Blind Source Separation with Distributed Microphone Pairs Using Permutation Correction by Intra-Pair TDOA Clustering

Takuma Ono, Shigeki Miyabe, Nobutaka Ono and Shigeki Sagayama
Graduate School of Information Science and Technology, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{tono, miyabe, onono, sagayama}@hil.t.u-tokyo.ac.jp

***Abstract*—In this paper, we present a novel framework of distributed microphone array for blind source separation (BSS), where stereo microphones or proximately-placed microphone pairs are distributed. Unlike distributing all microphones individually, the time difference of arrival (TDOA) in the paired channels can be robustly estimated without suffering spatial aliasing. Based on it, sound sources are separated by the frequency-domain independent component analysis (ICA) with the permutation correction by clustering the intra-pair TDOAs. The experimental results in real reverberant environment are also shown.**

I. Introduction

Blind source separation (BSS) is a technique to recover original source signals from mixtures without any mixing conditions, which has been actively investigated in array signal processing field. In particular, the frequency-domain independent component analysis (ICA) is one of the standard approach of BSS in convolutive mixing [1], [2] and closely-arrayed microphones are preferably used to avoid spatial aliasing [3]–[5]. However, setting closely-spaced microphone array in one position has limitation because we cannot locate the microphone array near to all sources to earn high signal-to-noise ratio, and some of sources can stand in the same direction.

In contrast, using spatially-distributed microphones would be more suitable with separating wide-spread sources. Improving separation performance with combining distributed multiple recording devices is also attractive scenario. Recently, the framework of distributed microphones has been investigated for speech recognition system [6] and teleconferencing technology [7], and it will facilitate speech-based man-machine interaction in prospective system such as "smart room" [8]. Several methods for localization or separation using distributed microphones has been also proposed [7], [9]–[11].

Since frequency-domain ICA finds independent components at every frequency bands, permutation inconsistency generally remains. A robust approach to solving permutation is clustering of the relative positions of sources and microphones, especially by exploiting time differences of arrival (TDOAs) [4]. However, in the distributed microphone array scenario, distances between microphones are often too large to estimate reliable TDOAs because of the spatial aliasing. Although there are several reports about utilizing magnitude ratios for the clustering [7], [10], the level difference is not as robust as TDOA for arbitrary arrangement of sources and microphones. Another approach is to use temporal structure of signals [12]. It is known that this approach alone is less robust than TDOA-based one, but integration with TDOA is effective [4]. The integration with temporal structure is out of focus in this paper.

In this paper, we present a novel framework of distributed microphone array for BSS, where stereo microphones or proximately-placed microphone pairs are distributed. The concept of distributing subarrays, which consist of closely spaced microphones for avoiding spatial aliasing, has been also used in [13], [14]. We focus on specifically distributing paired microphones because many recording devices have stereo channels and it would be very suitable with the context of the distributed microphone array. Based on it, sound sources are separated by the frequency-domain ICA with the permutation correction by clustering the intra-pair TDOAs since the TDOA in the paired channels can be robustly estimated. In addition, we also discuss using multiple stereo recording devices, where inter-pair channels are asynchronized. Finally, we compare the separation performance of the proposed method and several existing methods by experimental results in real environment.

II. BSS with distributed microphone pairs

A. Problem formulation

Let's consider that stereo microphones, which are pairs of closely-spaced two microphones, are distributed and N sources are observed by them. Let $M = 2M_p$ denote the number of microphones and the $(2i-1)$th and the $2i$th microphone be paired ($i = 1, \cdots, M_p$). We assume the overdetermined case ($M > N$) and the number of sources (N) is known a priori. First, suppose that all channels are synchronized. The problem here is to recover N sources from M observations.

