PAPER

A Single-Dimensional Interface for Arranging Multiple Audio Sources in Three-Dimensional Space*

Kento OHTANI[†a)], ***Student Member***, **Kenta NIWA**[††], ***Member***, ***and*** **Kazuya TAKEDA**[†], ***Fellow***

SUMMARY A single-dimensional interface which enables users to obtain diverse localizations of audio sources is proposed. In many conventional interfaces for arranging audio sources, there are multiple arrangement parameters, some of which allow users to control positions of audio sources. However, it is difficult for users who are unfamiliar with these systems to optimize the arrangement parameters since the number of possible settings is huge. We propose a simple, single-dimensional interface for adjusting arrangement parameters, allowing users to sample several diverse audio source arrangements and easily find their preferred auditory localizations. To select subsets of arrangement parameters from all of the possible choices, auditory-localization space vectors (ASVs) are defined to represent the auditory localization of each arrangement parameter. By selecting subsets of ASVs which are approximately orthogonal, we can choose arrangement parameters which will produce diverse auditory localizations. Experimental evaluations were conducted using music composed of three audio sources. Subjective evaluations confirmed that novice users can obtain diverse localizations using the proposed interface.

key words: *single-dimensional interface, auditory localization, head related transfer functions (HRTFs), spatial audio synthesis, matrix dimension reduction*

1. Introduction

Portable audio players and smartphones have allowed us to listen to music anywhere and at any time. Users generally download music from the internet which has been mixed by professionals, and they may adjust the overall frequency characteristics using frequency equalizers [2]. As an advanced music player application concept, several studies have proposed frameworks which allow users to individually vary the auditory localizations of different audio sources, e.g., vocals, guitar and drums, using approaches such as instrument equalizer [3], selective listening point audio [4] and interactive controller [5]. These music player application concepts allow users to re-mix the perceived source locations of music tracks arbitrarily for their own enjoyment. Since the user can change the arrangement of the audio sources as well as his/her listening position arbitrarily, mixing of music can now include manipulation of the positions of the audio sources or the listener. This technology could allow users to produce live music performances by remixing professionally recorded music signals to created new, spatially interesting compositions. This may also make it possible to create virtual concerts which allow users to experience video and corresponding, spatially correct audio of music performances from any selected viewpoint in the audience. When audio source signals and spatial localization effects e.g. head related transfer functions (HRTFs) [6] are assumed to be prepared, this is simply achieved by varying the arrangement parameters composed of position of listener and that of each audio source. Although arbitrary auditory localizations can be obtained through these frameworks, adjustment of the arrangement parameters to achieve the desired localizations may be difficult, especially for users who are unfamiliar with these systems. The issue of ease of arrangement parameter optimization in audio player devices has yet to be addressed by researchers.

In this paper, we propose a single-dimensional interface for arranging audio sources which allows users to switch between selected subsets of arrangement parameters, enabling them to easily obtain diverse auditory localizations. Since each audio source can be placed in an arbitrary position, the number of possible combinations of arrangement parameters is huge, even though the auditory localizations of many of these combinations may be quite similar. To measure the similarity of various auditory localizations, each arrangement parameter can be represented by a vector, which we call an auditory-localization space vector (ASV), and its variance-covariance matrix can then be calculated. If ASVs are similar, the dimensions of the ASV space constructed using the ASVs will be low. By choosing ASVs so that they are orthogonal, we can construct a representative subspace. We can then employ a single dimensional controller to allow listeners to smoothly change between preselected, sets of diverse arrangement parameters, so that the user's auditory localization settings can be drastically varied through the manipulation of a single-dimensional interface.

The rest of this paper is organized as follows. In Sect. 2, we explain an overview of auditory localization control using arrangement parameters. In Sect. 3, we propose a framework for a single-dimensional interface for arranging audio sources and describe its implementation. After experimentally investigating the effectiveness of the proposed interface in Sect. 4, we conclude this paper in Sect. 5.

Manuscript received January 18, 2017.
Manuscript revised May 8, 2017.
Manuscript publicized June 26, 2017.

[†]The authors are with Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8603 Japan.
[††]The author is with NTT Media Intelligence Laboratories, NTT Corporation, Musashino-shi, 180–0012 Japan.

*Part of this work is appeared in [1].

a) E-mail: ohtani.kento@g.sp.m.is.nagoya-u.ac.jp

DOI: 10.1587/transinf.2017EDP7028

2. Auditory Localization Control Using Arrangement Parameters

2.1 Spatial Audio Rendering

Let us assume that a music recording is composed of N audio sources, whose signals are prepared a priori. We use $S_n(\tau, k)$ to denote the n-th audio source signal at time-frame τ and frequency k. The signal $S_n(\tau, k)$ is multiplied by the spatial localization effect functions $G_n^{(\mathrm{L/R})}(k)$, where L/R denotes "left and right" between the n-th source position $\mathbf{p}_n$ and the left and right ears of a listener at $\mathbf{p}_0$. Functions $G_n^{(\mathrm{L/R})}(k)$ could be implemented using binaural effect functions (BEFs) because BEFs include effective cues for 3-dimensional localization. By synthesizing those signals, binaural signals $Y^{(\mathrm{L/R})}(\tau, k)$ are generated:

$$Y^{(\mathrm{L})}(\tau, k) = \sum_{n=1}^{N} G_n^{(\mathrm{L})}(k) S_n(\tau, k), \tag{1}$$

$$Y^{(\mathrm{R})}(\tau, k) = \sum_{n=1}^{N} G_n^{(\mathrm{R})}(k) S_n(\tau, k), \tag{2}$$

Since audio sources are generally placed in the horizontal plane, we assumed that $\mathbf{p}_n$ $(n = 0, \ldots, N)$ is defined in the two-dimensional space.

