PAPER

Selective Listening Point Audio Based on Blind Signal Separation and Stereophonic Technology

Kenta NIWA[†a)], **Takanori NISHINO**[††b)], ***and*** **Kazuya TAKEDA**[†c)], ***Members***

SUMMARY A sound field reproduction method is proposed that uses blind source separation and a head-related transfer function. In the proposed system, multichannel acoustic signals captured at distant microphones are decomposed to a set of location/signal pairs of virtual sound sources based on frequency-domain independent component analysis. After estimating the locations and the signals of the virtual sources by convolving the controlled acoustic transfer functions with each signal, the spatial sound is constructed at the selected point. In experiments, a sound field made by six sound sources is captured using 48 distant microphones and decomposed into sets of virtual sound sources. Since subjective evaluation shows no significant difference between natural and reconstructed sound when six virtual sources and are used, the effectiveness of the decomposing algorithm as well as the virtual source representation are confirmed.

key words: *acoustic field representation, blind source separation, frequency-domain independent component analysis (FD-ICA), spatial grouping*

1. Introduction

As an extension of multi-viewpoint image processing, free-viewpoint TV (FTV) systems [1], [2] that can generate scenes at an arbitrarily selected viewpoint have become an issue in MPEG standardization [3]. The goal of this research is to build a selective listening point (SLP) audio system that can be used for the audio part of the FTV system.

SLP audio is a spatial sound reproduction system characterized by three requirements: 1) microphones should be placed at distant locations from sound sources, 2) the system must work on the condition that the number and locations of the sound sources are unknown, 3) each sound source may move independently, and 4) the reproduced sound signals can be presented with ordinary equipment such as earphones, headphones and a stereo loudspeaker system. Therefore, simply applying an existing spatial audio reproduction method, such as binaural recording [4] or transaural audio [5] by boundary surface control with a speaker array [6], fails to achieve SLP audio. Figure 1 shows a block diagram of the SLP audio system.

In a previous work [7], we evaluated an SLP audio system that combined blind source separation (BSS) and binaural audio with a head-related transfer function (HRTF). In that system, BSS separated the mixtures of signals recorded at distant microphones, into independent source signals. Then a spatial impression was added to them through HRTFs between the selected listening point and the source locations. Through a preliminary experiment, we confirmed that even signal separation by BSS is not perfect, but after convolving the signals with transfer functions and remixing, natural spatial sound was reconstructed.

However, in that experiment we presumed that the number and locations of the sound sources were known. In this paper, we extend the previous work to eliminate the need for prior knowledge about either the number or locations of the sound sources.

Extension of the SLP algorithm mainly consists of three parts [8]. The first is finding virtual sound sources. Since accurate identification of real sound sources is not necessary in SLP audio, e.g., discriminating closely located sound sources is unnecessary, we roughly estimate the number of sound sources based on subspace analysis of the spatial correlation matrix [9]. BSS is applied in the obtained subspace to find the separation matrix for the estimated number of source signals, which we call virtual source signals.

The second part is localizing the signals. Since we use frequency-domain independent component analysis (FD-ICA) for signal separation, there is an ambiguity known as permutation in associating independent signal components with the correct sound source for every frequency bin. Instead of solving this permutation problem, in the proposed method, we cluster all of the virtual source signals into a predetermined number of groups across all frequency bins. Clustering, which is performed based on the acoustic transfer functions from the position of the virtual source signal to the microphones, is calculated from the pseudo-inverse of the separation matrix. The reconstructed signal from the group of virtual source signals, which we call the local signal mixture, represents either a signal of one source or a mixture of different source signals located in close positions.

The third part is determining the reference location of the local signal mixture. Here, we calculate the centroid of the groups in the virtual source subspace and transfer them back to the real geometrical space.

Through the above three steps, we can decompose multiple microphone signals into a set of virtual sound source information, i.e., the location and associated signals, which is the natural generalization of a typical 3D sound field rep-

Manuscript received July 10, 2008.
Manuscript revised October 24, 2008.

[†]The authors are with the Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8601 Japan.
[††]The author is with Center for Information Media Studies, Nagoya University, Nagoya-shi, 464–8603 Japan.
a) E-mail: niwa@sp.m.is.nagoya-u.ac.jp
b) E-mail: nishino@media.nagoya-u.ac.jp
c) E-mail: kazuya.takeda@nagoya-u.jp

DOI: 10.1587/transinf.E92.D.469

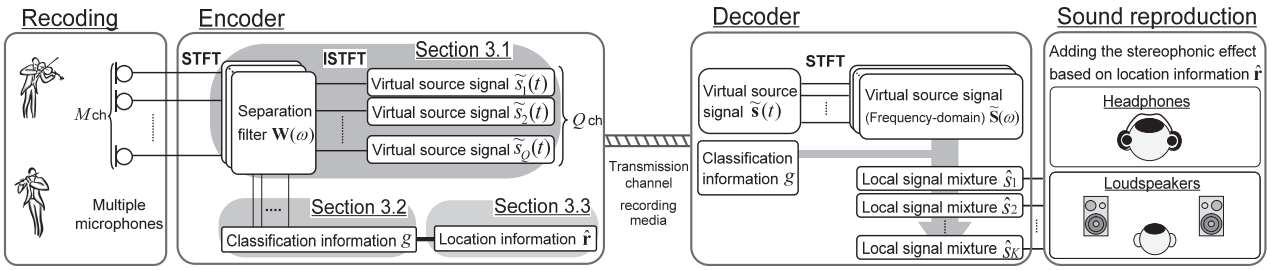


Fig. 1 Block diagram of selective listening point audio system. SLP audio system consists of four parts: 1) recording with multiple distant microphones, 2) encoder based on blind source separation, 3) decoder based on spatial audio technique, and 4) sound reproduction with ordinary audio device.

resentation. After decomposing, therefore, the local sound field at the selected listening point is flexibly presented. In this study, a binaural system based on an HRTF is used.

