FAST SOURCE SEPARATION BASED ON SELECTION OF EFFECTIVE TEMPORAL FRAMES

Yusuke Mizuno[1], *Kazunobu Kondo*[1,2], *Takanori Nishino*[3], *Norihide Kitaoka*[1], *and Kazuya Takeda*[1]

[1]Graduate School of Information Science, Nagoya University, Nagoya, Aichi, Japan, 464–8603

[2]Corporate Research & Development Center, Yamaha Corporation, Shizuoka, Japan, 438–0192

[3]Graduate School of Engineering, Mie University, Tsu, Mie, Japan, 514–8507

ABSTRACT

A faster computational method for performing frequency domain independent component analysis (FDICA) using a dodecahedral microphone array is proposed. Source separation with FDICA uses the spectrum of observed signals and estimates separation filters for each frequency. However, this technique is complex and requires high computational resources. In this paper, a method of selecting temporal frames which are effective for training the separation filters is proposed and evaluated. The log power spectrum and the kurtosis of amplitude distribution are employed as selection criteria. Performance was evaluated by comparing signal-to-interference performance with that of the conventional method. Experimental results showed that the proposed method reduced computation to 17.1 % of that required by the conventional method, and that separation performance of the proposed method is superior. Therefore, the proposed method can achieve faster computation with lower computational complexity, and its effectiveness can be confirmed.

***Index Terms*—** Dodecahedral microphone array, Frequency domain independent component analysis, Computational complexity reduction, Signal-to-interference improvement

1. INTRODUCTION

Frequency domain independent component analysis (FDICA) [1], which achieves source separation with only the assumption of independence between each source, is a blind source separation (BSS) method that extracts objective source signals without prior information. Although many studies on source separation target two sources, several sources often need to be separated in actual environments, because many speakers or noises are likely.

A source separation method using a dodecahedral microphone array (DHMA), as shown in Fig.1 is proposed[2]. Sixty microphones are set on the DHMA. This array can separate many sources with high accuracy, but a problem occurs because of the extremely high computational complexity involved. FDICA source separation requires huge computational resources, mainly for iterative learning on each frequency. The degree of computational complexity depends on the number of iterations, frequency bins, temporal frames and separated signals. Since estimation of the separation matrix requires a nonlinear correlation matrix between separated signals, computational complexity increases by the second power of the number of separated signals. A high-convergence algorithm combining beamforming and ICA [3, 4] has been proposed and the number of iterations is reduced. However, the computational complexity for constructing the beamformer increases according to the squared value of the number of sources. Therefore, another method that trains separation matrices on partial frequencies[5] has also been proposed, and reduction of computational complexity is achieved. However, calculation of frequency selection criteria, and estimation of substitute separation matrices for frequencies that are not selected, still required large computational resources.

Based on the above, this paper examines selection of temporal frames. Short time Fourier transforms (STFT) of observed signals are used for selecting temporal frames. If more effective temporal frames are selected, separation matrices can be trained using only these selected temporal frames, contributing to the reduction of computational complexity. In a previous study[5], the determinant of the spatial correlation matrix using observed signals was employed as the selection criteria, because the log power spectrum and the mixture number can be utilized. However, spatial correlation matrices

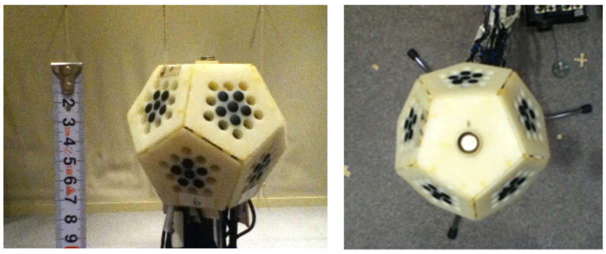


Fig. 1. Dodecahedral microphone array[2].

which are frequency dependent cannot be used as selection criteria for temporal frames. In this paper, in addition to the observed log power spectrum, kurtosis is employed as a criteria for evaluating mixture numbers. Because non-Gaussianity of temporal frames is decreased when many source signals are mixed, kurtosis of amplitude distribution, which is a method of evaluating non-Gaussianity, can be employed on each temporal frame to evaluate the mixture number. Although the method described above increases the computational complexity of the calculation of selection criteria, overall computational complexity is reduced considerably. Additionally, a trained separation matrix can be applied to temporal frames which are not selected for training, and the computational complexity required for interpolation of the un-trained frames is eliminated.