The BEFs are approximately modeled using measured HRTFs like [7], as in [8]. Since reverberations can be ignored when HRTFs are measured in anechoic chambers, $G_n^{(\mathrm{L/R})}(k)$ would be dependent on the relative position of an audio source between $\mathbf{p}_0$ and $\mathbf{p}_n$. When that relative position is represented by relative distance r_n and relative horizontal direction θ_n, $G_n^{(\mathrm{L/R})}(k)$ can be modeled as a function of r_n and θ_n as:

$$G_n^{(\mathrm{L/R})}(k) = G^{(\mathrm{L/R})}(k, r_n, \theta_n). \tag{3}$$

In many HRTF datasets, measurement was conducted by discretely placing a loudspeaker at a constant distance from a head and torso simulator (HATS) as in [7]. To generate $G^{(\mathrm{L/R})}(k, r_n, \theta_n)$ at an arbitrary direction and distance, we modeled it approximately by multiplying the distance attenuation by linearly angular-interpolated HRTFs as in [8],which can be represented as follows:

$$G^{(\mathrm{L/R})}(k, r_n, \theta_n) = W(k, r_n, r_m) \bar{H}^{(\mathrm{L/R})}(k, r_m, \theta_n) \tag{4}$$

where $W(k_n, r_n, r_m)$ and $\bar{H}^{(\mathrm{L/R})}(k, r_m, \theta_n)$ denote the distance attenuation and linearly angular-interpolated HRTFs, respectively. r_m represents the distances between the sound source and the HATS during HRTF measurement. As an implementation of $W(k, r_n, r_m)$, it is calculated so that the gain is inversely proportional to the relative distance:

$$W(k, r_n, r_m) = \frac{r_m}{r_n} \exp\left(j\omega_k \frac{(r_m - r_n)}{c} \right), \tag{5}$$

where ω_k and c denote the angular frequency of the k-th bin and sound velocity, respectively. Although it would be difficult to accurately obtain auditory localizations when audio sources are positioned close to the listener, it would be possible to roughly estimate the distance perspective using only distance attenuation.

2.2 Arrangement Parameters

Since the BEFs are modeled as a function of relative distance/direction, auditory localization would also be dependent on subsets of the relative distances/directions of the N audio sources. In this paper, set of relative distances/directions of the N audio sources is defined as an *arrangement parameter* as:

$$\Psi_m = \{r_{m,1}, \ldots, r_{m,N}, \theta_{m,0}, \ldots, \theta_{m,N}\} \ (m = 1, \ldots, M) \tag{6}$$

where M denotes the total number of possible arrangement parameters. We assumed that the acoustic field is discretely quantized by a 2-D horizontal grid, and that the cross-points of the grid are candidates for placement of the N audio sources. We define Δ as the grid interval length. The number of possible grid intersections for each x/y coordinate is denoted by L_x and L_y, respectively. Since $M = (L_x \times L_y)^N$ would be a huge number, it would be difficult for users to find their preferred arrangement parameters from among so many possible candidates. Therefore, we aim to construct a single-dimensional interface for arranging audio sources which will allow users to switch between selected, diverse arrangement parameters, allowing them to obtain a wide range of auditory localizations by manipulating a simple interface. Since the auditory localizations of many of the possible arrangement parameter combinations would be perceived as similar by the user, as shown in Fig. 1, the auditory localization space can be represented by a narrowed field of arrangement parameter space.

Under this assumption, a single-dimensional interface to switch between J kinds of arrangement parameters $\{\Psi_1, \ldots, \Psi_J\}$ may allow users to easily survey a diverse range of auditory localizations.

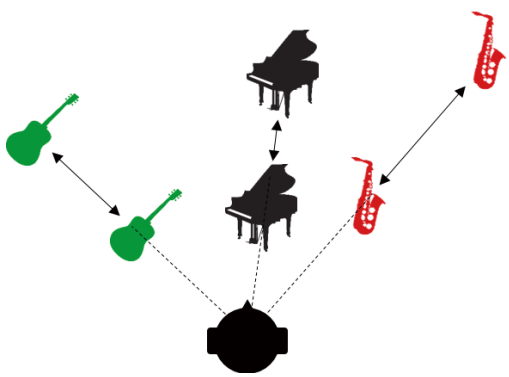

Fig. 1 An example in which a listener perceives similar auditory localizations even when the arrangement parameters are changed. The auditory localizations would not be effected by such variations because the left/right binaural signal power ratio will remain unchanged.

3. Proposed Single-Dimensional Interface for Arranging Audio Sources

3.1 Auditory-Localization Space Vector (ASV)

To model an auditory localization with N audio sources, an auditory-localization space vector (ASV) is defined as:

$$\begin{aligned}\mathbf{f} = [&G_1'^{(\mathrm{L})}(1), \cdots, G_1'^{(\mathrm{R})}(1), \cdots, G_1'^{(\mathrm{R})}(K), \cdots, \\ &G_n'^{(\mathrm{L})}(1), \cdots, G_n'^{(\mathrm{R})}(1), \cdots, G_n'^{(\mathrm{R})}(K), \cdots, \\ &G_N'^{(\mathrm{L})}(1), \cdots, G_N'^{(\mathrm{R})}(1), \cdots, G_N'^{(\mathrm{R})}(K)],\end{aligned} \tag{7}$$

where K represents the number of frequency bins. The auditory characteristics of ASVs can be represented by the magnitude of the response of BEFs:

$$G_n'^{(\mathrm{L/R})}(k) = \left|G_n^{(\mathrm{L/R})}(k)\right|^2. \tag{8}$$

Although inter-aural phase differences in the low frequency range may provide informative cues for localization, they are omitted in this paper.