The rest of the paper is organized as follows. The basic idea of SLP audio using BSS is described in Sect. 2. In Sect. 3, the proposed algorithm is detailed. After showing an experimental evaluation in Sect. 4, we conclude the paper in Sect. 5.

2. Selective Listening Point Audio Using Blind Source Separation

One of the simplest ways to define the 3D sound field is to specify the locations of the sound sources and the corresponding source signals:

$$\Omega = \{\mathbf{r}_n, s_n(t)\}, \qquad n = 1, \cdots, N, \tag{1}$$

where $\mathbf{r}_n$ and $s_n(t)$ denote the location and the signal of the n-th sound source. Given listening position $\mathbf{r}^{(R)}$, target sound $y(t)$ can be calculated by

$$y(t) = \sum_{n=1}^{N} h(\mathbf{r}_n, \mathbf{r}^{(R)}) * s_n(t), \tag{2}$$

or in the frequency domain

$$Y(\omega) = \sum_{n=1}^{N} H(\mathbf{r}_n, \mathbf{r}^{(R)}) \cdot S_n(\omega), \tag{3}$$

when the acoustic transfer function between $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ is given by $h(\mathbf{r}_\alpha, \mathbf{r}_\beta)$. Typically in the binaural audio case, column vector $\mathbf{h}(\mathbf{r}_\alpha, \mathbf{r}_\beta) = [h^{(\text{left})}(\mathbf{r}_\alpha, \mathbf{r}_\beta), h^{(\text{right})}(\mathbf{r}_\alpha, \mathbf{r}_\beta)]^T$ is used for the transfer function (HRTF). Therefore, the main problem of the SLP audio system is decomposing the multichannel signals captured through M distant microphones into source information Ω.

Potentially, BSS can be used for part of the decomposing by finding a set of independent signals $\hat{\mathbf{s}}(t)$. In particular, the frequency-domain ICA [10] combined with advanced methods for solving permutation ambiguity [11] is powerful under realistic acoustic conditions. However, since the assumption about the number of sources is crucial in BSS, accurate estimation of the independent source is difficult in such applications as SLP where the number of sound sources varies widely.

In a previous study, we evaluated the performance of SLP audio using BSS [7] under the assumption of the prior knowledge of the number and locations of the sound sources. Through the experiment, we found that imperfect separation does not cause serious problems in an SLP audio application because source signals are remixed in the target signal anyway. Therefore, to achieve an SLP audio system, we extend the BSS algorithm to operate it without any prior knowledge of sound sources, and build a decomposing algorithm that converts the multi channel signals into virtual source information.

3. Algorithm

3.1 Estimating Virtual Source Signals

Since the number of sound sources is unknown, we first roughly estimate them by subspace analysis on the spatial correlation matrix [12]:

$$\mathbf{R}(\omega) = E\{\mathbf{X}(\omega)\mathbf{X}(\omega)^H\}, \tag{4}$$

where $\mathbf{X} = [X_1(\omega), \cdots, X_M(\omega)]^T$ is the frequency domain representation of the signals captured at M distant microphones. H and E denote the conjugate transpose and the expectation operations, respectively. By decomposing $\mathbf{R}(\omega)$ into the form of $\mathbf{R}(\omega) = \mathbf{V}(\omega)\mathbf{\Lambda}(\omega)\mathbf{V}(\omega)^{-1}$ and truncating the dimensions whose eigen values are smaller than a predetermined threshold, we get Q eigen vectors of $\mathbf{R}(\omega)$ matrix, i.e., $\mathbf{V}'(\omega) = [\mathbf{v}_1(\omega), \cdots, \mathbf{v}_Q(\omega)]^T$. Although Q is an estimate of the source number, as we see below, the overall performance is not so sensitive to the accuracy of the estimate because most of the signals are remixed in the target signal. When Q is overestimated, the echoes of the original signal are identified as likely independent sources. $\mathbf{\Lambda}'$ denotes the truncated version of the diagonal eigen value matrix.

FD-ICA is performed on subspace signal $\mathbf{Z}(\omega) = [Z_1(\omega), \cdots, Z_Q(\omega)]^T$ given by

$$\mathbf{Z}(\omega) = (\mathbf{\Lambda}'(\omega))^{-1/2}\, \mathbf{V}'(\omega)\mathbf{X}(\omega). \tag{5}$$

The iterative learning rule below [13], [14] is used for estimating a separation matrix $\mathbf{U}(\omega)$ for subspace signal $\mathbf{Z}(\omega)$:

$$\mathbf{U}_{t+1} = \mathbf{U}_t + \mu \cdot \text{off-diag}\{E[\varphi(\mathbf{Z})\mathbf{Z}^H]\}\mathbf{U}_t, \tag{6}$$

where $\varphi(z) = \tanh(\beta \cdot \Re(z)) + j \cdot \tanh(\beta \cdot \Im(z))$ denotes an activating function.