B. Frequency-domain ICA

First, we summarize a typical frequency-domain ICA in the overdetermined case. In time-frequency representation, the observed signal $\mathbf{X}(\tau,\omega) = [X_1(\tau,\omega) \ \cdots \ X_M(\tau,\omega)]^T$, where $[\ ]^T$ denotes vector transpose, is denoted as

$$\mathbf{X}(\tau,\omega) = \begin{bmatrix} A_{11}(\tau,\omega) & \cdots & A_{1N}(\tau,\omega) \\ \vdots & \ddots & \vdots \\ A_{M1}(\tau,\omega) & \cdots & A_{MN}(\tau,\omega) \end{bmatrix} \mathbf{S}(\tau,\omega), \quad (1)$$

where n and m are source and microphone indices, respectively, τ is a frame index, $A_{mn}(\tau,\omega)$ is the frequency characteristic from the nth source to the mth microphone and $\mathbf{S}(\tau,\omega) = [S_1(\tau,\omega) \ \cdots \ S_N(\tau,\omega)]^T$ is the source signal.

In the overdetermined case, dimension reduction is first applied using principle component analysis (PCA) [15], which is performed by transforming the observed signal into the subspace signal as

$$\begin{aligned}\tilde{\mathbf{X}}(\tau,\omega) &= [\tilde{X}_1(\tau,\omega)\ \cdots\ \tilde{X}_N(\tau,\omega)]^T \\ &= W_{PCA}(\omega)\mathbf{X}(\tau,\omega),\end{aligned} \tag{2}$$

where W_{PCA} is a $M \times N$ matrix and obtained by the N eigenvectors corresponding to the N largest eigenvalues of the covariance matrix of $\mathbf{X}(\tau,\omega)$.

Then, source signals are estimated as

$$\mathbf{Y}(\tau,\omega) = W(\omega)\tilde{\mathbf{X}}(\tau,\omega). \tag{3}$$

The demixing matrix $W(\omega)$ is iteratively estimated by the update rule with natural gradient [16]:

$$W(\omega) \leftarrow W(\omega) + \mu(I - \mathrm{E}[\mathbf{g}(\mathbf{Y}(\tau,\omega))\mathbf{Y}^H(\tau,\omega)])W(\omega), \tag{4}$$

where μ is a step size, $\mathbf{g}$ is a nonlinear function, and $\mathrm{E}[\cdot]$ denotes expectation operation, which is practically replaced by time average over all frames.

Generally in frequency-domain ICA, scale ambiguity and permutation inconsistency remain. The scale is usually determined by projection back [12] and we also follow it. In the next section, we discuss the permutation correction as the main issue of this paper.

III. Permutation Correction by Clustering Intra-pair TDOA Vectors

A. Estimation of Intra-pair TDOA

TDOA-based permutation correction is one of the popular approach in overdetermined BSS, and it has been also well used in underdetermined BSS [17]. However, utilizing TDOA is not straightforward in the context of distributed microphone array. When microphones are distantly located, the correlation between their channels degrades and estimating TDOA becomes difficult. The spatial aliasing is also significant.

Our approach is to distribute paired microphones and using their intra-pair TDOAs. After frequency-domain ICA, with minimizing $\mathrm{E}\{\|A(\omega)\mathbf{Y}(\tau,\omega) - \mathbf{X}(\tau,\omega)\|\}^2$, the mixing matrix can be estimated as

$$\hat{A}(\omega) = \mathrm{E}\{\mathbf{X}(\tau,\omega)\mathbf{Y}(\tau,\omega)^H\}(\mathrm{E}\{\mathbf{Y}(\tau,\omega)\mathbf{Y}^H(\tau,\omega)\})^{-1}, \tag{5}$$

where $\hat{A}(\omega)$ is a $M \times N$ matrix and the nth row vector gives a steering vector for the nth source. Using it, the TDOA for the nth source within the ith pair of microphone pairs is obtained as

$$d_n^i(\omega) = \{\arg(\hat{A}_{(2i-1)n}(\omega)) - \arg(\hat{A}_{(2i)n}(\omega))\}/\omega, \tag{6}$$

where $\hat{A}_{mn}(\omega)$ is the (m,n)th element of $\hat{A}(\omega)$.