Other implementation examples used in experiments are explained in Sect. 3.3. ASVs could be represented as a function of an arrangement parameter as:

$$\mathbf{f}_m = \mathbf{f}(r_{m,1}, \cdots, r_{m,N}, \theta_{m,1}, \cdots, \theta_{m,N}) = \mathbf{f}(\Psi_m). \tag{9}$$

Since the combined number of possible arrangement parameters is M, an ASV space is composed of M ASVs, which can be represented as:

$$\mathbf{F} = [\mathbf{f}(\Psi_1), \cdots, \mathbf{f}(\Psi_M)]. \tag{10}$$

3.2 Representative Subspace Extraction from ASV Space

To construct a simple interface for arranging audio sources so that each user can obtain a diverse range of auditory localizations, several representative ASVs need to be extracted from the ASV space. Although M is a huge number, there are many subsets of arrangement parameters whose auditory localization characteristics are very similar to each other as shown in Fig. 1. Even if the arrangement parameters are varied, there are subsets of arrangements where the volume ratio among the audio sources is invariant. When the received volume is restricted, the auditory localizations of these different arrangements would be perceived as being identical.

Our basic method for extracting representative subspaces from the ASV space is to select J ($\ll M$) types of arrangement parameters which will line up in approximately the following pattern:

$$\bar{\mathbf{f}}_{\sigma(1)} \perp \bar{\mathbf{f}}_{\sigma(2)} \perp, \ldots, \perp \bar{\mathbf{f}}_{\sigma(J)}, \tag{11}$$

where $\sigma(1), \ldots, \sigma(J)$ are the indexes of the selected ASVs and $\bar{\mathbf{f}}_{\sigma(n)}$ is a normalized ASV, which is calculated as follows:

$$\bar{\mathbf{f}}_{\sigma(n)} = \frac{\bar{\mathbf{f}}_{\sigma(n)}}{\| \bar{\mathbf{f}}_{\sigma(n)} \|_2}. \tag{12}$$

where $\| \cdot \|_2$ is the L2 norm calculation. This normalization operation is conducted not to distinguish between the types of situations shown in Fig. 1, since the auditory localizations of the J selected types of arrangement parameters are unique. Thus, each user would obtain diverse auditory localizations by manipulating the proposed interface.

The extracted ASV subspace is denoted by:

$$\mathbf{T}_J = [\bar{\mathbf{f}}_{\sigma(1)}, \ldots, \bar{\mathbf{f}}_{\sigma(J)}]. \tag{13}$$

To measure similarity among the selected J ASVs, diagonality of the variance-covariance matrix of the extracted subspace matrix $\mathcal{D}(\mathbf{A}_J)$ is utilized as follows:

$$\mathcal{D}(\mathbf{A}_J) = \frac{\mathrm{trace}(\mathbf{A}_J)}{\sum_{i=1}^{D}\sum_{j=1}^{D} \mathbf{A}_J(i,j)}, \tag{14}$$

where $D = 2 \times N \times K$ represents the dimensions of the ASVs. Thus:

$$\mathbf{A}_J = \mathbf{T}_J^{\mathrm{T}} \mathbf{T}_J. \tag{15}$$

Diagonality $\mathcal{D}(\mathbf{A}_J)$ increases as differences in the characteristics of the J ASVs increase. Thus, our goal to extract representative subspaces can be achieved by selecting our ASVs so that $\mathcal{D}(\mathbf{A}_J)$ is increased. This problem is similar to subspace extraction when using principal component analysis (PCA). However, our problem would not be solved by simple applying PCA because the relationships between the calculated eigenvectors and the arrangement parameters cannot be described with a simple function. Thus, we need to construct an algorithm which will extract representative subspaces by sequentially selecting ASVs so as to increase $\mathcal{D}(\mathbf{A}_J)$. The details of this process are explained in Sect. 3.3.

3.3 Algorithm Implementation

Since there is no specific method for ASV component calculation, we used the three following implementation methods. An experimental comparison of these methods is described in Sect. 4.

1) Magnitude response of BEFs (ASV1)

The magnitude of the response of BEFs calculated in Eq. (8) is utilized as:

$$G_*^{(\mathrm{L/R})}(k, r_n, \theta_n) = G'^{(\mathrm{L/R})}(k, r_n, \theta_n). \tag{16}$$

2) Auditory-filter weighted BEFs (ASV2)

In order to take into account the auditory characteristics of humans, the auditory filter in [9] is multiplied by the magnitude of the response of the PSD of the BEFs as follows:

$$G_*^{(\mathrm{L/R})}(k', r_n, \theta_n) = \sum_k O(k', k) G'^{(\mathrm{L/R})}(k, r_n, \theta_n), \tag{17}$$

where $O(k', k)$ $(k' = 1, \ldots, K')$ denotes the auditory-filter weight between the k'-th auditory filter and the k-th frequency bin, and K' $(< K)$ represents the number of auditory filter-banks.

3) Source power spectrum weighted BEFs (ASV3)

The power spectrum of each audio source is weighted separately because power spectra of the signals of different musical instruments have distinct characteristics. When the power spectrum of the n-th audio source is denoted by $\bar{S}_n(k)$, ASV3 is calculated as follows:

$$G_*^{(\mathrm{L/R})}(k, r_n, \theta_n) = G'^{(\mathrm{L/R})}(k, r_n, \theta_n)\bar{S}_n(k). \quad (18)$$

As mentioned in Sect. 3.2, the calculated ASVs are normalized as shown in Eq. (12).

