PAPER Special Section on Digital Signal Processing

Blind Source Separation for Moving Speech Signals Using Blockwise ICA and Residual Crosstalk Subtraction

Ryo MUKAI[†a)], **Hiroshi SAWADA**[†], *Members*, **Shoko ARAKI**[†], *Nonmember*, *and* **Shoji MAKINO**[†], *Member*

SUMMARY This paper describes a real-time blind source separation (BSS) method for moving speech signals in a room. Our method employs frequency domain independent component analysis (ICA) using a blockwise batch algorithm in the first stage, and the separated signals are refined by postprocessing using crosstalk component estimation and non-stationary spectral subtraction in the second stage. The blockwise batch algorithm achieves better performance than an online algorithm when sources are fixed, and the postprocessing compensates for performance degradation caused by source movement. Experimental results using speech signals recorded in a real room show that the proposed method realizes robust real-time separation for moving sources. Our method is implemented on a standard PC and works in realtime.

key words: *blind source separation, independent component analysis, convolutive mixtures, realtime, spectral subtraction, post processing*

1. Introduction

Blind source separation (BSS) is a technique for estimating original source signals using only observed mixtures. The BSS of audio signals has a wide range of applications including noise robust speech recognition, hands-free telecommunication systems and high-quality hearing aids. In most realistic applications, the source location may change, and the mixing system is time-varying. Although a large number of studies have been undertaken on BSS based on independent component analysis (ICA) [1]–[5], only few studies have been made on BSS for moving source signals [6]–[9]. Indeed an online algorithm can track a time-varying system, however, in general, its performance is worse than a batch algorithm when the system becomes stationary. Although we are dealing with moving sources, we do not want to degrade the performance for fixed sources.

In this paper, we propose a robust real-time BSS method that employs frequency domain ICA using a blockwise batch algorithm in the first stage, and the postprocessing of crosstalk component estimation and non-stationary spectral subtraction in the second stage. When we adopt a blockwise frequency domain ICA, we need to solve a permutation problem for every block, and this is a time consuming process especially when the block length is short. We use an algorithm based on analytical calculation of null directions to solve the permutation problem quickly [10]. Another problem inherent to batch algorithms is an input-output delay. To reduce the delay, we use a technique for computing output signal without waiting for the calculation of the separating system to be completed. These techniques are useful for realizing low-delay real-time BSS.

The blockwise batch algorithm achieves better separation performance than an online algorithm for fixed source signals, but the performance declines for moving sources. As we pointed out in [11], the solution of ICA works like an adaptive beamformer, which forms a spatial null towards a jammer signal. This characteristic means that BSS using ICA is fragile as regards a moving jammer signal but robust with respect to a moving target signal. Utilizing this nature, we can estimate residual crosstalk components even when a jammer signal moves. To compensate for the degradation when a jammer signal moves, we employ postprocessing in the second stage. Figure 1 shows a block diagram of the proposed method for one output channel in one frequency bin. In contrast to the original spectral subtraction [12], which assumes stationary noise and periods with no target signal when estimating the noise spectrum, our method requires neither assumption because we use BSS in the first stage. A large amount of research has been undertaken on spectral subtraction for non-stationary noise conditions, and some researchers have proposed a combination comprising a microphone array and spectral subtraction [13], [14]. In our method, the jammer signal is mostly eliminated by the first stage, and the spectral subtraction is used for removing residual components which have small power. In addition, we can estimate the non-stationary spectrum accurately by utilizing signals separated in the first stage. Therefore the distortion of separated signals caused by over subtraction or under subtraction is small.

This paper is organized as follows. In the next section, we summarize the algorithm of frequency domain BSS for convolutive mixtures and formulate a blockwise batch algo-

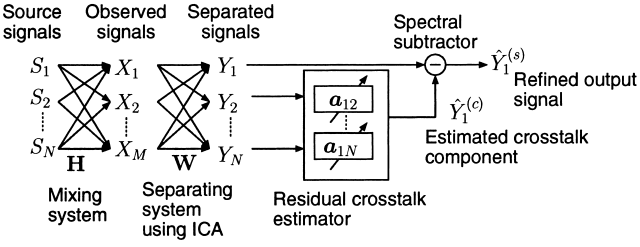

Fig. 1 Block diagram of proposed system for one output channel in one frequency bin.