This paper consists of four sections. In section 2, a global flow chart of the proposed temporal frame selection method is provided, and each step of the process is then described. In section 3, experimental evaluation of the proposed method, and comparison with the conventional method, are described. Section 4 concludes this paper.

2. PROPOSED METHOD

A global flow chart of 60 channel source separation with DHMA is shown in Fig.2. $\mathbf{X}(f,\tau)$ is given by STFT of 60 observed signals $\mathbf{x}(t)$ from the DHMA. The dimension of the observed signals is reduced using a subspace method[6]. Temporal frames are selected and the separation matrices are estimated by ICA iterative learning, using only the selected temporal frames. The permutation problem with FDICA is solved when we assume that the pseudo-inverse matrix of the separation matrix represents the propagation characteristics of signals from sources to microphones.

2.1. Subspace method

The dimensions of input and output signals are the same as in the ICA. Therefore, the dimension of the input signals must be reduced when the dimension of the output signals is larger. A subspace method employs eigenvalue decomposition of the spatial correlation matrix of observed signals as

$$\begin{aligned}\mathbf{R}_{xx}(f) &= E_\tau[\mathbf{X}(f,\tau)\mathbf{X}^H(f,\tau)] \\ &= \mathbf{V}(f)\mathbf{\Lambda}(f)\mathbf{V}(f)^T, \end{aligned} \tag{1}$$

where $(\cdot)^H$ represents conjugate transposition. $\mathbf{\Lambda}(f)$ is a diagonal matrix, and $\mathbf{V}(f)$ is composed of characteristic vector $\mathbf{v}(f)$ and

$$\mathbf{V}(f) = [\mathbf{v}_1(f), \ldots, \mathbf{v}_M(f)]. \tag{2}$$

The given characteristic vector $\mathbf{v}(f)$ is used only as a source of the number of desired separation signals, which is represented by $\mathbf{V}'(f)$. Likewise, the diagonal matrix $\mathbf{\Lambda}'(f)$ is a new matrix with a reduced number of dimensional characteristics, which is derived from a characteristic number diagonal matrix $\mathbf{\Lambda}(f)$. Finally, subspace signals are given as

$$\mathbf{Z}(f,\tau) = \mathbf{\Lambda}'(f)^{-1/2}\mathbf{V}'(f)\mathbf{X}(f,\tau). \tag{3}$$

2.2. Temporal frame selection

In order to reduce the computational complexity of two channel FDICA source separation, frequency band selection FDICA[5] has been proposed. Using this method, the determinant of the spatial correlation matrix, calculated from observed signals, is employed as the criteria for selecting frequency bands. Frequencies in which the determinant of the covariance matrix is large are selected. This is because such frequencies contain both signals' components and also because the log power spectrum is large. If there is only one source the determinant goes to zero. The separation matrix of a frequency which is not selected is substituted with the coefficient of beamforming, which results in lower computational complexity. In contrast, our proposed method involve a method of temporal frame selection.

Useful temporal frames are selected from the sequence of the spectrum using STFT, and a separation matrix is estimated using only the selected temporal frames. Using the previously described band selection FDICA method, the observed log power spectrum and mixture number are considered during selection. However, the observed log power spectrum and kurtosis of amplitude distribution are employed as the criteria for temporal frame selection. One proposal for es-

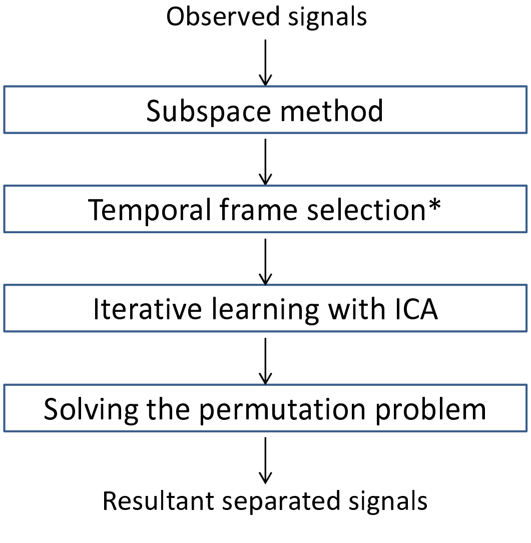

Observed signals
Subspace method
Temporal frame selection*
Iterative learning with ICA
Solving the permutation problem
Resultant separated signals

