14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SOURCE SEPARATION USING MULTIPLE DIRECTIVITY PATTERNS PRODUCED BY ICA-BASED BSS

Takashi Isa, Toshiyuki Sekiya, Tetsuji Ogawa and Tetsunori Kobayashi

Department of Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

ABSTRACT

In this paper, we propose a multistage source separation method constructed by combining blind source separation (BSS) based on independent component analysis (ICA) and segregation using multiple directivity patterns (SMDP) introduced in our previous paper. We obtain the directivity patterns needed in SMDP by ICA-based BSS. In the SMDP, simultaneous equations of amplitudes of sound sources are generated by using these multiple directivities. The solution of these equations gives good disturbance estimates. We apply spectral subtraction using these disturbance estimates and the speech enhancement of the target source is performed. We conducted experimentation in a real room in the source-number-given condition where there is no priori information about the sound sources and the characteristics of room acoustics. The experimental results of double talk recognition show that the proposed technique is effective in reducing the error rate by 30% compared to frequency domain BSS.

1. INTRODUCTION

Multi-talk recognition is indispensable to realize various applications of hands free speech recognition, for example, conversation systems of a humanoid robots, conference dictation systems and car-navigation interfaces.

Recently, blind source separation (BSS) within the framework of independent component analysis (ICA) has been studied actively as one of the approaches for speech segregation or enhancement. However, there is still room for improvement of BSS in an actual environment [1]. One of the approaches towards using BSS in real acoustics is a multistage processing [2] [3].

We propose a method of source separation using multiple directivity pattern produced by BSS based on ICA. This method is the integration of BSS and the method of segregation using multiple directivity patterns (SMDP).

SMDP, which we have proposed, is characterized by using multiple directivity patterns to estimate the disturbance spectrum [4]. In previous papers, multiple directivity patterns are obtained by microphone array processes such as a beamformer. In this paper, we utilize the ICA-based BSS in order to produce the directivities. The directivity patterns are scaled and permuted versions of the rows of matrix as the solution of ICA.

There are some advantages to utilize ICA instead of conventional microphone array processing. One of the advantages is that ICA solutions provide efficient directivities for the SMDP with a small number of microphones compared to beamformer. Another advantage is no need of localizing the sound source precisely in order to obtain directivity [4]

However, the signal filtered by the matrix as the solution of ICA can still be improved significantly. We estimate the source spectrum from the information of multiple filtered signals and the directivity patterns given by BSS. The separating matrix of ICA solution has the same property as the directivity pattern in the sense of having the spatial information. When the source positions and directivity patterns are estimated, we can calculate the proportion of each source spectrum contained in the signals filtered by ICA-based BSS. Simultaneous equations relating the amplitude of sound source spectrum are set up considering the proportion as the coefficients. We can consider the solution as the estimates of the sound source spectra.

To show the effectiveness of the separation algorithm, we carry out the speech recognition experiment not in simulations but in a real room.

In the following section 2 and 3, ICA-based BSS and SMDP is reviewed respectively. In section 4, how BSS and SMDP are integrated is described. In section 5, the conditions and results of the continuous speech recognition are described. We give the discussions and conclusions in section 6 and 7.

2. ICA-BASED BSS IN THE FREQUENCY DOMAIN

In this section, we review ICA-based BSS briefly.

2.1 Formulation of the Sound Field

We assume the environment where S sound sources exist and the sound field is observed by M microphones. We define the input vector $\mathbf{x}(\omega, t)$ as STFT of the input signal.

$$\mathbf{x}(\omega, t) = [X_1(\omega, t), \cdots, X_M(\omega, t)]^T.$$

$X_i(\omega, t)$ denotes the STFT coefficient at microphone i, discrete frequency ω, and frame t. The operator $[\cdot]^T$ represents the transposition. Using the transfer function, $\mathbf{x}(\omega, t)$ is written as follows.

$$\begin{aligned} \mathbf{x}(\omega, t) &= \mathbf{A}(\omega)\mathbf{s}(\omega, t) + \mathbf{n}(\omega, t) \\ \mathbf{A}(\omega) &= [\boldsymbol{a}_1(\omega), \cdots, \boldsymbol{a}_S(\omega)] \\ \mathbf{s}(\omega, t) &= [s_1(\omega, t), \cdots, s_S(\omega, t)]^T \\ \mathbf{n}(\omega, t) &= [N_1(\omega, t), \cdots, N_M(\omega, t)]^T \end{aligned} \quad (1)$$

where, $\boldsymbol{a}_j(\omega)$ denotes the transfer characteristics or steering vector from j-th source to the microphones at

discrete frequency ω. $s_j(\omega,t)$ denotes the spectrum of j-th source. $N_i(\omega,t)$ denotes the spectrum of the background noise and the reverberation at microphone i.