The separation matrix for the original microphone signals is given by

$$\mathbf{W} = \mathbf{U}\,(\mathbf{\Lambda}')^{-1/2}\,\mathbf{V}'. \tag{7}$$

Since there is the amplitude ambiguity in the separation matrix $\mathbf{W}(\omega)$, the projection back method [15] was used to solve this problem. The projection back method is one of the methods for solving this ambiguity, and the method generates the projected filter by using acoustic transfer functions among the virtual sources and one of the microphones. In our method, the separation matrix is projected by the average acoustic transfer function of the recorded environment.

Finally, Q independent signals $\tilde{\mathbf{S}}(\omega)$, called virtual source signals, can be calculated for each frequency bin by

$$\tilde{\mathbf{S}} = \mathbf{UZ} = \mathbf{U}\,(\Lambda')^{-1/2}\,\mathbf{V}'\mathbf{X} = \mathbf{WX}. \tag{8}$$

Note that we omit frequency index (ω) from $\tilde{\mathbf{S}}(\omega)$, $\mathbf{U}(\omega)$, $\mathbf{V}(\omega)$, $\mathbf{W}(\omega)$, $\mathbf{X}(\omega)$, $\mathbf{Z}(\omega)$, and $\mathbf{\Lambda}(\omega)$ in Eqs. (6) through (8). Inverse short-time Fourier transform (ISTFT) and overlap add will reproduce virtual source signals in time domain $\tilde{s}(t)$.

3.2 Grouping Virtual Signal Components

The pseudo-inverse of separation matrix $\mathbf{W}(\omega)$ represents the acoustic transfer functions from the source positions of the virtual source signals to M microphones. We denote the pseudo-inverse matrix by $\mathbf{W}^+(\omega) = [\mathbf{w}_1^+(\omega), \cdots, \mathbf{w}_Q^+(\omega)]$, where $\mathbf{w}_q^+(\omega)$ is a transfer function vector from the position of the q-th virtual source to M microphones, i.e., $\mathbf{w}_q^+(\omega) = [w_{1q}^+(\omega), \cdots, w_{Mq}^+(\omega)]^T$, for frequency ω. The phase component of that vector contains geometrical information of the virtual sources. In [11], [16]–[20], this geometrical information was used to solve the permutation problem of FD-ICA.

Since the estimate of the number of virtual sources is not accurate and the spectral components of the virtual source signals, i.e., $\tilde{S}_1(\omega), \cdots, \tilde{S}_Q(\omega)$, have permutation ambiguity across frequency indexes, we group closely located virtual sources and reconstruct the mixture of the signals of those virtual sources through clustering as follows.

The phase component of transfer function vector $\mathbf{w}_q^+(\omega)$ represents the relative arrival delay from the virtual sound source to each microphone element. Therefore, we define operation $\phi(\cdot)$ on the transfer function vectors to extract the relative phase at each microphone [20]:

$$\phi(\mathbf{w}_q^+(\omega)) = [\exp(j\xi_{q,1}), \cdots, \exp(j\xi_{q,M})]. \tag{9}$$

$\xi_{q,m}$ is the normalized delay given by

$$\xi_{q,m} = \frac{\arg(w_{q,m}^+(\omega))}{2\omega d/\pi c}, \tag{10}$$

where d is the array size in which the m-th microphone is located and $\arg(\cdot)$ operation calculates the relative phase angle in that array. As seen in the experiment below, we assume that each microphone is arranged as an element of one of the L arrays. We denote a set of microphones included in the l-th array by $\theta(l)$. A microphone array can catch acoustic characteristics such as a transfer function and a sound pressure distribution in the local acoustic field, and the acoustic characteristics corresponding to the location of the sound source are defined uniquely by using multiple arrays. By applying $\phi(\cdot)$, we can cancel the frequency dependency from $\mathbf{w}_q^+(\omega)$ and cluster the virtual sources across frequency bins, as shown in Fig. 2.

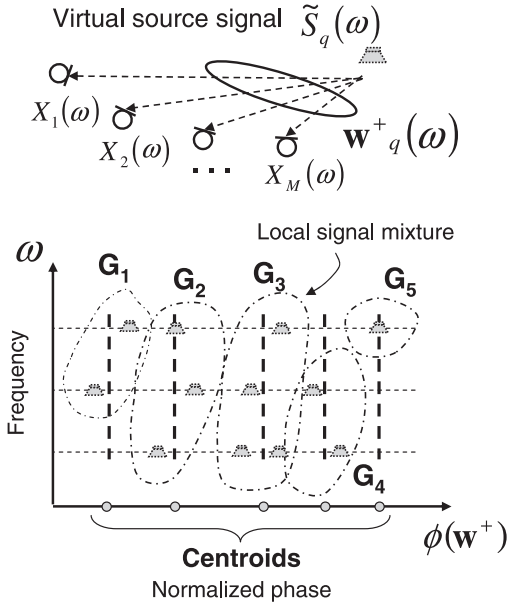


Fig. 2 Virtual source signals, their groups and centroids [8]. A column vector of the pseudo-inverse of separation matrix, i.e., $\mathbf{w}_q^+(\omega)$, represents acoustic transfer functions from q-th virtual source to microphones at frequency ω. The top figure shows that grouping transfer function vectors across frequencies and combining corresponding signal components give a mixture of closely located sound signals. $Q = 4$ virtual sources are clustered into $K = 5$ clusters shown at bottom.