Here we define an intra-pair TDOA vector in the kth frequency:

$$\mathbf{d}_{kn} = [d_n^1(\omega_k), \cdots, d_n^{M_p}(\omega_k)]^T. \tag{7}$$

This vector gives us definite clue for the permutation correction even in distributed microphone array since it should be obtained without spatial aliasing due to the close distance of the paired microphones.

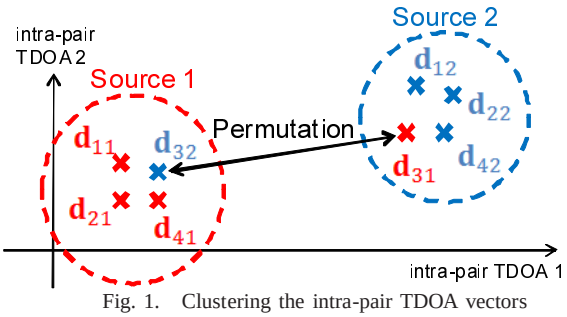

Fig. 1. Clustering the intra-pair TDOA vectors

B. Clustering Intra-pair TDOA vectors

Since the intra-pair TDOA vectors ideally do not depend on frequency but location of sources and microphones, the permutation could be corrected by grouping them, and for that, *k*-means algorithm can be applied in the M_p-dimensional euclidean space. In this clustering, considering that the TDOA at the frequency band with more significant energy is more reliable, we introduce weighting with reliability of each estimation. Let $\hat{S}_{mn}(\tau,\omega)$ be the estimation of the nth source at the mth microphone, which is obtained by Projection Back. Based on them, we define the weight coefficient as

$$\boldsymbol{\gamma}_{kn} = [\gamma_{kn}^1, \cdots, \gamma_{kn}^{M_p}]^T, \tag{8}$$

$$\gamma_{kn}^i = \mathrm{E}[|\hat{S}_{(2i-1)n}(\tau,\omega)| + |\hat{S}_{(2i)n}(\tau,\omega)|], \tag{9}$$

where $\boldsymbol{\gamma}_{kn}$ is the time average over all frames of the nth source amplitude at the kth frequency. Then the weighted square sum of the distances from the mean of all entries is minimized by the following iterative updates for all the sources and microphone pairs:

$$p_k(n) \leftarrow \underset{p_k(n)}{\mathrm{argmin}} \sum_k \sum_{n=1}^{N} \sum_{i=1}^{M_p} \gamma_{kp_k(n)}^i |d_{kp_k(n)}^i - m_n^i|^2 \tag{10}$$

$$m_n^i = \frac{\sum_k \gamma_{kp_k(n)}^i d_{kp_k(n)}^i}{\sum_k \gamma_{kp_k(n)}^i} \qquad (i = 1, \cdots, M_p), \tag{11}$$

where ^{i} of $\gamma_{kp_k(n)}^i, d_{kp_k(n)}^i$ is the ith row of a vector, $p_k(n)$ is a permutation pattern and m_n^i is the centroid of the nth source within the ith microphone pair, which represents the estimation of the intra-pair TDOA. Note that m_n^is take different values for different i depending on the location and the orientation of each pair. Grouping the intra-pair TDOAs in (10) and determination of the centroids in (11) corresponds to the permutation correction and intra-pair TDOA estimation, respectively.

IV. BSS with Multiple Stereo Recording Devices

In the previous section, we suppose that all microphones are synchronized. Next, in this section let's discuss BSS with multiple stereo recording devices, which means we assume that intra-pair channels are synchronized while inter-pair channels are asynchronized. In this paper, we also assume that the sampling frequencies of all devices are identical and their mismatch is negligible as a simple case.