2.2 ICA for the Source Separation

The purpose of BSS is to acquire the reconstructed signals $\mathbf{y}$ and the separating matrix $\mathbf{W}$ so that the rows of $\mathbf{y}$ are as mutually independent as possible in the following equation.

$$\mathbf{y}(\omega,t) = \mathbf{W}(\omega)\mathbf{x}(\omega,t). \tag{2}$$

From this, to simplify the expression, we omit the symbol ω and t. The output of ICA-based BSS is obtained by transforming the time-frequency-domain signal $\mathbf{y}$ into the time-domain signal.

However, we lack amplitude information of the source signals and their order. We must address these problems of scaling and permutation at each frequency for high performance. That is because a separated signal in the time domain is to be distorted and contains frequency components from other source signals. In this work, the scaling problem is solved by the mixing filter $\mathbf{W}^{-1}$, which is the inverse matrix of the ICA solution $\mathbf{W}$ [5]. The permutation problem is solved by the reference method with outputs of a basic binary mask technique as the reference signal [6].

3. SEGREGATION USING MULTIPLE DIRECTIVITY PATTERNS

In this section, we summarize the SMDP, Segregation using Multiple Directivity Patterns.

3.1 Estimation of the Sound Source Spectra

When a directivity pattern $\boldsymbol{f}_k\,(k=1,\cdots,Q)$ is given to the input vector $\mathbf{x}$ and DOAs can be estimated, the spectrum of filtered signal Y_k is written as follows.

$$\begin{aligned} Y_k &= \boldsymbol{f}_k^H \cdot \mathbf{x} \\ &= \boldsymbol{f}_k^H \cdot \boldsymbol{a}_1 s_1 + \cdots + \boldsymbol{f}_k^H \cdot \boldsymbol{a}_S s_S + \varepsilon_k \\ &= F_{k1}s_1 + \cdots + F_{kS}s_S + \varepsilon_k \end{aligned} \tag{3}$$

The operator $[\cdot]^H$ denotes the complex conjugate. F_{kj} represents the dot product between $\boldsymbol{f}_k^H$ and $\boldsymbol{a}_j$. We call F_{kj} directivity pattern. ε_k denotes the noise factor and error components which can not be modeled in equation (1) such as reverberation and errors of the steering vector $\boldsymbol{a}_j$. The power spectrum is derived from above equation (3) as the following.

$$\begin{aligned} |Y_k|^2 &= |F_{k1}|^2|s_1|^2 + \cdots + |F_{kS}|^2|s_S|^2 \\ &\quad +(F_{k1}s_1)\cdot(F_{k2}s_2)^H + (F_{k1}s_1)^H\cdot(F_{k2}s_2) \\ &\quad +(F_{k1}s_1)\cdot(F_{k3}s_3)^H + (F_{k1}s_1)^H\cdot(F_{k3}s_3) \\ &\quad +\cdots \end{aligned} \tag{4}$$

It is difficult to estimate the power spectrum from the data for a given short time because of the correlation among the sound sources. In other words, the effect of the dot-product terms $(F_{k1}s_1)\cdot(F_{k2}s_2)^H\cdots$ in equation (4) harms the estimates. We apply frame averaging with expectation to mitigate the effect of these terms. The average power spectrum is given as described below on the assumption that the sound sources are noncorrelated.

$$\langle|Y_k|^2\rangle = |F_{k1}|^2\langle|s_1|^2\rangle + \cdots + |F_{kS}|^2\langle|s_S|^2\rangle \tag{5}$$

The operator $\langle\cdot\rangle$ denotes the averaging for several frames. We can write a simultaneous equation using Q directivities.

