ENCODING LARGE ARRAY SIGNALS INTO A 3D SOUND FIELD REPRESENTATION FOR SELECTIVE LISTENING POINT AUDIO BASED ON BLIND SOURCE SEPARATION

Kenta NIWA†, Takanori NISHINO‡, Kazuya TAKEDA†

† Graduate School of Information Science, Nagoya University, Nagoya, Japan
‡ Center for Information Media Studies, Nagoya University, Nagoya, Japan

ABSTRACT

A sound field reproduction method which uses blind source separation and head-related transfer function is proposed. In the proposed system, multichannel acoustic signals captured at the distant microphones are encoded to a set of location/signal pairs of virtual sound sources based on frequency-domain ICA. After estimating the locations and the signals of the virtual sources, by convolving the controlled acoustic transfer functions with each signal, the spatial sound at the selected point is constructed. In the evaluation, the sound field made by 6 sound sources is captured using 48 distant microphones and is encoded into set of virtual sound sources. Subjective evaluation shows that there is no significant difference between natural and reconstructed sound when more than 6 virtual sources are used. Therefore the effectiveness of the encoding algorithm as well as the virtual source representation is confirmed.

Index Terms— Acoustic arrays, Acoustic fields, Acoustic beam steering, Array signal processing, Spatial filters

1. INTRODUCTION

Recently, as an extension of multi-viewpoint image processing, free viewpoint TV (FTV) systems [1] that can generate the scene at an arbitrarily selected viewpoint has become an issue in MPEG standardization [2]. 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.

The SLP audio is a spatial sound reproduction system characterized by four requirements: 1) microphones should be placed at the distant location from sound sources, 2) the number and the locations of the sound sources are unknown, 3) each sound source may move independently and 4) no special equipment is needed in the decoding side. Therefore, simply applying the existing spatial audio reproduction method, such as binaural recording [3] and transaural audio [4] by boundary surface controlling with speaker array [5], does not work well in SLP audio.

In the previous work [6], we evaluated an SLP audio system combining blind source separation (BSS) and binaural audio with a head-related transfer function (HRTF). In that system, BSS is used for separating the mixtures of signals, recorded at distant microphones, into independent source signals. Then a spatial impression is added to the signals 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 them with the transfer functions and remixing, natural spatial sound images can be reconstructed.

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

The extension of the SLP algorithm mainly consists of three parts. The first part is finding virtual sound sources. Since accurate identification of real sound sources is not necessary in SLP audio, e.g. no need to discriminate the closely located sound sources, we estimate the rough number of sound sources based on the subspace analysis of the spatial correlation matrix [7]. 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 signals. Since we use frequency-domain ICA (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. The clustering is performed based on the acoustic transfer functions from the position of the virtual source signal to microphones, that is calculated from the pseudo inverse of the separation matrix. The reconstructed signal from the group of virtual signals, which we call local signal mixture, represents either a signal of one source or mixture of different source signals located in close positions.

The third part is determining the reference location of the local signal mixture. In this part, 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 encode multiple microphone signals into a set of virtual sound source information, i.e., the location and the associated signal, which is the natural generalization of typical 3D sound field representation. After encoding, therefore, the local sound field at the selected listening point is flexibly presented. In this study a binaural system based on a HRTF is used.

In the rest of the paper, the basic idea of the SLP audio using BSS is described in Section 2. In Section 3, the proposed algorithm is detailed. After showing an experimental evaluation in Section 4, we conclude the paper in Section 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 corresponding source signals, i.e.,

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

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

$y(t)$ can be calculated by,

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

or in the frequency domain

$$Y(\omega) = \sum_{n=1}^{N} H(\mathbf{r}_n, \mathbf{r}^{(R)}) \cdot S_n(\omega), \tag{2}$$

when the acoustic transfer function between r_α and r_β is given by $h(\mathbf{r}_\alpha, \mathbf{r}_\beta)$. Typically in the binaural audio case, a 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 encoding the multichannel signals captured through M distant microphones into the source information Ω.