In this case, observed signals have different time offsets. However, the demixing filter could be obtained by frequency-domain ICA if each channel is aligned approximatively and

TABLE I
EXPERIMENTAL CONDITION

number of sources	3
number of microphones	4
room size	6.3 m ×7.7 m ×2.7 m
reverberation time	300 ms
sampling rate	16 kHz
frame shift	75% overlap
ICA algorithm	Infomax

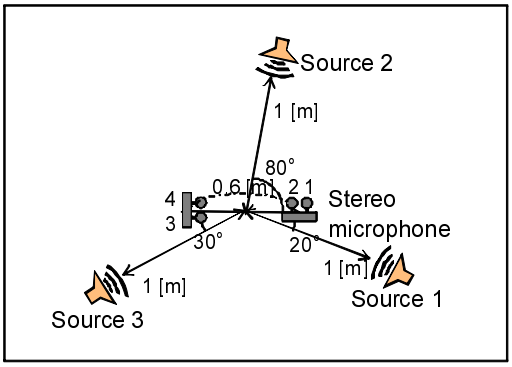

Fig. 2. Arrangement of sources and microphones in the experiment

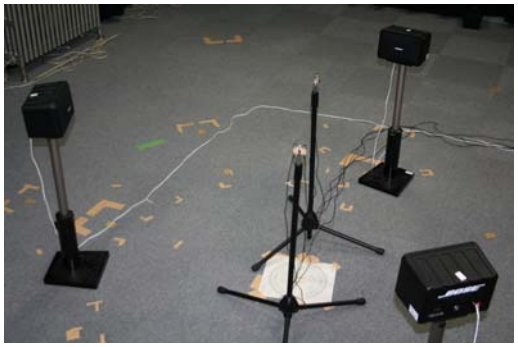

Fig. 3. A photo of real acoustic environment

the apparent TDOAs between channels for all sources are adequately small compared with the frame length. The rough alignment can be performed by shifting $x_i(t)$ by T_{i1} $(i = 2, \cdot, M)$ where

$$T_{i1} = \underset{t_m}{\operatorname{argmin}} R_{i1}(t_m), \tag{12}$$

and $R_{i1}(t_m)$ denotes the cross correlation between $x_j(t)$ and $x_1(t)$. Note that the permutation correction based on the intra-pair TDOA vector is not affected at all even when the intra-pair channels are exactly synchronized.

V. EXPERIMENTAL RESULTS

In this section, we evaluate the separation performance of the proposed method in real acoustic environment. Intending to simulate multiple recording in a meeting, two microphone pairs were placed inside three loudspeakers in a room with $T_{60} = 300$ ms as shown in Fig. 2 and Fig. 3, where the intra-pair and the inter-pair distances were 3 cm and 60 cm, respectively. The male and the female speech was used as source signals. Other experimental conditions were shown in Table I.

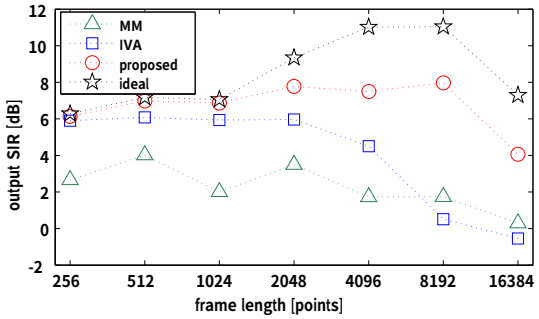

Fig. 4. Output SIRs for different frame lengths for 6s-length signal

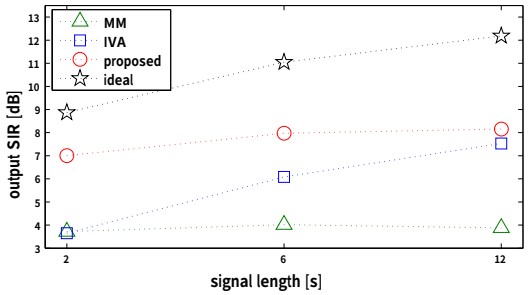

Fig. 5. Output SIRs for different signal lengths. Note that an appropriate frame length is chosen for each condition