$$\begin{aligned} \mathbf{y} &= \mathbf{F}\mathbf{s} + \boldsymbol{\varepsilon} \\ \mathbf{y} &= \left[\langle|Y_1|^2\rangle,\cdots,\langle|Y_Q|^2\rangle\right]^T \\ \mathbf{F} &= \begin{bmatrix} |F_{11}|^2 & \cdots & |F_{1S}|^2 \\ \vdots & \ddots & \vdots \\ |F_{Q1}|^2 & \cdots & |F_{QS}|^2 \end{bmatrix} \\ \mathbf{s} &= \left[\langle|s_1|^2\rangle,\cdots,\langle|s_Q|^2\rangle\right]^T \\ \boldsymbol{\varepsilon} &= [\varepsilon_1,\cdots,\varepsilon_Q]^T. \end{aligned} \tag{6}$$

The estimates of the sound source spectrum are found as the solutions of equation (6). When the number of directivities is larger than that of sound sources, the solutions minimize the squared error $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$.

$$\begin{aligned} \min_{\mathbf{s}}\ \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} \quad &\Rightarrow \quad \nabla_s\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = 0 \\ \bar{\mathbf{s}} &= (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y} \end{aligned} \tag{7}$$

3.2 Speech Enhancement

We derive the estimates of the target source spectrum and the disturbance source spectra in equation (7). However, we sacrifice the acoustic quality of the estimated target spectrum for frame averaging. In addition, the estimated spectrum cannot give satisfactory recognition accuracy. That is why the time resolution is degraded due to averaging step in equation (5) which is required to mitigate the effect of the correlation among sound sources. Accordingly, the disturbance spectrum is removed from short-time spectrum by spectral subtraction (SS). Let us assume that $\bar{s}_l$ is disturbance source spectrum estimated in equation (7). The symbol Y_j $(j\neq l)$ represents the target source short-time spectrum which is filtered by array processing or ICA-based BSS and is not averaged for several frames. The short-time spectrum of the target source $|\hat{s}_j|^2$ is obtained using the estimated disturbance spectrum $\langle|\bar{s}_l|^2\rangle$, $(l\neq j)$.

$$|\hat{s}_j|^2 = \begin{cases} |Y_j|^2 - \alpha\displaystyle\sum_{l\neq j}\langle|\bar{s}_l|^2\rangle, \\ \qquad\qquad \text{if } |Y_j|^2 - \alpha\displaystyle\sum_{l\neq j}\langle|\bar{s}_l|^2\rangle > \beta \\ \beta, \qquad\qquad \text{otherwise,} \end{cases} \tag{8}$$

The final output is the signal transformed time-frequency-domain signal $|\hat{s}_j|$ into the time-domain signal. To recover time-domain target signal, appropriate phase function has to be given. For example, we can use the phase of Y_j.

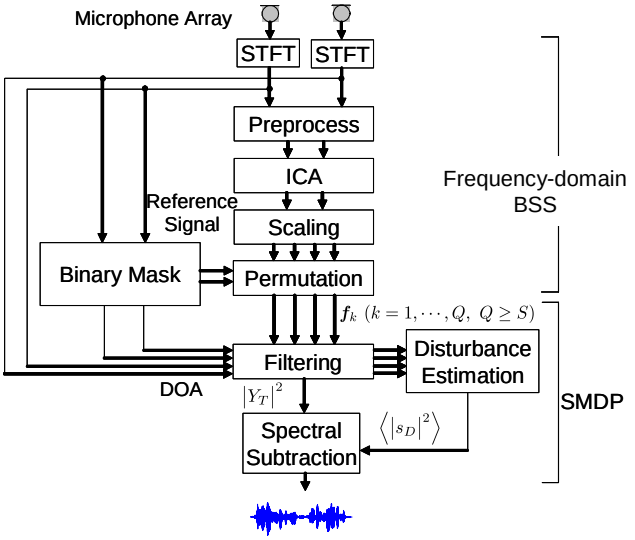


Figure 1: Diagram of proposed method

4. INTEGRATION OF SMDP AND ICA-BASED BSS

Figure 1 shows the diagram of proposed method. Our proposed method is composed of two main parts. One is BSS based on ICA in the frequency domain, the other is SMDP.

We have applied various beamformers to make directivities $\boldsymbol{f}_k$ in our previous works. In this paper, we utilize directivities obtained by the ICA-based BSS.

SMDP needs multiple directivities, signals filtered by the directivities, estimated DOAs and the steering vector. In this section, we explain how these components are acquired and how the ICA-based BSS is integrated into SMDP.

