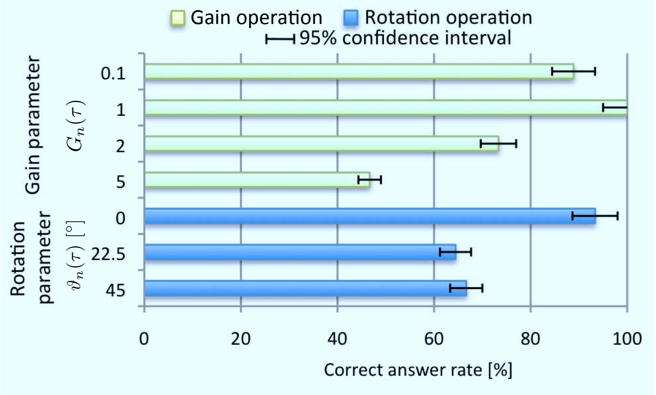

Fig. 3. Subjective evaluation results for operability of gain and rotation operations.

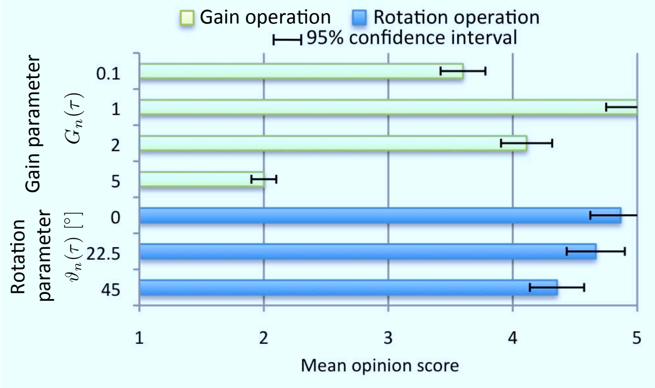

Fig. 4. Subjective evaluation results for sound quality of gain and rotation operations.

horizontal plane as that of the ears of the listener. The operated signal is obtained by restoring the single-channel signal using the SRV as

$$\hat{\boldsymbol{Z}}(f,\tau) = Y(f,\tau)\hat{\boldsymbol{C}}_{I(f,\tau)}(f,\tau). \tag{18}$$

IV. Evaluation of audio object operations

A. Experimental conditions

In this section, we assess the effectiveness of localization and the sound quality of an audio object after performing individual operations to verify the operation of the interactive controller via a subjective evaluation. In this experiment, we use three audio signals recorded by a professional musician and mixed using a real mixing console. The front direction corresponds to 0°, and the clockwise direction is denoted by a positive angle. The audio objects that are mixed are located at {-45°, 0°, 45°} in the real space. The reference signal in the subjective test is the unprocessed original signal. The number of SRVs is three, which is the same as the number of audio objects and the SRVs are set to three directions {-45°, 0°, 45°}, which are the same as the directions of the audio objects. The signal length is 5 s, fast Fourier transform (FFT) points is 8192, the number of shift points is 4096, the sampling rate is 44.1 kHz, the number of quantization bits is 16 bits and the maximum number of updates of the centroid is 50. The values of $G_n(\tau)$ are set to {0.1, 1, 2, 5}. The values of $\vartheta_n(\tau)$ are set to {0, 22.5, 45}. The test subjects are five adult males with normal audibility. We use a pair-comparison method with the reconstructed signal operated by (16) to evaluate the effectiveness of audio object localization. The test subjects evaluate how the perceptual gain of the audio object and the localization of each reconstructed signal changed by selecting one of the three phrases {became large, did not change, became small} and {rotated left, did not change, rotated right}, respectively. We use the mean opinion score [8] (MOS) method to evaluate the sound quality using the following five grades: {5: excellent, 4: good, 3: fair, 2: poor, 1: very poor}.

B. Results of subjective evaluation

Figures 3 and 4 show the correct answer rate and the perceived sound quality of the audio objects after localization control by the proposed method. First, as the gain parameter $G_n(\tau)$ increases, the perceptual audio gain changes. In addition, it is possible to adjust the gain of individual audio objects corresponding to each SRV while maintaining sufficient sound quality except for the case of $G_n(\tau) = 5$.

Next, as the rotation parameter $\vartheta_n(\tau)$ changes, the perceptual azimuth of the audio objects changes. In addition, the rotation operation maintains higher sound quality than the gain operation. The relations among the gains of the audio objects are not severely changed by the rotation operation, therefore the rotation operation can suppress the occurrence of spectrum distortion due to this operation by perceptual masking effects. However, the gain operation destroys these relations; thus, it appears that the gain operation has a harmful effect on the sound quality.