Fig. 2. FDICA source separation system that is employed in this paper. All steps are performed on the frequency domain. The step in the process which is modified by the proposed method is indicated by * .

timation of separation matrices with ICA is maximization of non-Gaussianity [7]. This idea is based on the central limit theorem. When random variables $x_1, \ldots, x_n$ follow mutually independent random distributions, the distributions $x = (x_1 + \cdots + x_n)$ become normal distributions if n is sufficiently large. Thus, when sources which are mutually independent mix, the amplitude distribution approaches a normal distribution and non-Gaussianity decreases. This non-Gaussianity can be evaluated by kurtosis. Kurtosis is defined as the fourth moment divided by the fourth power of the standard deviation, and as:

$$K = E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right], \tag{4}$$

where $E[\cdot]$ denotes expectation, μ is the mean and σ is the standard deviation. FDICA estimates the separation filters to maximize non-Gaussianity. This is because when a number of sources are mixed, the amplitude distribution closes in a Gaussian manner and non-Gaussianity diminishes. Thus the kurtosis also becomes diminished. For example, Fig.3 shows distribution and kurtosis of amplitude for each number of mixing sources. The mixture sources consist of different speech signals with a length of 4 seconds, mixed on the time domain. Since it has been confirmed that kurtosis becomes diminished when a number of sources are mixed, low kurtosis temporal frames are preferentially selected. Here, the kurtosis of each temporal frame is defined as the kurtosis of the samples in the window.

2.3. Iterative learning using ICA

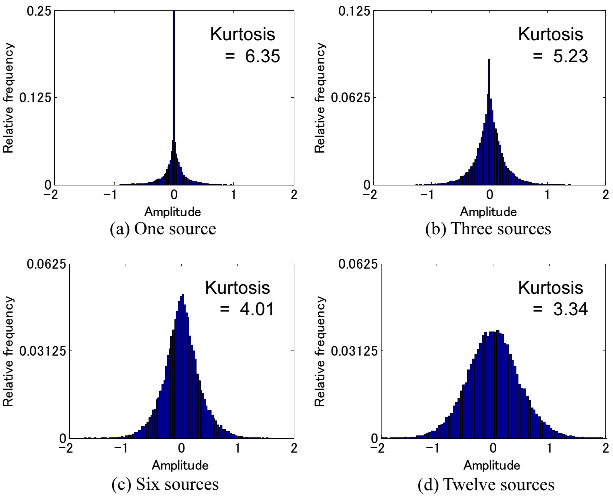
FDICA is a method of statistical analysis which employs only the assumption that source signals are mutually independent. Independence is a stronger characteristic than non-correlation and indicates non-linear non-correlation[7]. When s_1 and s_2 are mutually independent, non-correlation is satisfied. Additionally arbitrary non-linear transforms $\phi(s_1)$ and $\phi(s_2)$ are also not correlated. Thus, separation matrices are trained so that the covariance matrix of the separation signals and their non-linear transform, represented as:

$$\mathbf{R}_y(f) = E_\tau[\Phi(\mathbf{Y}(f,\tau))\mathbf{Y}^H(f,\tau)] \tag{5}$$

go to a diagonal matrix. $E_\tau[\cdot]$ denotes expectation about STFT frame index τ. To estimate separation matrices, natural gradient learning can be used as in:

$$\begin{aligned}\mathbf{W}(f) \leftarrow \alpha\big[&\mathrm{diag}(E_\tau[\Phi(\mathbf{Y}(f,\tau))\mathbf{Y}^H(f,\tau)]) \\ &- E_\tau[\Phi(\mathbf{Y}(f,\tau))\mathbf{Y}^H(f,\tau)]\big]\mathbf{W}(f),\end{aligned} \tag{6}$$

where α is the updating coefficient and $\mathrm{diag}(\cdot)$ denotes extraction of only the diagonal components. Estimation of the covariance matrix using separation signals $\mathbf{Y}$ and their non-linear transforms $\Phi(\mathbf{Y})$ is achieved by using the sample average, so a sufficient duration of observed signals is needed so that the covariance matrix can be estimated with accuracy. Statistical analysis using FDICA then captures the temporal sequence of the spectrum using the STFT of the sample for each frequency, and separation on each frequency can be achieved by using the covariance matrices of their non-linear transform. Thus it is possible to estimate the separation matrices with only selected temporal frames.