The similarity between phase vectors is defined by the sum of scalar products over arrays:

$$\mathrm{Sim}\left(\mathbf{w}_\alpha^+(\omega_\psi), \mathbf{w}_\beta^+(\omega_\varphi)\right) = \sum_{l=1}^{L} \left| \sum_{m\in\theta(l)} \phi(w_{\alpha,m}^+(\omega_\psi))^* \cdot \phi(w_{\beta,m}^+(\omega_\varphi)) \right|, \tag{11}$$

where $(\cdot)^*$ represents a complex conjugate. For example, we consider two virtual sources $\tilde{S}_\alpha(\omega_1)$ and $\tilde{S}_\beta(\omega_2)$, which propagated from one sound source, and these virtual sources should be clustered to the same group. There is the phase shift $\exp(j\theta)$ between two transfer function vectors $\phi(\mathbf{w}_\alpha^+(\omega_1))$ and $\phi(\mathbf{w}_\beta^+(\omega_2))$ corresponding to two virtual sources respectively. The calculation of absolute value in Eq. (11) removes this phase shift because $|\exp(j\theta)| = 1$. Therefore, the similarity measure in Eq. (11) is robust to the constant phase shift due to the ambiguity of the array position and the sound source. Based on this similarity measure, we cluster $Q \times D$ transfer function vectors into K clusters. D denotes the number of frequency bins. Note that K can be

more than Q. A grouping information g is calculated by

$$g(q,\omega) = \arg\max_{k} \mathrm{Sim}\left(\bar{\mathbf{w}}_k^+(\omega), \mathbf{w}_q^+(\omega)\right), \tag{12}$$

where $\bar{\mathbf{w}}_k^+ = [\bar{\mathbf{w}}_1^+, \cdots, \bar{\mathbf{w}}_K^+]$ is the centroid of k-th cluster. Centroids $\bar{\mathbf{w}}^+$ are needed to estimate the location of local signal mixture $\hat{s}$, and clusters are decided with the k-means algorithm.

Denoting the clustering results in which transfer function vector $\mathbf{w}_q^+(\omega)$ falls into the k-th category by $k = g(q,\omega)$, local signal mixture $\hat{S_k}(\omega)$ is given by

$$\hat{S_k}(\omega) = \sum_{q=1}^{Q} \delta_{k,g(q,\omega)} \mathbf{w}_q \cdot \mathbf{X}(\omega), \tag{13}$$

with $\delta_{i,j}$ as the Kronecker delta. Finally, ISTFT and overlap add will reproduce a mixture of locally located signals in time domain $\hat{s_k}(t)$.

3.3 Location Estimation

The reference location of k-th local signal mixture $\hat{s}_k$ can be estimated from the centroid of the k-th cluster of the transfer function vectors. Since we use a set of microphone arrays as the distributed sensors, the steering vector is used for converting the centroid to the signal source location.

For the l-th microphone array, a steering vector to location $\mathbf{r}$ is given by

$$\mathbf{a}_l(\mathbf{r}) = \left[\exp\left(j\frac{\pi|\mathbf{r}_{l,1}^{(r)} - \mathbf{r}|}{2d_l}\right), \cdots, \exp\left(j\frac{\pi|\mathbf{r}_{l,\theta(l)}^{(r)} - \mathbf{r}|}{2d_l}\right)\right], \tag{14}$$

where $\mathbf{r}_{l,i}^{(r)}$ represents the position of the i-th element of the l-th array and d_l denotes the array size.

As in the clustering case, the similarity between the K centroids of the transfer function vectors, $\{\bar{\mathbf{w}}_k^+\}_{k=1,\cdots,K}$, and a steering vector [21] can be calculated. We search for the location where the similarity becomes largest as the reference position of local signal mixture

$$\hat{\mathbf{r}}_k = \arg\max_{\mathbf{r}} \sum_{l=1}^{L} \left| \sum_{m\in\theta(l)} \phi(\bar{w}_{k,m}^+(\omega))^* \cdot a_{l,m}(\mathbf{r}) \right|. \tag{15}$$

The local signal mixture is a monaural signal that includes acoustical transfer functions among the virtual sound sources and one of the microphones. Finally, the estimated 3D sound field representation $\hat{\Omega} = \{\hat{\mathbf{r}}_k, \hat{s}_k(t)\}_{k=1,\cdots,K}$ is obtained.