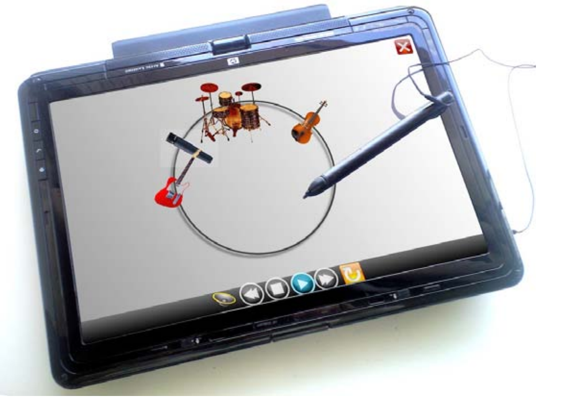

Fig. 5. Prototype of real-time interactive controller.

V. Interactive controller for audio object localization

In this section, we introduce the proposed interactive controller for audio object localization using an SRV operation. Figure 5 shows an overview of the controller. This controller is equipped with a capacitive touchscreen panel, and the listener can intuitively operate every audio object displayed on the touchscreen panel with a touch pen.

Figure 6 shows the geometry of two audio objects and their manipulative variables on the touchscreen panel. There is a subject icon on the center of the touchpanel screen and the front direction of the subject corresponds to 0°, and the clockwise direction is denoted by a positive angle. The default position of a displayed audio object in polar coordinates is described as follows:

$$r_{n_0} = 1, \tag{19}$$

$$\varphi_{n_0} = -2\left\{\left\langle \arg(\boldsymbol{C}_{\mathrm{amp},n}(f))\right\rangle_f - \frac{\pi}{4}\right\}, \tag{20}$$

where $\langle\cdot\rangle_f$ denotes the function that calculates the average degree over frequency f, and $\boldsymbol{C}_{\mathrm{amp},n}(f)$ and the $\arg(\cdot)$ operator are described as

$$\boldsymbol{C}_{\mathrm{amp},n}(f) = [|C_{(n,1)}(f)|, \ldots, |C_{(n,M)}(f)|]^{\mathrm{T}} \quad (n = 1, \ldots, N), \tag{21}$$

$$\arg(\boldsymbol{C}_{\mathrm{amp},n}(f)) = \arctan\left(\frac{C_{(n,L)}(f)}{C_{(n,R)}(f)}\right). \tag{22}$$

In (20), $\pi/4$ is subtracted from the argument of the channel power ratio of the audio object $\boldsymbol{C}_{\mathrm{amp},n}(f)$ so that $\varphi_{n_0} = 0$ when the channel power ratio is 1, that is, $\langle \arg(\boldsymbol{C}_{\mathrm{amp},n}(f))\rangle_f = \pi/4$. The result is then doubled to display the audio object on the touchscreen panel as an instrument icon. This transformation provides display positions consistent with the audio objects. From (16) and Fig. 6, the

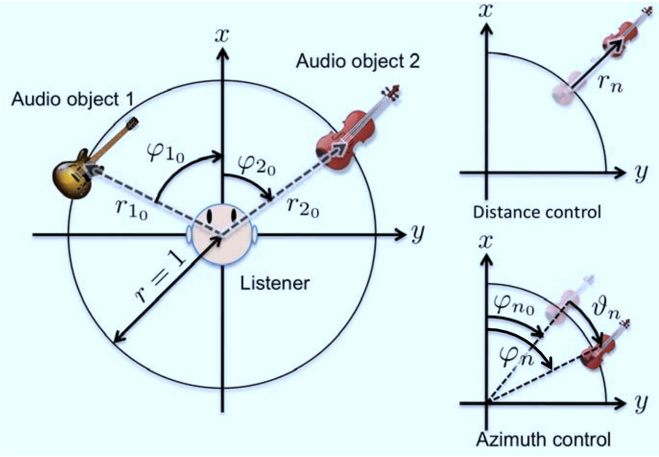

Fig. 6. Geometry of audio objects and their manipulative variables on the touchscreen panel.

relationship between the manipulative variables on the display and the operation on the audio objects is described as follows:

$$G_n(\tau) = \frac{1}{\left\{r_{n_0} + r_n(\tau)\right\}^2} = \frac{1}{\{1 + r_n(\tau)\}^2}, \tag{23}$$

$$\boldsymbol{R}_n(\tau) = \begin{bmatrix} \cos(\frac{\vartheta_n(\tau)}{2}) & -\sin(\frac{\vartheta_n(\tau)}{2}) \\ \sin(\frac{\vartheta_n(\tau)}{2}) & \cos(\frac{\vartheta_n(\tau)}{2}) \end{bmatrix}. \tag{24}$$

The total time required before getting the operation results is 40 to 50 ms, containing the delays of block buffer processing which is equivalent to shift length of STFT (the length is 2048 samples in this implementation), the audio-event handling and the antialiasing digital filter of DA converter in 48 kHz sampling frequency.