Manuscript received November 27, 2003.
Manuscript revised February 21, 2004.
Final manuscript received April 8, 2004.

[†]The authors are with NTT Communication Science Laboratories, NTT Corporation, Kyoto-fu, 619-0237 Japan.
a) E-mail: ryo@cslab.kecl.ntt.co.jp

rithm. In Sect. 3, we propose an algorithm to estimate and subtract residual crosstalk components in the separated signals. Section 4 presents experimental results using speech signals recorded in a room and show the effectiveness of the method in realizing robust real-time separation. Section 5 concludes this paper.

2. ICA Based BSS of Convolutive Mixtures

In this section, we briefly review the BSS algorithm that uses frequency domain ICA and formulate a blockwise batch algorithm including an online algorithm as a special case. We also describe a fast algorithm for solving permutation problems, which is necessary for real-time processing.

2.1 Frequency Domain ICA

When the source signals are $s_i(t)(i = 1, ..., N)$, the signals observed by microphone j are $x_j(t)(j = 1, ..., M)$, and the separated signals are $y_k(t)(k = 1, ..., N)$, the BSS model can be described by the following equations:

$$x_j(t) = \sum_{i=1}^{N} (h_{ji} * s_i)(t) \tag{1}$$

$$y_k(t) = \sum_{j=1}^{M} (w_{kj} * x_j)(t) \tag{2}$$

where h_{ji} is the impulse response from source i to microphone j, w_{kj} are the separating filters, and $*$ denotes the convolution operator.

A convolutive mixture in the time domain is converted into multiple instantaneous mixtures in the frequency domain. Therefore, we can apply an ordinary ICA algorithm in the frequency domain to solve a BSS problem in a reverberant environment. Using a short-time discrete Fourier transform (STDFT) for (1), the model is approximated as:

$$\mathbf{X}(\omega, n) = \mathbf{H}(\omega)\mathbf{S}(\omega, n), \tag{3}$$

where, ω is the angular frequency, and n represents the frame index. The separating process can be formulated in each frequency bin as:

$$\mathbf{Y}(\omega, n) = \mathbf{W}(\omega)\mathbf{X}(\omega, n), \tag{4}$$

where $\mathbf{S}(\omega, n) = [S_1(\omega, n), ..., S_N(\omega, n)]^T$ is the source signal in frequency bin ω, $\mathbf{X}(\omega, n) = [X_1(\omega, n), ..., X_M(\omega, n)]^T$ denotes the observed signals, $\mathbf{Y}(\omega, n) = [Y_1(\omega, n), ..., Y_N(\omega, n)]^T$ is the estimated source signal, and $\mathbf{W}(\omega)$ represents the separating matrix. $\mathbf{W}(\omega)$ is determined so that $Y_i(\omega, n)$ and $Y_j(\omega, n)$ become mutually independent.

To calculate the separating matrix $\mathbf{W}$, we use an optimization algorithm based on the minimization of the mutual information of $\mathbf{Y}$. The optimal $\mathbf{W}$ is obtained by the following iterative equation using the natural gradient approach [15]:

$$\mathbf{W}^{(i+1)} = \mathbf{W}^{(i)} + \mu[\mathbf{I} - \langle \Phi(\mathbf{Y})\mathbf{Y}^H \rangle]\mathbf{W}^{(i)}, \tag{5}$$

where i is an index for the iteration, $\mathbf{I}$ is an identity matrix, μ is a step size parameter, $\langle \cdot \rangle$ denotes the averaging operator, and $\Phi(\cdot)$ is a nonlinear function. Because the signals have complex values in the frequency domain, we use a polar coordinate based nonlinear function, which is effective for fast convergence especially when the number of input data samples is small [16]:

$$\Phi(\mathbf{Y}) = \tanh(g \cdot \mathrm{abs}(\mathbf{Y}))e^{J \arg(\mathbf{Y})}, \tag{6}$$

where g is a gain parameter that controls the nonlinearity.