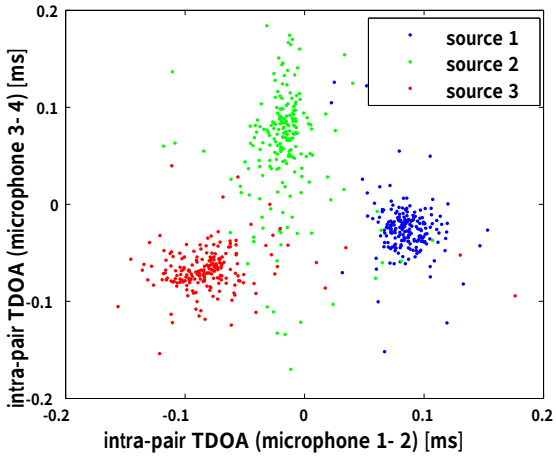

Fig. 6. Example of the clustered intra-pair TDOA vectors

In the first experiment, instead of using really asynchronous recordings, we used synchronous recordings (i.e., which means all microphones were connected to an audio board with synchronous inputs), and gave random time offsets between inter-pair channels for quantitative evaluation. We roughly aligned the simulatedly-asynchronous signals as discussed in section IV, applied the infomax-type frequency-domain ICA, and then, corrected the permutation by our method presented in section III.

Confined to usable approaches with distributed microphones, our method was compared with

- MM: the infomax-type frequency-domain ICA with a magnitude-based permutation correction (maximum-magnitude) in [7],

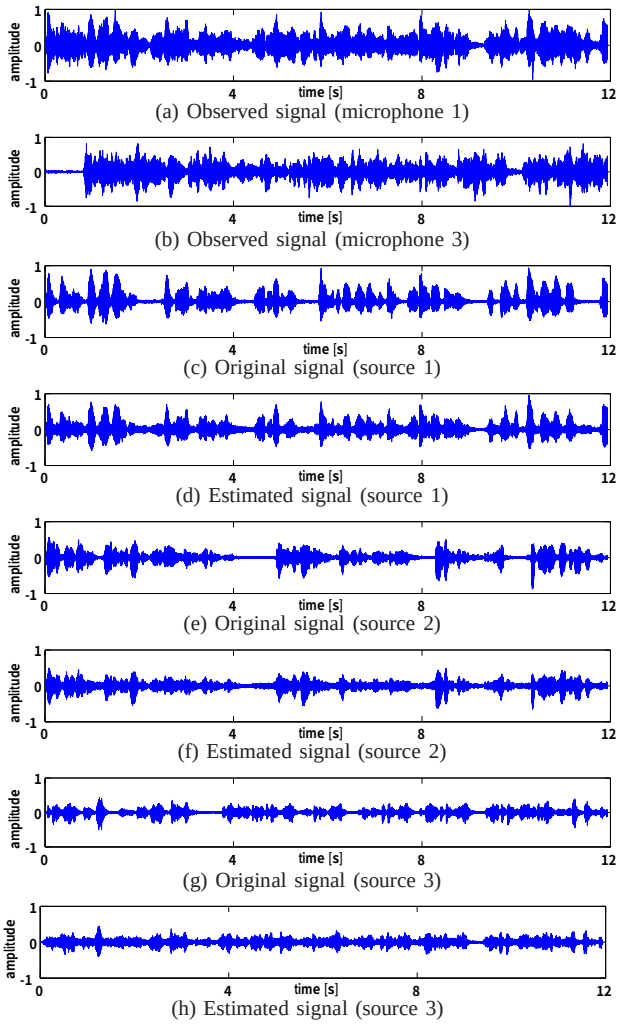

Fig. 7. Results by using multiple stereo recording devices

- IVA: a permutation-free blind source separation (independent vector analysis) in [18], and
- ideal: the infomax-type frequency-domain ICA with the ideal permutation correction to maximize SIR (Signal-to-Interference Ratio) where the original signals are given as reference.