4. Experimental Evaluation

4.1 Experimental Setup

Figures 3 and 4 show the experimental setup for the acoustic systems. Six 6-element arrays and a 12-element array

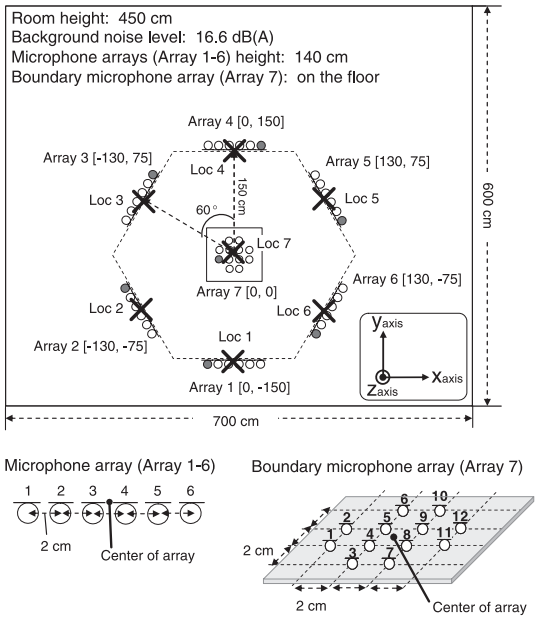


Fig. 3 Experimental setup. Seven microphone arrays surrounding six loudspeakers in a linear arrangement. One array is a 2D boundary array and located on the floor. The other six are linear arrays, located at a height of 140 cm. The black one in Arrays 1 to 6 indicates the first microphone of the array. The origin in this experimental setup is the center of Array 7.

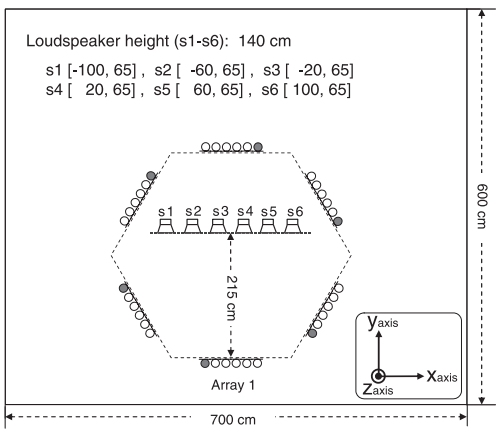


Fig. 4 Experimental setup of six loudspeakers. All loudspeakers are located at a height of 140 cm. Origin is same as Fig. 3.

are arranged to surround the six loudspeakers. All 48 sensors are omni-directional microphones (SONY ECM-77B). Six loudspeakers (BOSE ACOUSTMASS) are arranged in a linear form. Six source signals played at the loudspeakers are recorded at the 48 distant microphones in a synchronous manner with a sampling frequency of 40 kHz. The background noise level was 16.6 dB (A), and the reverberation time T_{60}, which was calculated by Schroeder integration [22], was 138 msec. As for the test signals, we recorded speech, popular music (Music 1: Winter games), and classical music (Music 2: Jupiter), as listed in Table 1. Duration of all test signals is 15 sec. Other conditions are listed in Table 2.

Table 1 Collection list with organizational sound sources.

	Speech (JNAS)	Music1: Winter games (D. Foster)	Music2: Jupiter (G. Holst)
s1	Female speech1	Bass	Strings1
s2	Male speech1	Brass	Strings2
s3	Female speech2	Drums	Percussions
s4	Male speech2	Piano	Timpani
s5	Female speech3	Strings	Brass1
s6	Male speech3	Orchestra Hit	Brass2

Table 2 Parameters of SLP audio system.

Sampling frequency, F_s	40 kHz
Number of microphone arrays, L	7
Number of microphones, M	48
Number of sources, N	6
Length of STFT, D	2048 pt (51.2 msec)
Frame shift of STFT	512 pt (12.8 msec)
Window function	Hamming
Number of virtual sources, Q	2, 3, $\cdots$, 20
Number of clusters, K	2, 3, $\cdots$, 20

In this study, we assume a binaural system based on the HRTF as a sound reproduction system. The obtained virtual sources are convolved with HRTFs on the appropriate direction. The measured HRTFs were used after interpolation [23] to add a spatial impression and to the estimated local signal mixtures. Spatial impressions such as sound source distance and direction are added by the HRTFs. Other spatial impressions such as reverberation are given by the estimated local signal mixtures. They include the average acoustic transfer function in the environment because we used the projection back method [15] that generates the projected matrix by using acoustic transfer functions among virtual sources and one of the microphones. The HRTFs were measured with a head-and-torso simulator (B&K 4128) and these data can be downloaded at [24].

4.2 Evaluation Results

Objective and subjective tests were conducted to evaluate the performances of representing the sound field.

4.2.1 Objective Results

The number of virtual sources Q influences sound quality and transinformation. As Q decreases, sound quality is worsened.