2.2 Scaling and Permutation

Once we have completed the ICA for all frequencies, we need to solve the permutation and scaling problems. Since we are handling signals with complex values, the scaling factors are also complex values. Thus the scaling can be divided into phase scaling and amplitude scaling. We use a direction of arrival (DOA) based method to solve the permutation and phase scaling problems. The permutation problem is solved so that the DOAs of the separated signals are aligned, and the phase scaling problem is solved so that the phase response of the estimated source direction becomes zero.

The DOA of the i-th separated signal $\theta_i(\omega)$ can be calculated analytically as [10]:

$$\theta_i(\omega) = \arccos \frac{\arg([\mathbf{W}(\omega)^{-1}]_{ji}/[\mathbf{W}(\omega)^{-1}]_{j'i})}{\omega c^{-1}|d_j - d_{j'}|}, \tag{7}$$

where $[\cdot]_{ji}$ denotes ji-th element of the matrix, c is the speed of sound, and d_j represents a location of microphone j. This method does not require the directivity pattern to be scanned, thus we can solve the permutation problem quickly.

The amplitude scaling problem is solved by using a slightly modified version of the method described in [17]. We calculate the inverse of the separating matrices $\mathbf{W}(\omega)^{-1}$, and decide the scaling factors so that the norms of each column of $\mathbf{W}(\omega)^{-1}$ become uniform.

2.3 Low Delay Blockwise Batch Algorithm

In order to track the time-varying mixing system, we update the separating matrix for each time block $B_m = \{t : (m-1)T_b \leq t < mT_b\}$, where T_b is the block size, and m represents the block index ($m \geq 1$).

Koutras et al. have proposed a similar method in the time domain [7]. When T_b equals the STDFT frame length, this procedure can be considered an online algorithm in the frequency domain.

We use the separating matrix of the previous block as the initial iteration value for a new block, i.e., $\mathbf{W}_{m+1}^{(0)}(\omega) = \mathbf{W}_m^{(N_I)}(\omega)$, where N_I is the number of iterations for (5). We use a set of two null beamformers as the initial matrix $\mathbf{W}_1^{(0)}(\omega)$ for the first block.

The batch algorithm has an inherent delay, because the

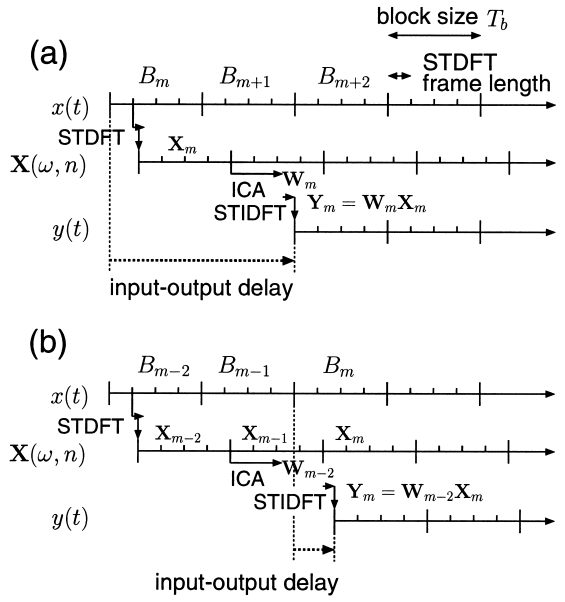

Fig. 2 Input-output delay of (a) BSS using ordinary blockwise batch algorithm, and (b) BSS without waiting for calculation of $\mathbf{W}_m$.