An algorithm for sequentially selecting J ASVs so that they are approximately orthogonal is developed as follows. First, two ASVs are selected from among the ${}_M\mathrm{C}_2$ available arrangement parameter combinations so as to maximize diagonality $D(\mathbf{A}_2)$. Next, one of the remaining ASV candidates is sequentially inserted into $\mathbf{T}_j$ after its diagonality is calculated as follows:

$$\mathbf{T}_j = [\mathbf{f}_{\sigma(1)}, \mathbf{f}_{\sigma(2)}, \ldots, \mathbf{f}_{\sigma(j)}]. \quad (19)$$

$$\sigma(j) \leftarrow \arg\max\{D(\mathbf{A}_j)\}, \quad (20)$$

where $j-1$ ASVs are assumed to be selected in advance. This process is sequentially continued until J kinds of ASVs are selected.

To determine the dimension of representative subspace, i.e. J, may be a difficult problem. As a pre-investigation to determine J, eigenvalue analysis of $\mathbf{A}_M$ was conducted. Although the details of simulating conditions are explained in Sect. 4, $N = 3$ audio sources are assumed to be placed in 2-D horizontal grid and the number of its cross points is 81 $(= 9 \times 9)$, i.e., $M = 81^3 = 531{,}441$.

Figure 2 shows the relationships between the number of eigenvalues (top twenty only) and the cumulative contribution ratio. The ASVs for this pre-investigation were calculated by applying ASV1. Since the sum of the first ten eigenvectors covers over 95% of the total sum of all of the eigenvalues, we found that the ASV space could be well represented using $J = 10$ orthogonal eigenvectors. Although the ASVs selected with our algorithm were not perfectly orthogonal to each other, they are approximately orthogonal. Note that J should be the number of dominant ASVs needed to cover the ASV space, and that the optimal value for J would vary in relation to the number of audio sources.

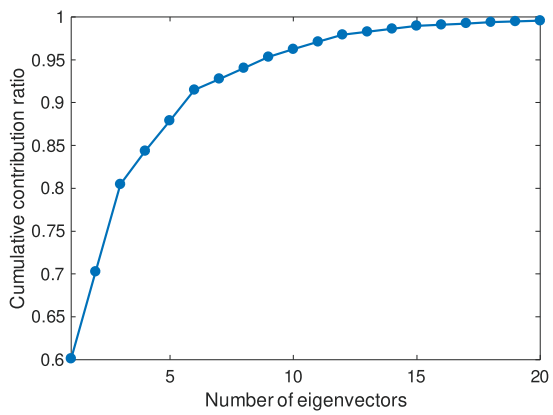

Fig. 2 Pre-investigation of relationships between number of eigenvalues and cumulative contribution ratio of the top twenty eigenvectors. This analysis confirms that the cumulative contribution ratio reaches over 95% when ten or more appropriate eigenvectors are utilized. Note that cumulative contribution ratio of the top twenty eigenvectors is 0.996.

In order to smoothly change between the selected arrangement parameters $\sigma(1), \ldots, \sigma(J)$ by manipulating a single-dimensional interface, a permutation is conducted to avoid drastically varying the auditory localization. The Euclidian distances between audio sources for all possible pairs of J arrangement parameters are calculated as shown in Eq. (21), and they are sorted as $\sigma'(1), \ldots, \sigma'(J)$ to minimize the total distance of N moving audio source paths as shown in Eq. (22). The distance between the $\sigma(i)$-th and the $\sigma(j)$-th arrangement parameters is measured as follows:

$$d(\sigma(i), \sigma(j)) = \sum_{n=1}^{N} \| \mathbf{p}_{\sigma(i),n} - \mathbf{p}_{\sigma(j),n} \|_2, \quad (21)$$

$$\sigma'(1), \ldots, \sigma'(J) = \underset{\sigma(1),\ldots,\sigma(J)}{\arg\min} \sum_{j=1}^{J-1} d(\sigma(j), \sigma(j+1)) \quad (22)$$

To allow the user to smoothly change between auditory localizations, the positions of $\mathbf{p}_{\sigma'(i),n}$ and $\mathbf{p}_{\sigma'(j),n}$ are interpolated using a cubic spline function [10].

4. Experiments

To investigate whether diverse auditory localization can be obtained with the proposed single-dimensional interface, several evaluations were conducted.

4.1 Experimental Setup

We placed audio sources at the intersections of a 9×9 horizontal grid. The range of x/y-coordinates is, respectively, from -3 m to 3 m and from 0 m to 6 m, and $\Delta_{x/y}$ is 0.66 m. Using an HRTF database [7] with the specifications shown in Table 1, ASVs were calculated using the three methods described in Sect. 3.3 (ASV1, ASV2 and ASV3). Two songs composed of three audio sources were downloaded from [11], which contained audio sources listed in Table 2. Although there is a vocal track included in "song 1", there is no vocal in "song 2". From $M = (9 \times 9)^3 = 531{,}441$ possible arrangement parameters, a subset of arrangement parameters in which the sound source positions were extremely

Table 1 HRTF measurement conditions.

Sampling frequency	48 kHz
HATS	B&K 4128
Microphones	Sony ECM-77B
Reverberation time	0.15 s
HRTF angular interval	5°
HRTF length	10.7 ms (512 pt.)

Table 2 Music songs used in experiments.