The sound quality was evaluated by comparing reference signal $\mathrm{ref}_{SQ}(t)$ to the signal obtained by reducing-order filter $y_{SQ}(t)$:

$$\mathrm{ref}_{SQ}(t) = \sum_{n=1}^{N} s_n(t), \tag{16}$$

$$y_{SQ}(t) = \mathrm{IFFT}\left[\sum_{q=1}^{Q} \mathbf{w}_q(\omega)\mathbf{X}(\omega)\right]. \tag{17}$$

The cepstrum distance was used for measuring the similarity between them:

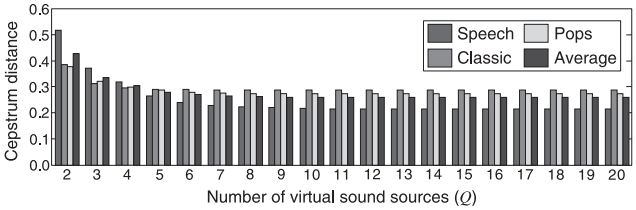


Fig. 5 Results of evaluating sound quality with cepstrum distance.

$$\mathrm{D}_{\mathrm{cep}} = E\left(\sqrt{\sum_{k=1}^{D}[c_{y_{SQ}}(k) - c_{\mathrm{ref}_{SQ}}(k)]^2}\right), \tag{18}$$

where $c_{y_{SQ}}(k)$ is the k-th order cepstrum of $y_{SQ}(t)$ and $c_{\mathrm{ref}_{SQ}}(k)$ is the k-th order cepstrum of $\mathrm{ref}_{SQ}(t)$. Since the cepstrum distance is one of the methods for measuring sound quality, lower $\mathrm{D}_{\mathrm{cep}}$ gives us good sound quality. Therefore, we employ this distance as a criterion of deciding the parameter Q.

Figure 5 shows the results of evaluating sound quality with cepstrum distance. There is no difference in the cepstrum distance when the number of virtual sources Q is more than six. This result corresponds to the number of real sources. Therefore, condition $Q = 6$ is used in the following evaluation.

Since the number of clusters K corresponds to the number of divisions of the acoustic field, the degree of mixing with each other is low for the large K. A low degree of mixture produces good performance of division into each sound signal. The number of clusters K influences sound localization and the sound localization is improved as K increases. Sound localization performance was obtained by calculating the difference between reference signal $\mathrm{ref}_{LQ}(t)$ and local signal mixtures $y_{LQ}(t)$:

$$\mathrm{ref}_{LQ}(t) = \sum_{n=1}^{N} h(\mathbf{r}_n, \mathbf{r}^{(R)}) * s_n(t), \tag{19}$$

$$y_{LQ}(t) = \sum_{k=1}^{K} h(\hat{\mathbf{r}}_k, \mathbf{r}^{(R)}) * \hat{s}_k(t), \tag{20}$$

where $\mathbf{r}_n$ is the position information and $\hat{\mathbf{r}}_k$ is the estimated position information. The sound localization is achieved by the interaural time difference (ITD) and level difference (ILD), however, it is difficult to calculate the ITD for multiple sound sources. Therefore, we calculate the ILD and employ this distance as a criterion for deciding parameter K. The ILD is calculated as an inter-channel level difference (ICLD).

$$\mathrm{ICLD}_x = 10\log_{10}\frac{\sum_t^N x_R(t)^2}{\sum_t^N x_L(t)^2} \quad [\mathrm{dB}], \tag{21}$$

where $x_R(t)$ and $x_L(t)$ are transduced signals with sound equipment such as headphones. R and L denote right and left channel, respectively. The difference between the reference signal and local signal mixtures was calculated using an ICLD:

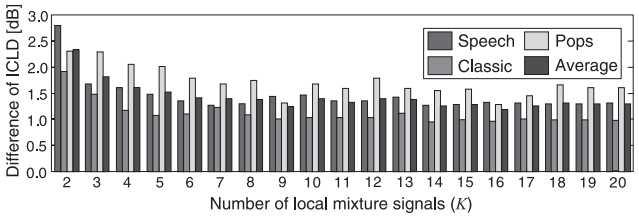


Fig. 6 Results of evaluating sound localization with inter-channel level difference.

$$\mathrm{D_{ICLD}} = E\left(\left|\mathrm{ICLD}_{\mathrm{ref}_{LQ}} - \mathrm{ICLD}_{y_{LQ}}\right|\right) \ \text{[dB]}. \tag{22}$$

Figure 6 shows the results of evaluating sound localization with ICLD. The best performance was obtained for $K = 16$. Therefore, subjective tests were conducted under conditions where $Q = 6$ and $K = 16$.

4.2.2 Subjective Results

For $Q = 6$ and $K = 16$, subjective tests were performed using the XAB method, which is the standard method for evaluating the sound quality of an audio signal with very low degradation [25]. Three stimuli of X, A, and B were presented to the subjects. These stimuli were made from three test signals: speech, popular music and classic music. These test signals were divided into three parts, each with a duration of 5 sec. Seven locations of sound sources were assumed on the center of microphone array. Thus the sets of stimuli were 63 (9 signals × 7 locations). In our experiments, stimulus X was the reference signal (Eq. (19)). Either A or B was the same signal as X, and the other was the comparison signal (Eq. (20)). However, subjects did not know which signals were reference or comparison. Subjects evaluated the degradations between X and A, and between X and B. Answers about the degradations of sound quality and sound localization were required every set of X-A-B. The evaluation grades are shown in Table 3. The obtained grade was converted to subjective difference grade (SDG):

$$\mathrm{SDG} = G_{\mathrm{ev}} - G_{\mathrm{ref}}, \tag{23}$$

where G_{ev} is a grade between the comparison and reference signals, and G_{ref} is a grade between both reference signals. SDG ranged from -4 to 0, and each grade is also shown in Table 4.