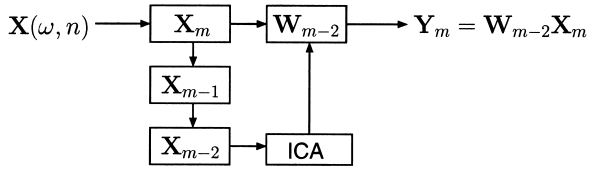

Fig. 3 Data flow of low delay blockwise batch algorithm.

calculation of $\mathbf{W}$ needs to wait for the arrival of a data block. Moreover, the calculation itself also takes time (Fig. 2(a)). However, when the calculation is completed within T_b and we use $\mathbf{W}_{m-2}$ for separation of the signals in B_m, we can avoid the delay for waiting and calculation (Fig. 2(b)). This technique can reduce the input-output delay and is suitable for low-delay real-time applications. Figure 3 shows a data flow diagram of the proposed method. The block of the observed signal $\mathbf{X}_m$ is queued for the ICA process. Concurrently, $\mathbf{X}_m$ is separated by $\mathbf{W}_{m-2}$, which is ready before the arrival of $\mathbf{X}_m$. Accordingly, the block of the separated signal $\mathbf{Y}_m$ can be calculated with low delay.

It seems that this method fails when a source signal moves, but it is actually robust for the moving target signal, which is shown in Sect. 4.3. Unfortunately, this method suffer performance deterioration when a jammer signal moves. To cope with this problem, we propose a postprocessing method using crosstalk component estimation and non-stationary spectral subtraction which reduces the performance deterioration.

3. Residual Crosstalk Subtraction

In this section, we examine the nature of separated signals obtained by the frequency domain ICA described in the previous section. We then propose an algorithm to estimate and subtract residual crosstalk components in these signals. We have examined this algorithm with fixed speech signals and confirmed the effectiveness of the proposed method in [18].

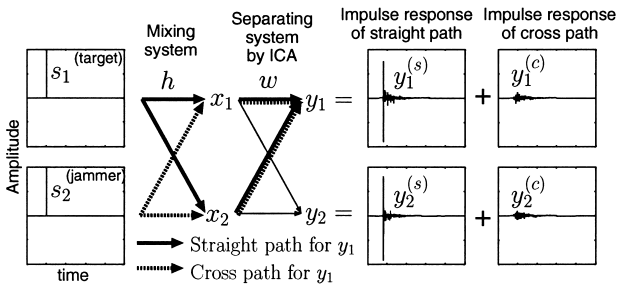

Fig. 4 Impulse responses of straight path and cross path.

3.1 Straight and Crosstalk Components of BSS

When we denote the concatenation of a mixing system and a separating system as $\mathbf{G}(\omega)$, i.e., $\mathbf{G}(\omega) = \mathbf{W}(\omega)\mathbf{H}(\omega)$, each of the separated signals Y_i obtained by BSS can be described as follows:

$$Y_i(\omega, n) = \sum_{j=1}^{N} G_{ij}(\omega) S_j(\omega, n). \tag{8}$$

We decompose Y_i into the sum of straight component $Y_i^{(s)}$ derived from target signal S_i and crosstalk component $Y_i^{(c)}$ derived from jammer signals $S_j (j \neq i)$. Then, we have

$$Y_i(\omega, n) = Y_i^{(s)}(\omega, n) + Y_i^{(c)}(\omega, n) \tag{9}$$

$$Y_i^{(s)}(\omega, n) = G_{ii}(\omega) S_i(\omega, n) \tag{10}$$

$$Y_i^{(c)}(\omega, n) = \sum_{j \neq i} G_{ij}(\omega) S_j(\omega, n). \tag{11}$$

We denote estimation of $Y_i^{(s)}$ and $Y_i^{(c)}$ as $\hat{Y}_i^{(s)}$ and $\hat{Y}_i^{(c)}$, respectively. Our goal is to estimate the spectrum of $Y_i^{(c)}$ using only $Y_j (1 \leq j \leq N)$ and obtain $\hat{Y}_i^{(s)}$ by subtracting $\hat{Y}_i^{(c)}$ from Y_i.

In our previous research [19], we measured the impulse responses of the straight and cross paths of a BSS system. As a result, we found that the direct sound of a jammer can be almost completely removed by BSS, and also that residual crosstalk components are derived from the reverberation (Fig. 4). We utilize these characteristics of separated signals to estimate the crosstalk components.