4.1 Directivities obtained by ICA

In the equation (2), the ICA solution $\mathbf{W}$ at a frequency is composed of the directivivetiy patterns when its permutation and scaling ambiguity are addressed.

$$\mathbf{W} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_S]^T \tag{9}$$

The separating matrix $\mathbf{W}$ is characterized by the spatial information [9]. That means the rows of $\mathbf{W}$ direct a null (which is the point of low gain) towards the sound sources (section 6 has the figures of its practical example). It is equivalent to the directivity patterns created by beamformer.

Basically, the number of the rows of $\mathbf{W}$ is that of sources S. It means we can use only S directivities. It is sufficient for SMDP to use S directivities but we can apply the SMDP using more directivities ($Q > S$). In the case of $Q > S$, we can set up redundant simultaneous equations between amplitudes of source spectra and multiple directivity patterns. Source spectra are estimated as the least squares solutions of these equations. To formulate redundant simultaneous equation, we must apply multiple techniques e.g. Delay and Sum beamformer, Minimum Variance Distortionless Response beamformer and so on [4]. We determine the coefficient of the equation with these directivities.

4.2 Filtered Signals

The term 'Filtered signals' represent the signals processed by the $\mathbf{W}$, in other words, the outputs of the ICA-based BSS in the frequency domain. It follows that the scaling and permutation of outputs must be solved. When the conventional beamformer is selected to organize the directivity patterns, designing directivity patterns and outputting the filtered signal are two different process in beamformer. On the contrary, the ICA outputs the separating matrix and the separated, filtered, signal. In the Figure 1, the process of 'Filtering' and 'ICA' are dipicted as different boxes due to the simplicity.

4.3 Estimate of DOA

SMDP needs estimates of DOA. That is because directivity pattern F_{kj} is a function as the DOA. We must estimate DOA to determine the coefficient of the simultaneous equation.

$$F_{kj}(\omega, \theta) = \boldsymbol{f}_k^H(\omega) \cdot \boldsymbol{a}_j(\omega, \theta) \tag{10}$$

It is possible to estimate the DOAs from the information of ICA solution [9]. However we used the binary mask output as the DOA estimates. There are two reasons. First, the SMDP does not need accurate DOA. The estimates of DOA in this stage is largely unaffected the performance of speech separation. The estimates of DOA from binary mask are sufficient for SMDP. The error of estimates in the stage to make the directivities by beamformer have a more negative effect on the performance [4]. It is the advantage of ICA over the beamformer that ICA does not require source localization to make directivities. Second, DOA is easily formed from binary-mask in the process of producing the reference signal to solve the permutation. Reference method is more effective in solving the permutation than the DOA-based method [6]

4.4 Steering Vector

In the SMDP, the transfer characteristics needs to be settled. The steering vectors which we use in this work are calculated at intervals of 10 degree in the range of -90 to 90 degree from the microphones on the assumption of far-field such as below.

$$\boldsymbol{a}_j(\omega) = \left[e^{-\jmath\omega d_1 \cos\theta_j}, \cdots, e^{-\jmath\omega d_M \cos\theta_j}\right]^T \tag{11}$$

The symbol $\jmath$ is the imaginary unit, d_i is the position of sensor, and θ_j is the DOA according to the j-th source. In the conventional array processing, steering vectors are observed such as the impulse response. It is also possible to assume the near-field when we obtain a information about the distance to the sound source. However, we use calculated one in order to avoid relying on environmental acoustics.

5. EXPERIMENT

To show the effectiveness, we applied the proposed method to double-talk recognition and evaluated the condition where the number of sources was given and the room acoustics was unknown.

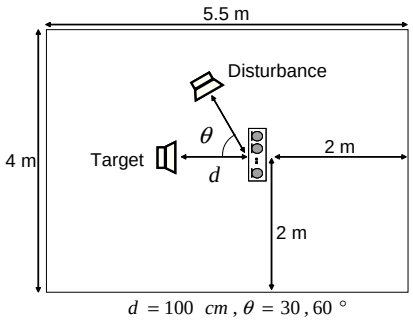


Figure 2: Recording condition. (We recorded evaluation data in a real room. The reverberation time was changed by drawing curtains.)