(a) One source
(b) Three sources
(c) Six sources
(d) Twelve sources

Fig. 3. Distribution and kurtosis of amplitude for various numbers of mixed sources.

2.4. Solving the permutation problem

FDICA estimates the separation matrices on each frequency, but the order of output signals between frequencies is irregular. This is called a permutation problem, and many methods for solving permutation problems have been proposed[8, 9]. Solutions can be divided into two groups. One type of solution uses the correlation between neighboring frequencies, and the other uses the arrival direction of the sources. In this paper, the permutation problem is solved based on arrival direction using the DHMA[2, 10]. The pseudo-inverse matrix of the separation matrix describes propagation characteristics from sources to microphones, and the permutation problem can be solved by clustering signals coming from the same direction. A number of features obtained using the DHMA are used for clustering propagation characteristics.

3. SOURCE SEPARATION EXPERIMENT

3.1. Experimental description

A source separation experiment using impulse responses from each source to each microphone was conducted. The DHMA shown in Fig.1 is used as the sound receiving device. This device does not cause spatial aliasing until 24 kHz because the distance between microphones is 7 mm. The sampling rate is 40 kHz. Window size is 1024 points (25.6 msec) and shift size is 256 points (6.4 msec). The FFT point is 1024. The reverberation time is 138 msec. in a low reverberation room. The number of sources is 12, and the experimental set-up is shown in Fig.4. The scaling problem of FDICA is solved using the projection back method[11].

To check the effectiveness of this method, L_s frames are selected from the total L_t frames given by STFT of observed signals. The L_s frames are selected following each selection criteria. For temporal frame selection, log power spectrum is normalized to a mean of 0 and a standard deviation of 1 as:

$$\bar{P} = \frac{P - \mu_P}{\sigma_P}, \tag{7}$$

where μ_P is the mean of the log power spectrum and σ_P is the standard deviation of the log power spectrum. Likewise, normalized kurtosis is calculated as:

$$\bar{K} = \frac{K - \mu_K}{\sigma_K}, \tag{8}$$

where μ_K is the mean of kurtosis and σ_K is the standard deviation of kurtosis. From the above, temporal frames with high $\bar{P}$ or low $\bar{K}$ are selected, respectively. When considering both selection criteria, another selection value is defined as

$$C = \beta\bar{P} - (1 - \beta)\bar{K}, \tag{9}$$

where β is a weight coefficient. Temporal frames that have high C are selected. In the conventional method, separation filters are estimated using the beginning L_s. SIR (Signal to Interference Ratio) improvement and processing time are compared. Since the observed signals are 4 sec. in length, the total number of frames is $L_t = 600$, with $L_s = 100, 150, \ldots,$ 600 representing the number of frames which are selected.

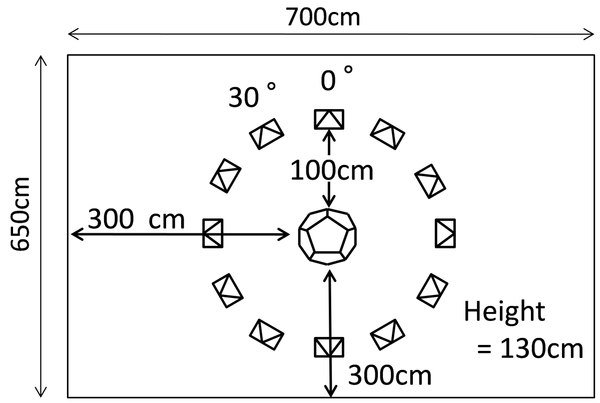


Fig. 4. Experimental set-up.

3.2. Evaluation method

SIR improvement is employed for evaluation of separated signals. SIR is the degree of interference of undesired signals with the desired signal. SIR improvement is represented as

$$\mathrm{SIR}_{\mathrm{improvement}_\xi} = \mathrm{OutputSIR}_\xi - \mathrm{InputSIR}_\xi. \tag{10}$$

Here InputSIR is the ratio of desired signal $x_\xi(t)$ to undesired signal $x_s(t)(s \neq \xi)$ in an observed signal, and is described as

$$\mathrm{InputSIR}_\xi = 10\log_{10}\left[\frac{\sum_t x_\xi(t)^2}{\sum_t\{\sum_{s\neq\xi} x_s(t)\}^2}\right]. \tag{11}$$

OutputSIR is the ratio of desired signal in the separation signal $y_{\xi\xi}(t)$ to undesired signal $y_{\xi s}(t)(s \neq \xi)$ and is represented as

$$\mathrm{OutputSIR}_\xi = 10\log_{10}\left[\frac{\sum_t y_{\xi\xi}(t)^2}{\sum_t\{\sum_{s\neq\xi} y_{\xi s}(t)\}^2}\right]. \tag{12}$$

In this paper, the average SIR improvement score using 12 sources is employed for evaluation.