3.2 Model of Residual Crosstalk Component Estimation

Figure 5 shows an example of a narrow band power spectrum of straight and crosstalk components in separated signals obtained by a two-input two-output BSS system. The crosstalk component $Y_1^{(c)}$ is in Y_1 and the straight component $Y_2^{(s)}$ is in Y_2. Both components are derived from source signal S_2; $Y_1^{(c)}$ is derived from the reverberation of S_2 and

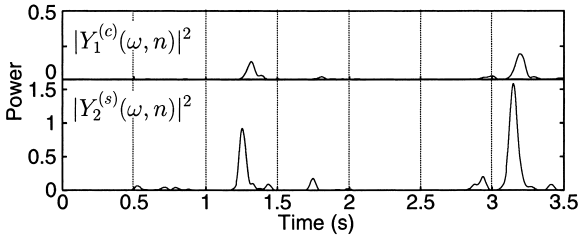

Fig. 5 Example of narrow band power spectrum of straight and crosstalk components ($\omega = 320$ Hz).

$Y_2^{(s)}$ is mainly derived from the direct sound of S_2. Accordingly, for the narrow band signal in each frequency bin, the crosstalk component $Y_1^{(c)}$ can be approximated by the output of the filter whose input is the straight component of the other channel $Y_2^{(s)}$.

We extend this approximation to multiple signals by introducing filters $\boldsymbol{a}_{ij}(\omega,n) = [a_{ij0}(\omega,n), ..., a_{ijL-1}(\omega,n)]^T$ for each frequency bin ω and combination of channels i and j ($i \neq j$), where L is the length of filters.

Furthermore, we use Y_j as an approximation of $Y_j^{(s)}$, because $Y_j^{(s)}$ is actually unknown. Therefore, the model for estimating residual crosstalk components is formulated as follows:

$$|Y_i^{(c)}(\omega,n)|^\beta \approx \sum_{j\neq i}\sum_{k=0}^{L-1} a_{ijk}(\omega,n)|Y_j^{(s)}(\omega,n-k)|^\beta \quad (12)$$

$$\approx \sum_{j\neq i}\sum_{k=0}^{L-1} a_{ijk}(\omega,n)|Y_j(\omega,n-k)|^\beta \quad (13)$$

where the exponent $\beta = 1$ for the magnitude spectrum and $\beta = 2$ for the power spectrum.

3.3 Adaptive Algorithm and Spectrum Estimation

Figure 6 shows a block diagram of the proposed method for one output channel. We estimate filters $\boldsymbol{a}_{ij}$ described in the previous section by using an adaptive algorithm based on the normalized LMS (NLMS) algorithm [20].

For each i, the filters $\hat{\boldsymbol{a}}_{ij}(\omega,n)$ are adapted so that the sum of the output signals becomes $|Y_i^{(c)}(\omega,n)|^\beta$ for input signals $|Y_j^{(s)}(\omega,n)|^\beta$ ($1 \leq j \leq N$, $j \neq i$). Unfortunately, $|Y_i^{(c)}(\omega,n)|$ and $|Y_j^{(s)}(\omega,n)|$ are unknown, so they are substituted by $|Y_i(\omega,n)|$ and $|Y_j(\omega,n)|$, respectively. We assume that $|Y_i^{(s)}(\omega,n)|$ can be approximated by $|Y_i(\omega,n)|$ when $|Y_i(\omega,n)|$ is large and $|Y_i^{(c)}(\omega,n)|$ can be approximated by $|Y_i(\omega,n)|$ when $|Y_i(\omega,n)|$ is small. This assumption is based on the characteristics of narrow band signals where $Y_i^{(s)}$ and $Y_j^{(s)}$ seldom have large power simultaneously, especially when the source signals are speech signals. A detailed analysis of overlapping frequency components of speech signals can be found in [21] and [22].

Since not all $|Y_i^{(c)}(\omega,n)|$ and $|Y_i^{(s)}(\omega,n)|$ can be approximated by $|Y_i(\omega,n)|$, only a subset of the filters is updated at each iteration. To formulate a selective update algorithm, we introduce sets of channel index numbers, $\mathcal{I}_S(\omega,n) = \{i : |Y_i(\omega,n)| \approx |Y_i^{(s)}(\omega,n)|\}$ and $\mathcal{I}_C(\omega,n) = \{i : |Y_i(\omega,n)| \approx |Y_i^{(c)}(\omega,n)|\}$. This means that $|Y_i^{(s)}(\omega,n)|$ can be approximated by $|Y_i(\omega,n)|$ for $i \in \mathcal{I}_S(\omega,n)$ and $|Y_i^{(c)}(\omega,n)|$ can be approximated by $|Y_i(\omega,n)|$ for $i \in \mathcal{I}_C(\omega,n)$.