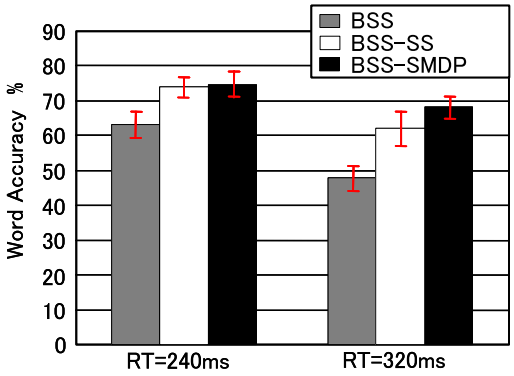


Figure 3: Evaluation of proposed method. (Each thick bar represents the average performance. Line on the bar represents the maximum and minimum performance.)

5.1 Experimental Setup

We recorded speech data to enable continuous speech recognition. The speech was sampled at 32 kHz. Two microphones with spacing of 3 cm were used. Figure 2 shows the recording condition. The reverberation time (RT) could be changed to 240 ms and 320 ms by drawing heavy curtains or not. The loudspeaker for the target source was arranged in front of the microphones. Another loudspeaker for the disturbance was placed at an angle with the target. The angle of disturbance was 30 degrees and 60 degrees from the target. Evaluation data was recorded for a total of four different conditions. As for target utterances, we selected a total of 100 sentences spoken by 23 male speakers from ASJ-JNAS continuous speech corpus [7]. In the same way, we selected speech data to play as disturbance utterances which were different from the target. Each utterance was adjusted to almost the same duration and energy. The SNR was almost 0 dB.

5.2 Speech Processing

5.2.1 ICA-based BSS

Analysis condition is described below. Frame length was 64 ms, frame shift was 8 ms with Hamming window.

JADE (joint approximate diagonalization of eigen matrices) extended to complex values was applied [8]. The scaling problem was solved by the mixing matrix and the permutation was solved by the reference method using the outputs of binary-mask as the reference signal.

5.2.2 SMDP

Analysis condition was the same as that of BSS. The SMDP parameter, the number of the frame to average was three. The SS parameter, α and β, was determined through preliminary experiments so that the word accuracy was the highest in each environment (reverberation time and the angle of the disturbance).

5.2.3 Speech Recognition

The parameters of the acoustic features were 12-dimensional MFCC, ΔMFCC and Δpower. Pre-emphasis was done with $1 - 0.97z^{-1}$. Frame length was 25 ms and frame shift was 10 ms by applying Hamming window.

The acoustic models were trained with 20 K sentences spoken by about 100 male speakers from ASJ-JNAS corpus. The training data was recorded with a close-talk microphone. The language models were the trigram language models using lexicon of 20 K vocabulary size. In this experiment, the speech data was sampled at 32 kHz while the acoustic models were trained with the speech data sampled at 16 kHz. Thus segregated speech was converted to 16 kHz sampling rate and converted to acoustic features.

5.3 Evaluation

We compared the performance of three methods to confirm the effectiveness of the proposed method. One is the BSS-only method, second is the BSS-SS method. BSS-only method is normal BSS based on ICA. BSS-SS method is the integration of BSS and spectral subtraction (SS), which is also our original method. In this method, estimated disturbance, one output of ICA, is directly subtracted from estimated target, the other output of ICA. That means we used $|Y_l|^2$ instead of $\langle|\bar{s}_l|^2\rangle$ in equation (8).

5.4 Results

Figure 3 shows the continuous speech recognition results. The performance of the BSS only method was not sufficient for the given reverberation time. BSS-SS method was effective in enhancing the target signal compared to BSS. However, our proposed method performed the best out of the three when considering the long reverberation. The comparison of the SMDP with commonly used beamformer have been described in our works before. For comparison, the directionally constrained minimization of power (DCMP) adaptive array performs about 50 % in the recognition accuracy in the condition of short reverberation time using eight microphones.

6. DISCUSSIONS

Figure 4 shows the actual directivity patterns produced by ICA-based BSS at frequency 3200 Hz under the condition of two sound sources. This figure was plotted $\boldsymbol{f}_k^H \cdot \boldsymbol{a}_j$ for any θ in equation (11). In the case of two sound sources, the filtered signal Y_1 corresponding to the target is described as follows.

$$\langle |Y_1|^2 \rangle = |F_{11}|^2 \langle |s_1|^2 \rangle + |F_{12}|^2 \langle |s_2|^2 \rangle + \varepsilon_1 \quad (12)$$