Potentially, BSS can be used for a part of the encoding by finding a set of independent signals $\hat{\mathbf{s}}(t)$. In particular, the frequency-domain ICA [8] combined with advanced methods for solving permutation ambiguity [9] is powerful under realistic acoustic conditions. However, it is known that the assumption on the number of sources is crucial in BSS therefore, 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 an SLP audio using BSS [6] under the assumption of prior knowledge of the number and locations of the sound sources. Through the experiment, we found that the imperfect separation does not cause serious problems in an SLP audio application because source signals are remixed in the target signal anyway. Therefore in order to realize the SLP audio system, we extend the BSS algorithm so as to operate it without any prior knowledge of sound sources, and build an encoding algorithm that converts the multi-channel signals into virtual sound source information.

3. ALGORITHM

3.1. Estimating virtual source signals

Since the number of sound sources is unknown, we first estimate the rough number of the sound sources by subspace analysis on the spatial correlation matrix [10]

$$\mathbf{R}(\omega) = E\{\mathbf{X}(\omega)\mathbf{X}(\omega)^H\}, \tag{3}$$

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 denotes the conjugate transpose and the expectation operations, respectively. By decomposing the $\mathbf{R}(\omega)$ 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 the 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, overall performance is not so sensitive to the accuracy of the estimate because the most of the signals are remixed in the target signal. [1] $\mathbf{\Lambda}'$ denotes the truncated version of diagonal eigen value matrix.

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

$$\mathbf{Z}(\omega) = \left(\mathbf{\Lambda}'(\omega)\right)^{-1/2} \mathbf{V}'(\omega)\mathbf{X}(\omega). \tag{4}$$

[1] When Q is overestimated, the echoes of the original signal are likely identified as independent sources.

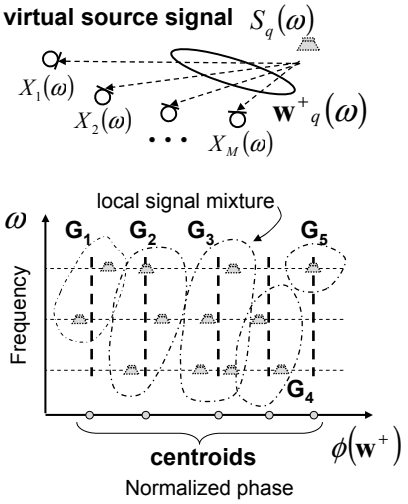

Fig. 1. Virtual source signals, their groups and centroids. A column vector of the pseudo inverse of the separation matrix, i.e., $\mathbf{w}_q^+(\omega)$ represents acoustic transfer functions from the q^{th} virtual source to microphones at the frequency ω. (top) Grouping transfer function vectors across the frequencies and combining the corresponding signal components give mixture of closely located sound signals. In this case, $Q = 4$ virtual sources are clustered into $K = 5$ clusters.(bottom)

The iterative learning rule below [11, 12] is used for estimating the separation matrix $\mathbf{U}(\omega)$ for the subspace signal $\mathbf{Z}(\omega)$:

$$\mathbf{U}_{t+1} = \mathbf{U}_t + \mu \cdot \text{off-diag}\{E[\varphi(\mathbf{Z})\mathbf{Z}^H]\}\mathbf{U}_t, \tag{5}$$

where $\varphi(z) = \tanh(\gamma \cdot \Re(z)) + j \cdot \tanh(\gamma \cdot \Im(z))$ denotes the activating function. Finally, Q independent signals $\mathbf{S}(\omega)$, which we call virtual source signals, can be calculated for each frequency bin by

$$\mathbf{S} = \mathbf{UZ} = \mathbf{U}\left(\Lambda'\right)^{-1/2}\mathbf{V}'\mathbf{X} = \mathbf{WX}. \tag{6}$$

Thus, the separation matrix for the original microphone signals is given by

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

Note that we omit the frequency index, (ω), from $\mathbf{S}(\omega)$, $\mathbf{U}(\omega)$, $\mathbf{V}(\omega)$, $\mathbf{W}(\omega)$, $\mathbf{X}(\omega)$, $\mathbf{Z}(\omega)$ and $\mathbf{\Lambda}(\omega)$ in equations (5) through (7).