Eleven subjects (ten males and one female) examined the sound quality and sound localization, respectively. The evaluated signals at the seven listening points, Loc 1 to 7 shown in Fig. 3, were generated. The duration of every stimulus was 5 sec. Stimuli were presented by intra-concha earphones (Etymotic research ER-4B).

Figures 7 and 8 show the sound quality and sound localization, respectively. Figures 7 (b) and 8 (b) indicate that good representation of the sound field was obtained by source information Ω. The average SDG of sound quality and sound localization was -0.22 and -0.23, respectively. There is less difference between the reference and comparison signal.

Table 3 Grades and qualities of XAB test.

Grade	Quality
5	Inaudible
4	Audible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Table 4 Grades and qualities of subjective difference grade.

SDG	Quality
0	Inaudible
-1	Audible, but not annoying
-2	Slightly annoying
-3	Annoying
-4	Very annoying

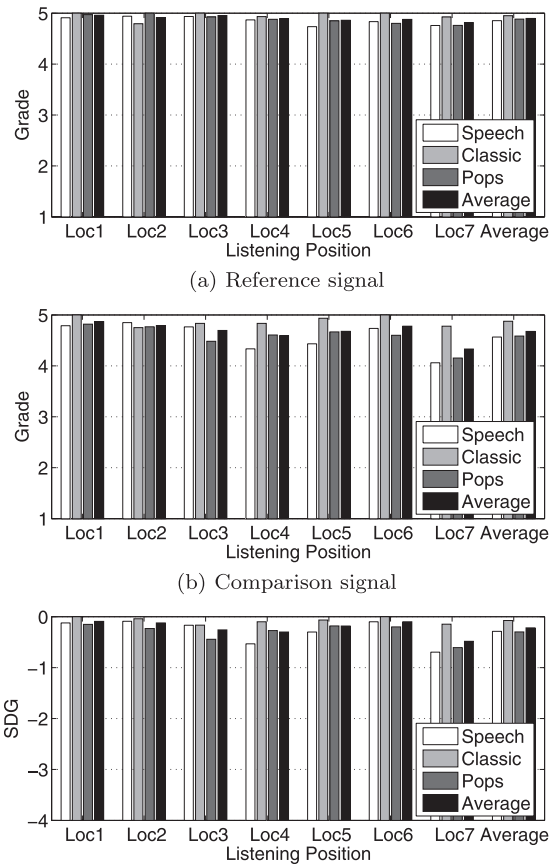
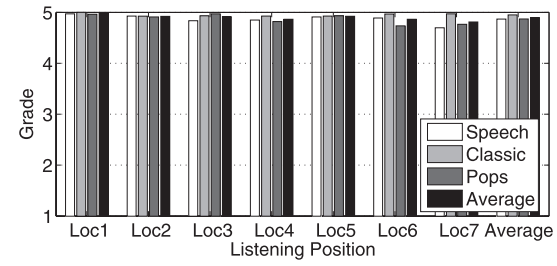
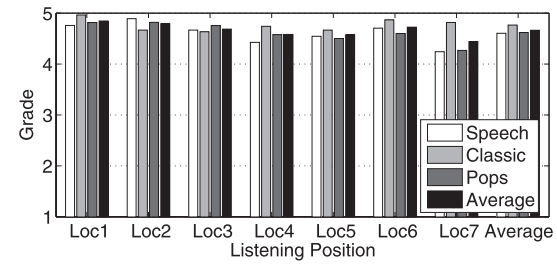


Fig. 7 Results of subjective test for sound quality for $Q = 6$ and $K = 16$. Loc 1 to 7 are listening positions shown in Fig. 3.

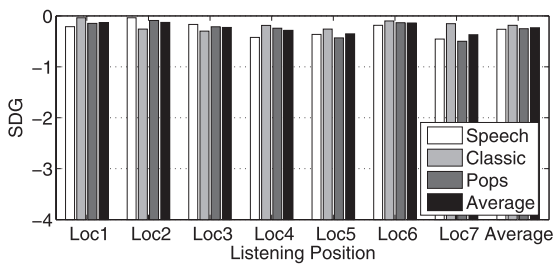
The results confirmed that the proposed decomposing method as well as 3D sound field representation based on virtual sound sources is effective for an SLP audio system. Subjective experiments were conducted in the case of $Q = 6$ and $K = 16$ which were decided by objective measures D_{cep} and D_{ICLD}, respectively. It suggests that smaller D_{cep} and D_{ICLD} are one of the effective criteria for deciding Q and K, however, more investigation is needed in future works.



(a) Reference signal



(b) Comparison signal



(c) Subjective difference grade

Fig. 8 Results of subjective test for sound localization for $Q = 6$ and $K = 16$. Loc 1 to 7 are listening positions shown in Fig. 3.

5. Summary and Future Works

In this paper, we proposed and evaluated a new spatial audio scheme: a selective listening point audio system. In the system, a 3D acoustic field is represented by a set of signal sources with their locations and associated signals. We developed a method to decompose the multichannel signals recorded at distant positions into this representation based on BSS technologies.

For evaluation, the proposed method was applied to decompose the signals captured through 48 distant microphones into a set of virtual signals. Subjective evaluation showed the effectiveness of the proposed method, revealing that the spatial impression of the resultant spatial sound is as high as the natural reference sounds. The number of local signal mixtures is more influential than the number of virtual sources. However, when the number of local signal mixtures is higher than that of the real sound sources, there is no significant difference in the spatial impression of the sound. These results suggest an important insight into information reduction achieved in the proposed system, which is one of our most crucial future works.