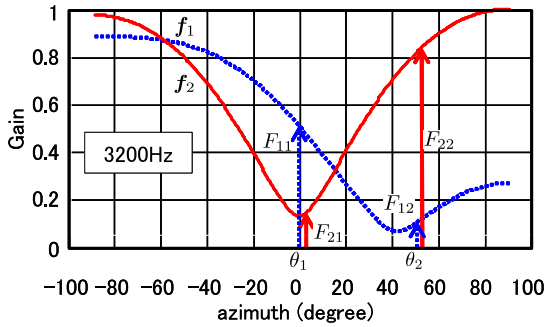
The symbol F_{11} expresses the proportion of the target source power to the disturbance by directivity $\boldsymbol{f}_1$. Similarly, F_{12} means the proportion of the disturbance source to the target. Equation (12) shows that the output of the ICA-based BSS still contains other source components. That is because the weight of disturbance, F_{12}, is never zero and there are many reflections in a real room . The process of estimating sound source spectra in SMDP works in reducing the components from other sources. To solve equation (6) is equivalent of the minimizing the negative influence of other sources.

In this work, we utilized steering vectors calculated on the assumption of far-field not relying on room acoustics. Above means that we expect SMDP to remove the direct sound component of other source. ICA deal with the reflections in a statistical manner in the frequency-domain. However, we assume the power of reflection is itself small. Additionally, reflections come through different directions than DOAs. The power of reflection become smaller with multiplying the directivity (or beamformer) by inputs. It is reasonable to approximate the real transfer characteristic with calculated one.

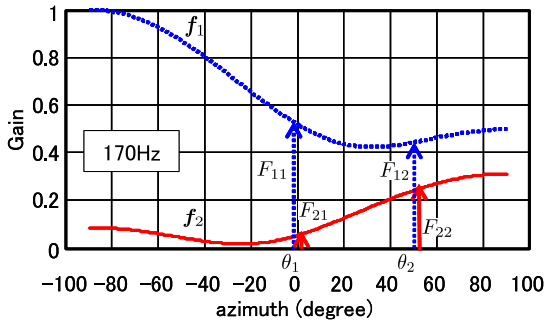
To cope with the deterioration of time resolution, SS can be carried out. In other words, SMDP eliminate the disturbance components which has less other components from the target source. That is the difference between the BSS outputs and SMDP outputs. We consider that above as the reason why BSS-SMDP performed more effective than BSS-SS.

7. CONCLUSION

We proposed a multistage processing with integration of BSS and SMDP for a speech segregation system. Proposed method achieved 70 % word accuracy in double-talk recognition of 20 K vocabulary in source-number-given condition. From the comparison of the ICA-based BSS method, the great advantage of proposed method was shown, particularly in long reverberation environment.



(a)



(b)

Figure 4: Directivity pattern produced by ICA solution (a) at frequency 3200 Hz, (b) at frequency 170 Hz. ($\boldsymbol{f}_1$ is the directivity which enhances the target source which is placed at zero degree and suppresses the disturbance which is placed at 60 degree. $\boldsymbol{f}_2$ has the reverse characteristics. θ_1 and θ_2 are the estimates of DOA. F_{kj} represents the coefficients of the equation (6).)

REFERENCES

[1] M. Z. Ikram et al., "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," Proc. ICASSP2000, vol. 2, pp. 1041–1044.

[2] R. Mukai et al., "Removal of residual cross-talk components in blind source separation using time-delayed spectral subtraction," Proc. ICASSP2002, vol. 2, pp. 1789–1792.

[3] S. Ukai et al., "Blind source separation combining SIMO-model-based ICA and adaptive beamforming," Proc. ICASSP2005, vol. 3, pp. 85–88.

[4] T. Sekiya et al., "Speech recognition in the blind condition based on multiple directivity patterns using a microphone array," Proc. ICASSP2005. vol. 1, pp. 373–376.

[5] N. Murata et al., "An approach to blind source separation based on temporal structure of speech signals," Neurocomputing, vol. 41, pp. 1–24, 2001.

[6] T. Isa et al., "A Method for Solving the Permutation Problem of Frequency-Domain BSS Using Reference Signal" Proc. EUSIPCO2006, submitted.

[7] K. Itou et al., "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," Proc. ICSLP98, pp.3261-3264.

[8] J. F. Cardoso et al., "Blind beamforming for non Gaussian signals," IEE Proc., vol. F140, pp. 362–370, 1993.

[9] H. Sawada et al., "A Robust and precise method for solving the permutation problem of frequency-domain blind source separation," IEEE Trans. ASSP, vol. ASSP-12, pp. 530–538, 2004.