ID	Title	Artist	Audio sources		
song 1	I'm Alright	Angels In Amplifiers	vocal	drums	others (acoustic/electric guitars)
song 2	Milk Cow Blues	Angela Thomas Wade	bass	drums	others (acoustic guitar, fiddle and piano)

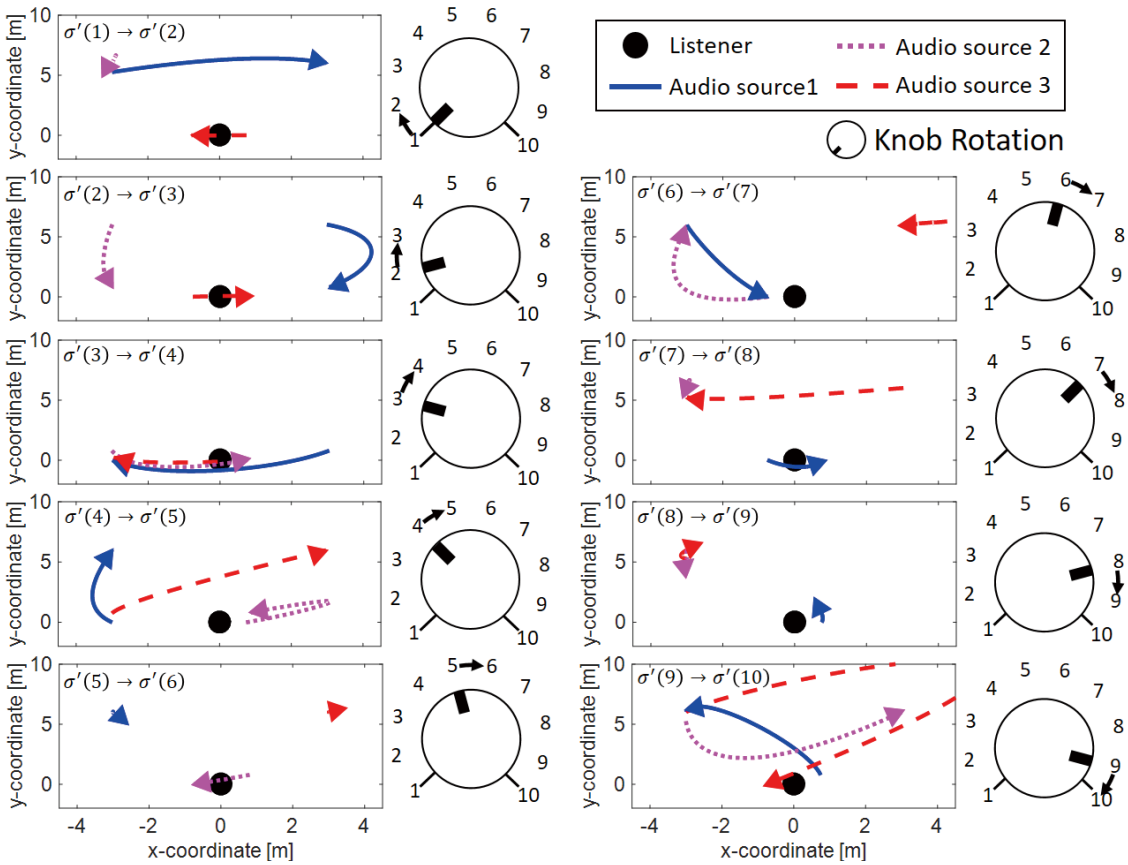

Fig. 3 Audio source locations corresponding to object arrangement parameters selected at various user interface positions. The black dot denotes the fixed position of the listener, while other marks denote positions of audio sources $\mathbf{p}_{m,n}$. Trajectories denote paths for each audio source between adjacent arrangement parameters.

asymmetric, i.e., all of the audio sources were on the right or left side of the listener, were eliminated. After this elimination of unnatural arrangement parameters, 334,368 sets of arrangement parameters remained. $J = 10$ were selected using our proposed algorithm as described in Sect 3.

The selected arrangement parameters and their trajectories in response to manipulation of the single-dimensional interface are shown in Fig. 3. The black dot in Fig. 3 denotes the listener's fixed position, while the other marks denote the relative positions of the audio sources $\mathbf{p}_{m,n}$. The starting points of each line denote $\mathbf{p}_{\sigma'(i),n}$ and the tip of the arrows denotes $\mathbf{p}_{\sigma'(i+1),n}$. By turning the control knob, the user can smoothly change between the J arrangement parameters, moving each audio source as shown in Fig. 3. Because cubic spline interpolation was applied, the trajectories for each sound source were curved and go beyond the defined range in some cases. These results confirm that each audio source moves independently across a wide range of localizations. Trajectories of the audio sources for each of the three kinds of ASVs are shown in Fig. 4. However, the localizations obtained using these trajectories are difficult to evaluate from these figures, therefore, objective and subjective evaluations were conducted which will be explained in the following sections.

4.2 Objective Evaluations of Extracted Representative Subspaces

To investigate whether approximately orthogonal ASVs were actually selected when using the proposed algorithm, their diagonality $\mathcal{D}(\mathbf{A}_J)$ was calculated using Eq. (14). Using the three ASV implementations, representative subspaces and their diagonalities were calculated. As a comparison, arrangement parameters were randomly selected 1,000 times and the results were averaged. To adjust the experimental condition, $J = 10$ arrangement parameters were randomly selected in this case. Figure 5 shows the diagonality of the ASVs selected using each calculation method. For

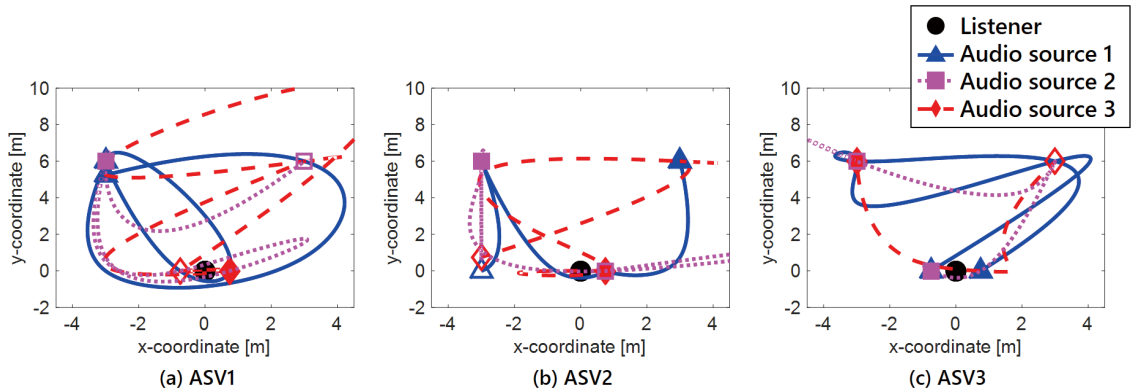

Fig. 4 Trajectory of each audio source. Differences between ASV implementations can be compared.