3.2. Grouping Virtual Signal Components

The pseudo-inverse of the 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., $\mathbf{w}_q^+(\omega) = [w_{1,q}^+(\omega), \cdots, w_{M,q}^+(\omega)]^T$, for the frequency ω. The phase component of that vector contains geometrical information of the virtual sources. In [9, 13], 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 spectral components of the virtual source signals, i.e., $S_1(\omega), \cdots, S_Q(\omega)$, have permutation ambiguity across frequency

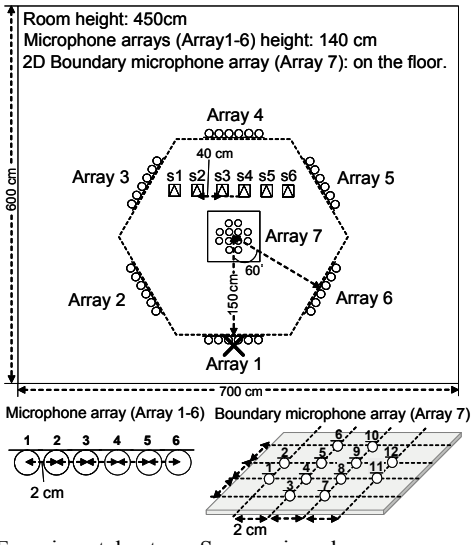

Fig. 2. Experimental setup. Seven microphone arrays surrounded the six loudspeakers in a linear arrangement. One of the arrays is 2D boundary array and located on the floor. The other six are linear arrays and located at a height of 1.4 m.

indices, we try to group closely located virtual sources and reconstruct the mixture of the signals of those virtual sources through clustering as follows.

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

$$\phi(\mathbf{w}_q^+(\omega)) = [\exp(j\xi_{q,1}), \cdots, \exp(j\xi_{q,M})]. \quad (8)$$

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

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

where d is the array size of which the m^{th} microphone located and $\arg(\cdot)$ operation calculates the relative phase angle in that array. As we see in the experiment below, we assume that each microphone is arranged as an element of one of L arrays. We denote a set of microphones included in the l^{th} array by $\theta(l)$. By applying the $\phi(\cdot)$, we can cancel the frequency dependency from $\mathbf{w}_q^+(\omega)$ therefore, we can cluster the virtual sources across frequency bins as shown in Figure 1.

The similarity between phase vectors is defined by the sum of scalar products over arrays, i.e.,

$$\sum_{l=1}^{L} \left| \sum_{m\in\theta(l)} \phi(w_{\alpha,m}^+(\omega))^* \cdot \phi(w_{\beta,m}^+(\omega)) \right|, \quad (9)$$

because this measure is robust to the constant phase shift due to the ambiguity of the array position. $(\cdot)^*$ represents complex conjugate. Based on this similarity measure, we cluster $Q \times D$ transfer function vectors into K clusters. D denotes the number of frequency bins. It should be noted that K can be more than Q.

Table 1. 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 clusters, K	2, 6, 10, 14
Number of virtual sources, Q	2, 6, 10

Denoting the clustering results in which the 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

$$\hat{S}_k(\omega) = \sum_{q=1}^{Q} \delta_{k,g(q,\omega)} \mathbf{w}_q \cdot \mathbf{X}(\omega), \quad (10)$$

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

3.3. Location estimation

The reference location of the 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 the location $\mathbf{r}$ is given by

$$\mathbf{a}_l(\mathbf{r}) = [\exp(j\frac{\pi|\mathbf{r}_{l,1}^{(r)} - \mathbf{r}|}{2d_l}), \cdots, \exp(j\frac{\pi|\mathbf{r}_{l,\theta(l)}^{(r)} - \mathbf{r}|}{2d_l})], \quad (11)$$

where $\mathbf{r}_{l,i}^{(r)}$ represents the position of the i^{th} element of the l^{th} array. 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 can be calculated. We search for the location where the similarity becomes largest as the reference position of the local signal mixture,

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

Finally, the 3D sound filed representation $\hat{\Omega} = \{\hat{\mathbf{r}}_k, \hat{s}_k(t)\}_{k=1,\cdots,K}$ is obtained.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