Many other issues need further study, including the optimal array arrangement and performance under more reverberant and/or noisy conditions. Among these problems, dealing with a non-stationary sound field, e.g., moving sources, is one of the most important.

A demonstration of SLP audio can be downloaded from:
http://www.sp.m.is.nagoya-u.ac.jp/~niwa/slpademo-e.html

Acknowledgment

This work was supported by a Grant-in-Aid for Scientific Research (No. 18300064).

References

[1] T. Fujii and M. Tanimoto, "Free-viewpoint TV system based on the ray-space representation," SPIE ITCom, vol.4864-22, pp.175–189, Aug. 2002.
[2] N. Fukushima, K. Niwa, T. Yendo, T. Fujii, M. Tanimoto, T. Nishino, and K. Takeda, "Free viewpoint and listening-point video generation by using multi viewpoint and multi listening-point data capturing system," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J91-D, no.8, pp.2039–2041, Aug. 2008.
[3] ISO/IEC JTC1/SC29/WG11 (N9168), July 2007.
[4] J. Blauert, Spatial Hearing (revised ed.), MIT Press, 1996.
[5] J. Bauck and D.H. Cooper, "Generalized transaural stereo and applications," J. Audio Eng. Soc., vol.44, no.9, pp.683–705, Sept. 1996.
[6] S. Ise, "The boundary surface control principle and its applications," IEICE Trans. Fundamentals, vol.E88-A, no.7, pp.1656–1664, July 2005.
[7] K. Niwa, T. Nishino, and K. Takeda, "Development of selectable viewpoint and listening point system for musical performance," ICA 2007, PPA-06-011, Sept. 2007.
[8] K. Niwa, T. Nishino, and K. Takeda, "Encoding large array signals into 3D sound field representation for selective listening point audio based on blind source separation," ICASSP 2008, pp.181–184, March 2008.
[9] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," IEEE Trans. Speech Audio Process., vol.11, no.3, pp.204–215, May 2003.
[10] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, vol.22, no.1-3, pp.21–34, Nov. 1998.
[11] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," EURASIP J. Applied Signal Processing 2003, pp.1135–1146, Nov. 2003.
[12] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," IEEE Trans. Acoust. Speech Signal Process., vol.33, no.2, pp.387–392, April 1985.
[13] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," Neural Comput., vol.7, no.6, pp.1129–1159, Nov. 1995.
[14] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," Int. Workshop on ICA and BSS, pp.371–376, Jan. 1999.
[15] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," Int. Sym. on Nonlinear Theory and its Application, pp.923–926, 1998.
[16] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," ICASSP 2000, pp.3140–3143, June 2000.

[17] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind signal separation on speech signals," Proc. Int. Sym. ICA and BSS, pp.505–510, April 2003.
[18] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Near-field frequency domain blind source separation for convolutive mixtures," ICASSP 2004, vol.IV, pp.49–52, May 2004.
[19] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Solving the permutation problem of frequency-domain bss when spatial aliasing occurs with wide sensor spacing," ICASSP 2006, vol.V, pp.77–80, May 2006.
[20] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," IEEE Trans. Audio, Speech, and Language Processing, vol.15, no.5, pp.1592–1604, July 2007.
[21] D.H. Johnson and D.E. Dudgeon, Array signal processing: Concepts and techniques, Prentice Hall, 1993.
[22] M.R. Schroeder, "New method of measuring reverberation time," J. Acoust. Soc. Am., vol.37, no.3, pp.409–412, March 1965.
[23] T. Nishino, S. Mase, S. Kajita, K. Takeda, and F. Itakura, "Interpolating HRTF for auditory virtual reality," Third Joint Meeting ASA and ASJ, 1pSP6, pp.1261–1266, Dec. 1996.
[24] http://www.sp.m.is.nagoya-u.ac.jp/HRTF/
[25] Methods for the subjective assessment of small impairments in audio systems including multi-channel sound systems, ITU-R BS.1116, 1997.

Kenta Niwa received the B.E. and M.E. degrees from Nagoya university in 2006 and 2008, respectively.

Takanori Nishino received the B.E., M.E., and Dr. Eng. degrees from Nagoya university in 1995, 1997 and 2003, respectively. He was assistant professor at the Faculty of urban science, Meijo university from 2000 to 2003. He is currently assistant professor at the Center for information media studies, Nagoya university. His research interest is spatial audio. He is a member of the ASJ, ASA, and IPSJ.

Kazuya Takeda received the B.S. degree, the M.S. degree, and the Dr. of Engineering degree from Nagoya University, in 1983, 1985, and 1994 respectively. In 1986, he joined ATR (Advanced Telecommunication Research Laboratories), where he involved in the two major projects of speech database construction and speech synthesis system development. In 1989, he moved to KDD R&D Laboratories and participated in a project for constructing voice-activated telephone extension system. He has joined Graduate School of Nagoya University in 1995. Since 2003, he is a professor at Graduate School of Information Science at Nagoya University. He is a member of The ASJ, IEEE, and IPSJ.