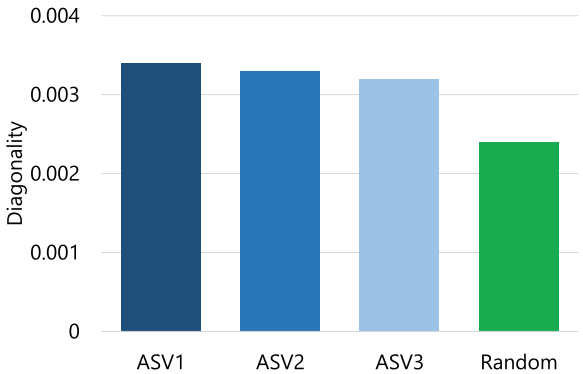

Fig. 5 Diagonality of subspaces. When using the proposed methods, the diagonality of subspaces were increased compared with a random arrangement of parameters (averages of 1,000 trials).

all three of the calculated ASVs (ASV1, ASV2 and ASV3), the diagonality of the ASVs selected was increased when compared with random selection of arrangement parameters. Although the highest score was obtained using ASV1, there were no significant differences between the calculated ASVs.

To investigate how representative the selected arrangement parameters are of an ASV space, a two-dimensionally compressed ASV space was calculated based on the t-SNE method [12], which allows dimension reduction while maintaining the relative distance between ASVs. The results are shown in Fig. 6. In t-SNE method, a student's t-distribution is used as the method of relative distance calculation. The many faint dots shown in Fig. 6 represent all of the available two-dimensionally compressed ASVs. When $J = 10$ ASVs were randomly selected, they were often concentrated around the origin of the two-dimensionally compressed ASV space. On the other hand, when the arrangement parameters were selected using the proposed methods, they were widely distributed within their ASV spaces. Thus, the proposed methods are effective for selecting a diverse sample of arrangement parameters so that users will obtain a wide variety of auditory localizations.

To measure how representative of an ASV space the selected arrangement parameters are as a score, a coverage ratio C is calculated as follows:

$$C = 1 - \frac{1}{M} \sum_{j \neq i} |\mathcal{S}_i \cap \mathcal{S}_j|, \tag{23}$$

where $\mathcal{S}_j$ denotes a set of dots around x_j defined by:

$$\mathcal{S}_j = \{x | d_{ASV}(x, x_j) < r, x \in \mathcal{S}\}, \tag{24}$$

$$\mathcal{S} = \{x | \text{all possible } M \text{ arrangement parameters}\}. \tag{25}$$

The x_j denotes the dot corresponding to the j-th arrangement parameter, $|\mathcal{S}_j|$ denotes the number of elements included in $\mathcal{S}_j$, $d_{ASV}(\cdot)$ is the Euclidian distance between two dots in the two-dimensionally compressed ASV space, and r denotes the radius of the j-th circle when the center is defined as x_j. The value of C becomes 1 when all of the ASVs are included in J circles. Since C varies with r, r must be determined appropriately. As a pre-investigation, we placed J dots evenly in a two-dimensionally compressed ASV space and examined the relationship between C and r. Since C became less than 1 when r was less than 40, $r = 40$ was used in this evaluation. Figure 7 shows the coverage ratios for the representative subspace when using the proposed methods as well as when using randomly selected arrangement parameters (averaged scores of 1,000 trials). From these results, it was confirmed that C increased when using the proposed methods, especially when ASV2 was utilized.

To investigate why the coverage ratio was the highest with ASV2, we applied eigenvalue analysis to the ASV space. Figure 8 shows the relationships between the number of eigenvalues and the cumulative contribution ratio. Since the cumulative contribution ratio with ASV2 became larger than that of other ASV implementations, the highest coverage ratio was obtained.

4.3 Subjective Evaluation

Subjective evaluation tests were conducted to investigate how diverse the selected auditory localizations would be perceived to be by listeners. Ten subjects evaluated the same binaurally rendered music signals of the same song remixed into ten different auditory localizations as stimuli. The ten