One example implementation for determining $\mathcal{I}_S(\omega,n)$ and $\mathcal{I}_C(\omega,n)$ is $\mathcal{I}_S(\omega,n) = \{i : i = \mathrm{argmax}_i |Y_i(\omega,n)|\}$ and $\mathcal{I}_C(\omega,n) = \overline{\mathcal{I}_S(\omega,n)}$. Another example is $\mathcal{I}_S(\omega,n) = \{i : |Y_i(\omega,n)| > threshold\}$ and $\mathcal{I}_C(\omega,n) = \overline{\mathcal{I}_S(\omega,n)}$.

The filters $\hat{\boldsymbol{a}}_{ij}$ are updated for $i \in \mathcal{I}_C(\omega,n)$ and $j \in \mathcal{I}_S(\omega,n)$. The update procedure is given by

$$\hat{\boldsymbol{a}}_{ij}(\omega,n+1) = \begin{cases} \hat{\boldsymbol{a}}_{ij}(\omega,n) + \dfrac{\eta}{\delta + \|\boldsymbol{u}_j(\omega,n)\|^2}\boldsymbol{u}_j(\omega,n)e_{ij}(\omega,n) \\ \qquad (\text{if } i \in \mathcal{I}_C(\omega,n), \text{ and } j \in \mathcal{I}_S(\omega,n)) \\ \hat{\boldsymbol{a}}_{ij}(\omega,n) \quad (\text{otherwise}) \end{cases}, \quad (14)$$

where $\boldsymbol{u}_j(\omega,n) = [|Y_j(\omega,n)|^\beta, |Y_j(\omega,n-1)|^\beta, ..., |Y_j(\omega,n-L+1)|^\beta]^T$ is a tap input vector and $e_{ij}(\omega,n) = |Y_i(\omega,n)|^\beta - \sum_{j\neq i}\hat{\boldsymbol{a}}_{ij}^T(\omega,n)\boldsymbol{u}_j(\omega,n)$ is an estimation error. Here, η is a step size parameter and δ is a positive constant to avoid numerical unstability when $\|\boldsymbol{u}_j\|$ is very small.

We apply the estimated filters to the model (13), and obtain an estimation of the power of residual crosstalk components:

$$|\hat{Y}_i^{(c)}(\omega,n)|^\beta \approx \sum_{j\neq i}\hat{\boldsymbol{a}}_{ij}^T(\omega,n)\boldsymbol{u}_j(\omega,n). \quad (15)$$

Finally, we obtain an estimation of the straight component as $\hat{Y}_i^{(s)}$ by the following spectral subtraction procedure:

$$\hat{Y}_i^{(s)}(\omega,n) = \begin{cases} (|Y_i(\omega,n)|^\beta - |\hat{Y}_i^{(c)}(\omega,n)|^\beta)^{1/\beta}\dfrac{Y_i(\omega,n)}{|Y_i(\omega,n)|} \\ \qquad (\text{if } |Y_i(\omega,n)| > |\hat{Y}_i^{(c)}(\omega,n)|) \\ 0 \quad (\text{otherwise}) \end{cases}. \quad (16)$$

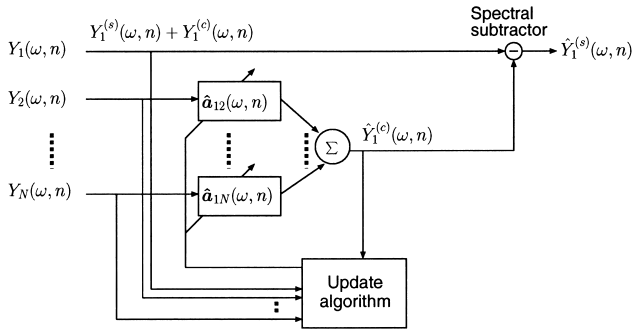

Fig. 6 Adaptive filters and spectral subtractor to estimate $Y_i^{(s)}$ (for $i = 1$).