VI. Evaluation of interactive controller

A. Experimental conditions

In this section, we assess the sound quality, the operability and the GUI usability of localization of an audio object after performing single operation of the interactive controller via a subjective evaluation. In this experiment, we use three audio signals recorded by a professional musician and mixed using a real mixing console, locating at {-45°, 0°, 45°} in the real space. The number of SRVs is three, which is the same as the number of audio objects and the SRVs are initialized to three directions {-45°, 0°, 45°}. FFT points is 8192, the number of shift points is 4096, the sampling rate is 48 kHz, the number of quantization bits is 16 bits, and the maximum number of updates of the centroid is 50. The user-operable range of $G_n(\tau)$ is set to {1/1.5–1.5 (±1.76 dB), 1/2–2.0 (±3.01 dB), 1/3–3.0 (±4.77 dB)} and the range of $\varphi_n(\tau)$ is set to $\{-\pi/2\text{–}\pi/2\}$. The test subjects are 18 adult males and females with normal audibility. We use the MOS to evaluate the sound quality, the operability and the GUI usability of localization by the interactive controller; the following five grades are asked: {5: excellent, 4: good, 3: fair, 2: poor, 1: very poor}.

B. Results of subjective evaluation

Figure 7 shows the subjective evaluation results for sound quality, the operability and the GUI usability of localization of the audio objects after performing signal operations by the interactive controller. In Fig. 7 (a), as the range of gain parameter increases, the perceptual audio quality decreases, whereas the rotation operation maintains the higher sound quality. Figures 7 (b) and (c) conclude that the qualities on the sound localization and the GUI are almost good and satisfactory to the users because these scores are around 4 regardless of the type of operations. These results clarify that this interactive controller enables the listener to change the gain and the localization of audio objects without sound degradation except the excessive gain operation. As a result of the development of this system, various operations based on this interactive controller and decoder enable the seamless control of audio objects. The usability of the developed system is demonstrated at the following URL:

`http://spalab.naist.jp/aocdemo.html`.

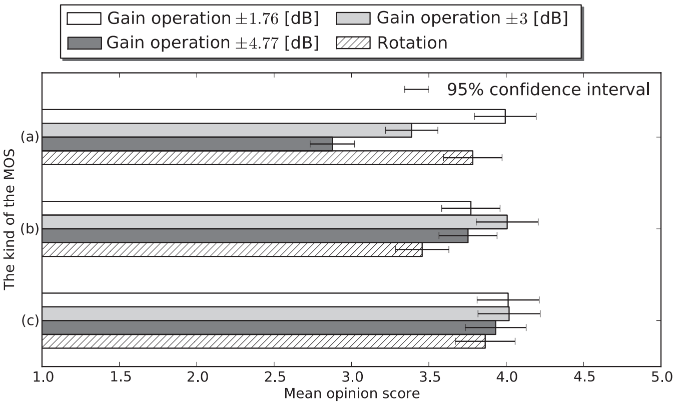

Fig. 7. Subjective evaluation results using MOS to verify the gain and rotation operations of the interactive controller. (a) Results of the sound quality after each operation. (b) Results of the operability of each operation. (c) Results of the GUI usability.

VII. Conclusion

In this paper, we proposed a new interactive controller for audio object localization based on spatially representative vector operations. First, we introduced the configuration of the proposed system. Second, we described the mathematical principles of our system in detail, proving that the spatially representative vectors are closely related to the centroids of cosine-distance weighted *k*-means, and we described the audio object localization control method using the derived vectors. Next, we assessed the effectiveness of localization and the sound quality of an audio object after performing single operations to verify the operation of the interactive controller via a subjective evaluation. The results of the experiments clarified that the interactive controller enables the listener to change the localization of audio objects without sound degradation if the gain operation is not extreme.

VIII. Acknowledgment

This work was partly supported by the MIC Strategic Information and Communications R&D Promotion Programme in Japan.

References

[1] ISO/IEC FCD 23003-2.2, 2009 (Under development).

[2] J. Herre and S. Disch,"New concepts in parametric stereo coding in MPEG-4," *Proc. ICME*, pp. 2–4, 2007.

[3] L. Villemoes, J. Herre, J. Breebaart, G. Hotho, S. Disch, H. Purnhagen, K. Kjörling, "MPEG-surround: the forthcoming ISO standard for spatial audio coding," *Proc. AES*, 2006.

[4] C. Faller and F. Baumgarte, "Binaural cue coding - Part II: schemes and applications," *IEEE Trans. Speech Audio Process.*, Vol.11, pp. 520–531, 2003.

[5] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[6] S. Miyabe, K. Masatoki, H. Saruwatari, K. Shikano, T. Nomura, "Temporal quantization of spatial information using directional clustering for multichannel audio coding," To appear in 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2009), New Paltz, USA, pp. 261–264, 2009.

[7] S. Suzuki, S. Miyabe, N. Kamado, H. Saruwatari, K. Shikano, and T. Nomura, "Audio object individual operation and its application to earphone leakage noise reduction," Proc. Int. Symp. Communications, Control and Signal Processing (ISCCSP2010), Th. 5.6, 2010.

[8] ITU-T Recommendation P.800 Annex B, 1996.