The separation performance was evaluated by the gain of SIR.

Fig. 4 and Fig. 5 show the separation performance in the different frame length and the different signal length, respectively. In all conditions, the proposed method showed superior performance to other two methods. The difference between IVA and our method is small when the frame length is short or the signal length is long, but our method is especially strong for the long frame length or the short signal length. This nature would be very desirable in block-wise processing, which is necessary for separating moving sources or compensating drifted time offset between channels. We also show an example of clustered intra-pair TDOAs in the frequency band lower than 4000 Hz for 12s-length real recorded signals in Fig. 6. We can see that the intra-pair TDOA vectors are appropriately grouped in the 2-dimensional euclidean space.

In the second experiment, we used two asynchronous stereo recording devices (digital voice recorders, SANYO ICR-PS603RM), and other experimental conditions were the same as the first experiment. Fig. 7 shows the recorded signals by the different devices (microphone 1 and 3), the pre-recorded sources at microphone 1 and the estimated sources from the mixtures by the proposed method. We can see that the proposed method worked well and the sources were separated even when the recordings were really asynchronous and had different time offsets.

VI. Conclusion

In this paper, the framework of distributing microphone pairs for BSS based on the frequency-domain ICA and the permutation correction suitable with it were presented. In the proposed method, the permutation is corrected based on clustering TDOAs within the paired channels. The BSS using multiple stereo recording devices was also discussed. In the experiments, the proposed method showed the better separation performance than other existing methods and it also worked well in multiple stereo recordings.

References

[1] A. Hyvärinen, J. Karhumen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.
[2] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," *Proc. ICA*, pp. 365–371, 1999.
[3] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Near-Field Frequency Domain Blind Source Separation for Convolutive Mixtures," *Proc. ICASSP*, pp.49–52, 2004.
[4] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *Proc. ICASSP*, pp. 3140–3143, 2000.
[5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust approach to the permutation problem of frequency-domain blind source separation," *Proc. ICASSP*, pp. 381–384, 2003.
[6] Y. Zhao, S. Shin, E. Robledo-Arnuncio, and B.H. Juang, "A study on recognizing distorted speech over local distributed transducer networks," *Proc. ICASSP*, pp. 4181–4184, 2009.
[7] J.P. Dmochowski, Z. Liu, and P.A. Chou, "Blind source separation in a distributed microphone meeting environment for improved teleconferencing," *Proc. ICASSP*, pp. 89–92, 2009.
[8] C. Busso, S. Hernanz, C. Chu, S. Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S. Narayanan, "Smart room: Participant and speaker localization and identification," *Proc. ICASSP*, pp. 1117–1120, 2005.
[9] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," *Proc. WASPAA*, pp.161–164, 2009.
[10] E. Robledo-Arnuncio and B.H. Juang, "Blind source separation of acoustic mixtures with distributed microphones," *Proc. ICASSP*, pp. 949–952, 2007.
[11] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," *Proc. IWAENC*, 2008.
[12] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.
[13] 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," *Proc. ICASSP*, pp.181–184, 2008.
[14] F. Nesta and M. Omologo, "Cooperative Wiener-ICA for source localization and separation by distributed microphone arrays," *Proc. ICASSP*, pp.181–184, 2010.
[15] S. Winter, H. Sawada, and S. Makino, "Geometrical interpretation of the PCA subspace approach for overdetermined blind source separation," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–11, 2006.
[16] S. Amari, "Natural gradient works efficiently in learning," *Neural Compututation*, vol. 10, pp. 251–276, 1998.
[17] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, pp. 1833–1847, 2007.
[18] T. Kim, H.T. Attias, S. Lee, and T. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 70–79, 2007.