Figure 2 shows the experimental setup for the acoustic systems. Six 6-element arrays and a 12-element array are arranged so as to surround six loudspeakers. All of the 48 sensors are omni-directional microphones (SONY-ECM77B). 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

Table 2. Collection list with organizational sound sources

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

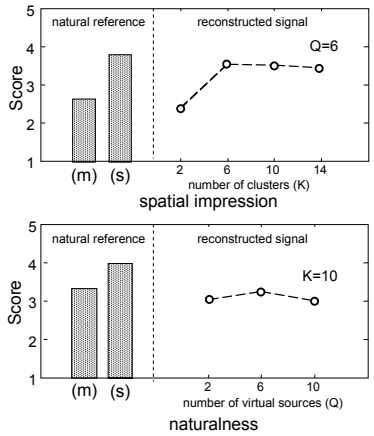

Fig. 3. Experimental evaluation results. Mean scores of spatial impression (top) and naturalness (bottom) compared to the natural references.

40 kHz. As for the test signals, we recorded speech, a popular music piece and a classical music piece as listed in Table 2. In order to add a spatial impression to the estimated local signal mixtures, the measured HRTFs are used after interpolation [14]. Other conditions are listed in Table 1.

4.2. Evaluation Results

A preliminary evaluation of the reconstructed sound is performed. Subjective scores on "spatial impression" and "naturalness" of the generated binaural signals for the center front position (shown by 'x' in Figure 2) were measured. Six male students are involved in the evaluation. In addition to the generated SLP sounds, the subjects evaluated two reference sounds, i.e., monaural (m) and stereo (s) audio sounds recorded at the same position. We have tested the proposed method by changing the number of virtual sources (Q) from 2 to 10 and the number of local signal mixtures (K) from 2 to 14 in order to evaluate the effect of these parameters on the spatial impression and the naturalness of the signal, respectively. The averaged scores are plotted in Figure 3.

The spatial impression of the generated signal is as high as that of the stereo sound when more than 6 local signal mixtures are used. Under that condition, the average MOS is better than the recorded monaural signal by 0.8 point. On the other hand, the naturalness of the reconstructed signal is lower than that of the recorded monaural signal by 0.3.

From the results, we confirmed that the proposed encoding method as well as the 3D sound field representation based on virtual sound sources are effective for the SLP audio system.

5. SUMMARY AND FUTURE WORKS

In this paper we proposed and evaluated a new spatial audio system, 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 have developed a method to encode the multichannel signal recorded at the distant positions into this representation based on the BSS technologies.

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

There are many other issues that need further study including the optimal array arrangement and performance under more reverberant and/or noisy conditions. Among those problems, we think that dealing with a non-stationary sound field, e.g., moving sources, is one of the most important future works.

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

6. 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, 2002.

[2] "ISO/IEC JTC1/SC29/WG11 (N9168)," July 2007(Lausanne, Switzerland).

[3] J.Blauert, "Spatial hearing (revised ed.)," pp. 372–392, 1996.

[4] J. Bauck and D. H. Cooper, "Generalized transaural stereo and applications," *J. Audio Eng.Soc.*, vol. 44, no. 9, pp. 683–705, 1996.

[5] S.Ise, "The boundary surface control principle and its applications," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1656–1664, 2005.

[6] K.Niwa, T.Nishino, and K.Takeda, "Development of selectable viewpoint and listening point system for musical performance," *ICA2007*, PPA-06-011, 2007.

[7] 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. on Speech and Audio Proc.*, vol. 11, no. 3, pp. 204–215, 2003.

[8] P.Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1–3, pp. 21–34, 1998.

[9] 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 Sig. Proc.*, pp. 1135–1146, 2003.

[10] M.Wax and T.Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[11] A.Bell and T.Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[12] 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, 1999.

[13] 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, 2006.

[14] Takanori Nishino, Sumie Mase, Shoji Kajita, Kazuya Takeda, and Fumitada Itakura, "Interpolating hrtf for auditory virtual reality," *The Third Joint Meeting ASA and ASJ, (1pSP6)*, pp. 1261–1266, 1996.