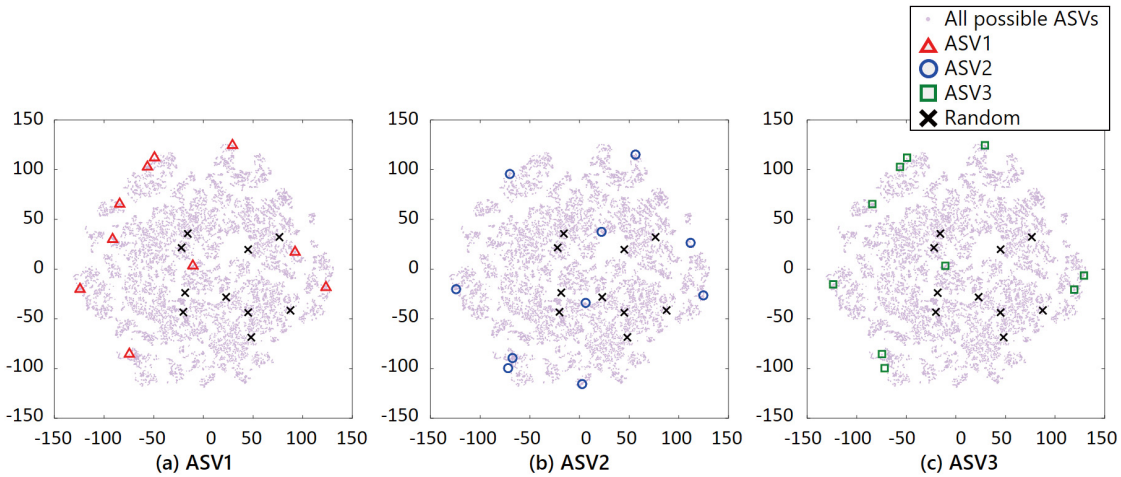

Fig. 6 Two-dimensionally compressed ASV spaces using the t-SNE method. The many faint dots represent all of the available two-dimensionally compressed ASVs. When $J = 10$ ASVs were randomly selected, they were often concentrated around the origins of the ASV spaces, but arrangement parameters selected using the proposed methods were widely distributed around the ASV spaces.

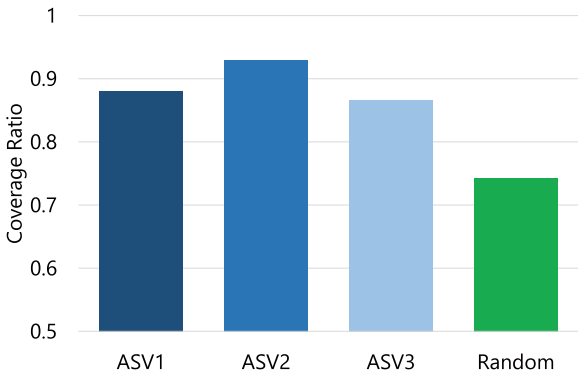

Fig. 7 Results of coverage ratio experiment. "Random" bar represents the average of 1,000 random selections.

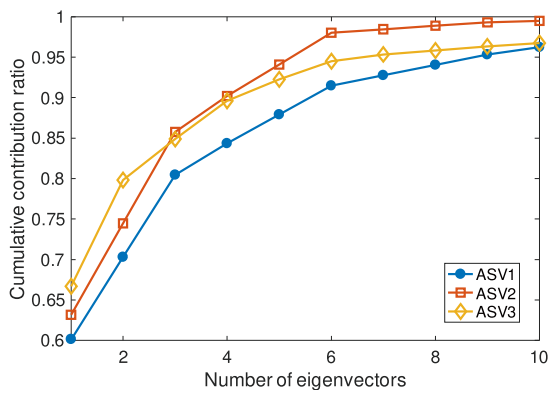

Fig. 8 Eigenvalue analysis of an ASV space. Overall, the cumulative contribution ratio was highest when calculated with ASV2.

arrangement parameters were selected using the proposed method and subjects accessed them using the proposed interface. Subjects played the 10 different localizations for an arbitrary amount of time and in an arbitrary order. After listened to all 10 localizations, they estimated the number of localizations Q which they perceived to be significantly different from the previous ones. Thus, subjects estimated the number of unique localizations Q from one to ten.

Each subject listened to binaurally rendered music signals through headphones (Audio-Technica ATH-900) in a sound-proof chamber (D value: D-85) with a 14.4 dB A-weighted back ground noise level. During the binaural spatial audio rendering, there was no individualized adaptation of the HRTFs for the subjects. Since the coverage score of the ASV space was highest when ASV2 was used to select arrangement parameter subsets, it was utilized as the proposed method. As a comparison method, $J = 10$ kinds of arrangement parameters were also selected randomly. Three trials were conducted using each of the two methods (ASV2 and random), i.e., each subject listened to a total of $10 \times 2 \times 3 = 60$ differently mixed music signals. In this experiment, "song 1" described in Table 2 was used. The ten subjects, all in their twenties, had all conducted similar evaluations about three month previously, and before testing had attended training sessions. The other experimental conditions were the same as during our objective evaluations.

The results of the subjective evaluation are shown in Fig. 9. The vertical axis represents the average number of arrangements Q whose auditory localizations were considered to be significantly different from the others by our subjects. Higher scores were obtained, when many of subjects felt they had obtained diverse auditory localizations. The horizontal axis denotes the method used for arrangement parameter extraction. These results confirm that a higher Q score was obtained when subjects used with the proposed method. A t-test was used to confirm that the difference between methods was statistically significant since the p-value was less than 0.01. Although we used $J = 10$ arrangement parameters for each evaluation, the averaged Q was less than five, thus an average of only five different arrangements dis-

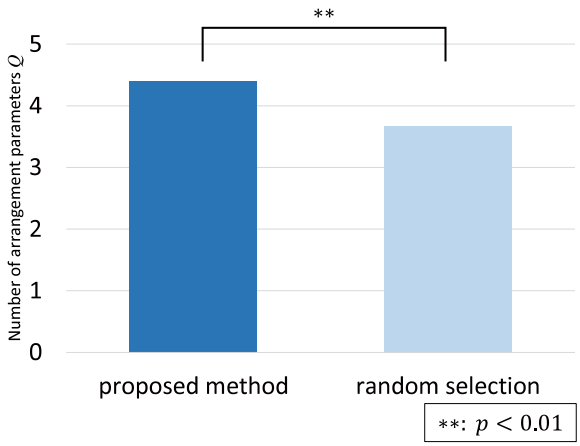

Fig. 9 Results of subjective evaluations. Vertical axis represents the average of the subjects' answers. Higher values mean the method provided a wider variety of localizations.

cernable by our subjects. Therefore, deciding how to determine J is a problem which remains to be resolved.