4. Experiments and Discussions

4.1 Experimental Conditions

To examine the effectiveness of the proposed method, we carried out experiments using speech signals recorded in a room. The reverberation time of the room was 130 ms. We used two omni-directional microphones with an inter-element spacing of 4 cm. The layout of the room is shown in Fig. 7. The target source signal was first located at A, and then moved to B at a speed of 30 deg/s. The jammer signal was located at C and moved to D at a speed of 40 deg/s.

The step size parameter μ in (5) affects the separation performance of BSS when the block size changes. We carried out preliminary experiments and chose μ to optimize the performance for each block size. Other conditions are summarized in Table 1. The frame shift and the filter length L in the postprocessing part were decided so that the filter could cover the reverberation.

To update filters $\hat{\boldsymbol{a}}_{ij}(\omega, n)$, we used the following simple selective update policy:

if $|Y_1(\omega, n)| > |Y_2(\omega, n)|$
then $\mathcal{I}_S(\omega, n) = \{1\},\ \mathcal{I}_C(\omega, n) = \{2\}$
else $\mathcal{I}_S(\omega, n) = \{2\},\ \mathcal{I}_C(\omega, n) = \{1\}$.

We assumed the straight component $y_1^{(s)}$ as a signal, and the difference between the output signal and the straight component as interference. We defined the output signal-to-interference ratio (SIR_O) in the time domain as follows:

$$\mathrm{SIR}_O \equiv 10 \log \frac{\sum_t |y_1^{(s)}(t)|^2}{\sum_t |y_1(t) - y_1^{(s)}(t)|^2} \ (\mathrm{dB}). \tag{17}$$

Similarly, the input SIR (SIR_I) is defined as,

$$\mathrm{SIR}_I \equiv 10 \log \frac{\sum_t \sum_{i=1}^{2} |(h_{i1} * s_1)(t)|^2}{\sum_t \sum_{i=1}^{2} |(h_{i2} * s_2)(t)|^2} \ (\mathrm{dB}). \tag{18}$$

We use $\mathrm{SIR} = \mathrm{SIR}_O - \mathrm{SIR}_I$ as a performance measure. This measurement is consistent with the performance evaluation of BSS in which the crosstalk component is assumed as interference. We measured SIRs with 30 combinations of source signals using three male and three female speakers, and averaged them.

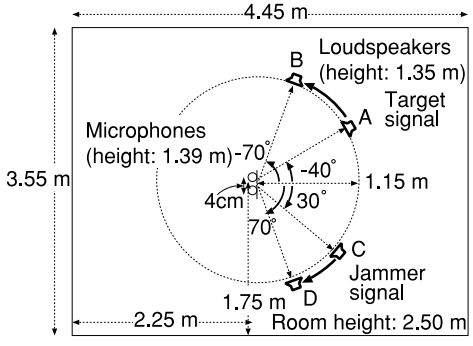

Fig. 7 Layout of room used in experiments.

Table 1 Experimental conditions.

Common	Sampling rate = 8 kHz Window = hanning Reverberation time T_R=130 ms
ICA part	Frame length T_{ICA} = 1024 point (128 ms) Frame shift = 256 point (32 ms) g = 100.0 μ = optimized for block size T_b Number of iterations N_I = 100
Post processing part	Frame length T_{SS} = 1024 point (128 ms) Frame shift = 64 point (8 ms) Filter length L = 16 β = 2 δ = 0.01 η = 0.1

4.2 Performance for Fixed Sources

Although we are dealing with moving sources, we do not want the performance for fixed sources to deteriorate. First, we measured the BSS performance using ICA without postprocessing. Figure 8 shows the average and standard deviation of SIR for fixed sources (the target is at A and the jammer at C in Fig. 7). This indicates that the blockwise batch algorithm outperforms the online algorithm (in which μ is tuned to optimize the performance), when we use the update Eq. (5). In addition, the deviation of the batch algorithm is smaller than that of the online algorithm. This is why we adopt the blockwise batch algorithm in the first stage. We used T_b = 1.0 sec. in the following experiments.