3.3. Experimental results

The results using each method are shown in Fig.5. The results show that FDICA source separation can work without using temporal frames successively. By selecting high log power spectrum or low kurtosis temporal frames respectively, the number of frames needed to achieve the same values decreases in comparison to the conventional method. Higher values of SIR improvement can be achieved by employing kurtosis as a criteria, and by considering both log power spectrum and kurtosis, even higher SIR improvement values can be achieved. For example, the number of frames needed to achieve 24 dB can be reduced by 50%. In this paper, the weight β is 0.4 and emphasis is placed on the kurtosis.

Next, table 1 shows the relationship between the number of temporal frames to used for estimation and processing time. We considered the processing time using the conventional method, which requires estimating the settings for the filters using all 600 frames, to be 100%. When the number of temporal frames decreases, computational complexity also decreases. This result supports the theory that computational complexity is proportionally affected by the number of temporal frames. Thus it can be said that the number of temporal frames has a huge effect on processing time for estimating separation filters. In addition, the computational complexity for calculating our selection criteria, which are log power spectrum and kurtosis, are very small.

4. CONCLUSION

In this paper, the computational complexity of different FDICA source separation methods were discussed. We proposed a method in which separation filters are trained with

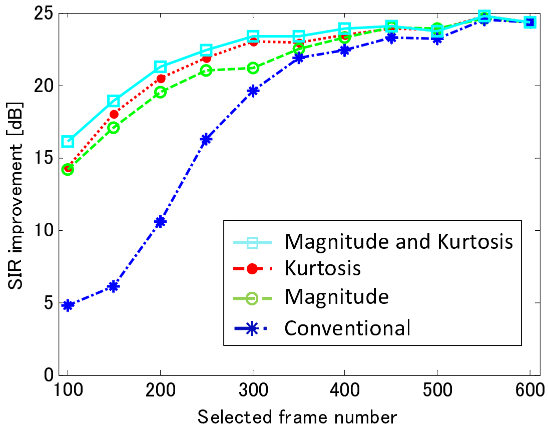


Fig. 5. Experimental results.

Table 1. Processing time. We considered the processing time using all 600 frames, to be 100%.

L_s	100	150	200	250	300
Time [%]	17.1	25.4	33.6	41.7	49.2
L_s	350	400	450	500	550
Time [%]	56.8	63.8	75.4	83.4	90.7

only selected temporal frames. Our proposed method uses the observed log power spectrum and kurtosis of amplitude distribution as selection criteria. Our separation experiment showed that higher SIR improvement values can be achieved with fewer temporal frames than when using the conventional method. We also found that it is especially effective to consider kurtosis and to select frames which include many sources. Even higher values can be achieved by considering both log power spectrum and kurtosis. Future work involves examining criteria for selection of temporal frames using other features of observed signals.

5. REFERENCES

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

[2] M. Ogasawara, T. Nishino, and K. Takeda, "Blind source separation based on acoustic pressure distribution and normalized relative phase using dodecahedral microphone array," *EUSIPCO 2009*, pp. 1413–1417, 2009.

[3] L.C. Parra and C.V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[4] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ica and beamforming," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[5] K. Kondo, Y. Takahashi, S. Hashimoto, H. Saruwatari, T. Nishino, and K. Takeda, "Efficient blind speech separation suitable for embedded devices," *EUSIPCO 2011*, pp. 2319–2328, 2011.

[6] M.Wax and T.Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Speech and Audio Processing*, 1985.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley-Interscience, 2001.

[8] S. Ikeda and N. Murata, "An approach to blind source separation of speech signals," *ICANN'98*, vol. 2, pp. 761–766, 1998.

[9] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[10] M.Ogasawara, T.Nishino, and K.Takeda, "Blind source separation using dodecahedral microphone array under reverberant conditions," *IEICE Trans. Fund. Electron. Comm. Comput. Sci.*, vol. 94, no. 3, pp. 897–906, 2011.

[11] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *International Symposium on Nonlinear Theory and Its Application*, vol. 3, pp. 923–926, 1998.