5. Conclusion

A single-dimensional interface for arranging audio sources which would allow users to easily obtain diverse auditory localizations was proposed. ASVs were utilized to represent the characteristics of various auditory localizations for each arrangement parameter. Since many of the possible ASVs were similar to each other, we hypothesized that an ASV space could be represented with a limited number of dimensions, which was confirmed through pre-investigations.

In order to select ASVs whose auditory localizations were diverse, we proposed an algorithm for sequentially selecting arrangement parameters so that the corresponding ASVs would be approximately orthogonal to each other. A single-dimensional interface for arranging audio sources was proposed which allowed users to select various auditory localizations by switching between sequential arrangement parameters through the manipulation of a single-dimensional interface. To avoid drastical variations between auditory localizations, arrangement parameter sorting and audio source path smoothing were conducted. Objective evaluations were conducted to confirm that diverse ASVs were selected by measuring the increase in the diagonality of a subspace when the proposed methods were used. Additionally, the results of subjective tests showed that users felt they could obtain diverse auditory localizations when manipulating the proposed interface.

One topic for further study is how to extract audio source signals from stereo music. Statistical approaches for signal extraction have been the focus of recent studies, but we hope to improve upon them so as to obtain distortion-less output signals. In this paper, we omitted inter-aural phase differences when constructing ASVs. However, in the low frequency range these differences may provide useful cues for localization, and thus this will be another focus of future work. Some problems also remain in the area of binaural synthesis. Although simple distance attenuation was multiplied using angularly interpolated HRTFs in this study, it may be necessary to take the effect of room reflection, reverberation and/or individuality of HRTFs (i.e., compensation for physical differences among subjects) into account to achieve accurate auditory localization.

Acknowledgements

We would like to thank Takanori Nishino of Mie University for his insightful advice and suggestions regarding this research.

References

[1] K. Ohtani, K. Niwa, and K. Takeda, "Single dimensional control of spatial audio object arrangement," Proc. 12th Western Pacific Acoustics Conference (WESPAC), pp.456–461, 2015.
[2] M.N.S. Swamy and K.S. Thyagarajan, "Digital bandpass and bandstop filters with variable center frequency and bandwidth," Proc. IEEE, pp.1632–1634, 1976.
[3] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. Okuno, "Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models," Proc. 9th International Society for Music Information Retrieval (ISMIR 2008), pp.133–138, 2008.
[4] K. Niwa, T. Nishino, and K. Takeda, "Development of selectable viewpoint and listening point system for musical performance," Proc. 19th International Congress on Acoustics (ICA2007), 6 pages, 2007.
[5] N. Kamado, H. Nawata, H. Saruwatari, K. Shikano, and T. Nomura, "Interactive controller for audio object localization based on spatial representative vector operation," Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC2010), 4 pages, 2010.
[6] J. Blauert, Spatial hearing: The psychophysics of human sound localization, Chap. 2, pp. 36–200, MIT Press, Cambridge, 1996.
[7] Takeda Lab., Nagoya Univ., "Head related transfer functions database," http://www.sp.m.is.nagoya-u.ac.jp/HRTF/ [Accessed online; 21 Oct. 2016].
[8] T. Nishino, S. Kajita, K. Takeda, F. Itakura, "Interpolating head related transfer functions in the median plane," Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '99), pp.167–170, 1999.
[9] T. Irino and R.D. Patterson, "A dynamic compressive gammachirp auditory filterbank," IEEE Trans. Audio Speech Language Process., vol.14, no.6, pp.2222–2232, 2006.
[10] C.D. Boor, A Practical Guide to Splines, Springer Verlag, New York, 1978.
[11] Cambridge Music Technology, "Free on-line mixing resources," http://www.cambridge-mt.com/ms-intro.htm [Accessed online; 20 Oct. 2016].
[12] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," Journal of Machine Learning Research, vol.9, pp.2579–2605, 2008.

Kento Ohtani received his B.E. and M.E. degrees from Nagoya University in 2013 and 2015, respectively. He is currently pursuing a Ph.D. degree at Nagoya University. His research interests include acoustic signal processing, spatial audio and informed source separation. He is a student member of the Acoustical Society of Japan (ASJ) and the Institute of Electronics, Information and Communication Engineers (IEICE).

Kenta Niwa received his B.E. and M.E. degrees from Nagoya University in 2006 and 2008, respectively. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 2008, he has been engaged in research on microphone array signal processing. He received his Ph.D. degree from Nagoya University in 2014 and is currently a Research Engineer at NTT Media Intelligence Laboratories. He was awarded the Awaya Prize by the ASJ in 2010. He is a member of IEEE, ASJ and IEICE.

Kazuya Takeda received his B.E. and M.E. degrees in Electrical Engineering and his Doctor of Engineering degree from Nagoya University, Nagoya Japan, in 1983, 1985 and 1994, respectively. From 1986 to 1989 he was with the Advanced Telecommunication Research laboratories (ATR), Osaka, Japan. His main research interest at ATR was corpus based speech synthesis. He was a Visiting Scientist at MIT from November 1987 to April 1988. From 1989 to 1995, he was a researcher and research supervisor at KDD Research and Development Laboratories, Kamifukuoka, Japan. From 1995 to 2003, he was an associate professor of the faculty of engineering at Nagoya University. Since 2003 he has been a professor in the Department of Media Science, Graduate School of Information Science, Nagoya University. His current research interests are media signal processing and its applications, which include spatial audio, robust speech recognition and driving behavior modeling.