4.3 Moving Target and Moving Jammer

Before considering the result obtained with the postprocessing method, we investigate the BSS performance for moving sources using the blockwise batch algorithm. Figure 9 shows the SIR for a moving target (solid line) and for a moving jammer (dotted line). We can see that the SIR is not degraded even when the target moves. By contrast, jammer movement causes a decline in the SIR.

This can be explained by the directivity pattern of the

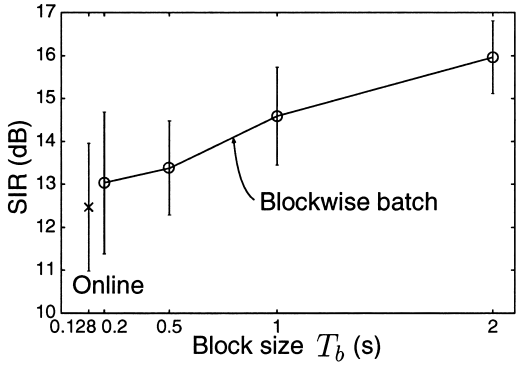

Fig. 8 Average and standard deviation of SIR for fixed sources.

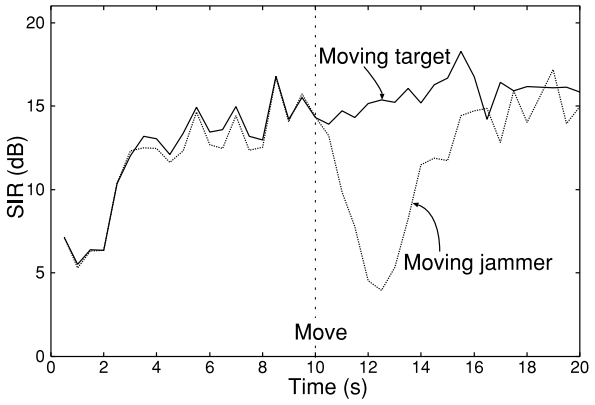

Fig. 9 SIR of blockwise batch algorithm without postprocessing. Target and jammer signals moved at 10 sec. (T_b = 1.0 sec.)

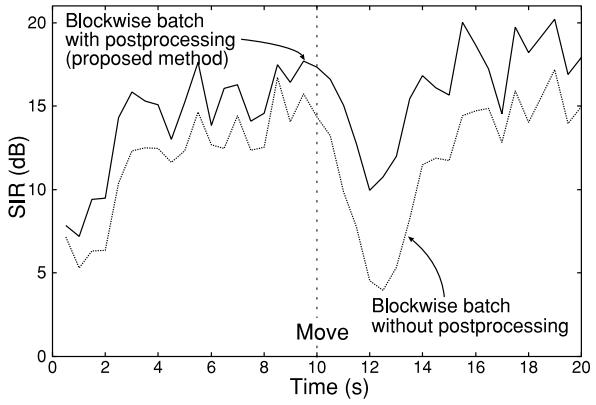

Fig. 11 Effect of postprocessing. Jammer signal moved from C to D at 10 sec. (T_b = 1.0 sec.)

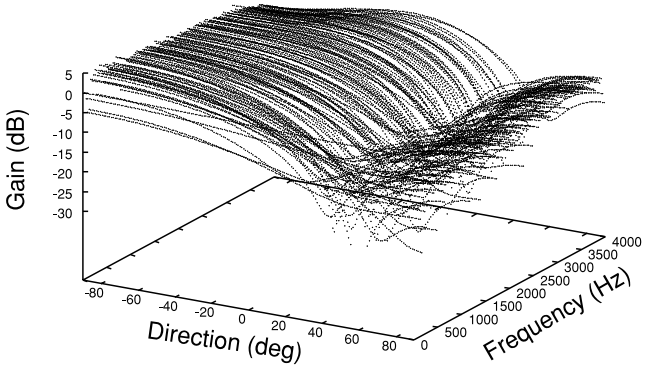

Fig. 10 Directivity pattern of separating system obtained by frequency domain ICA.

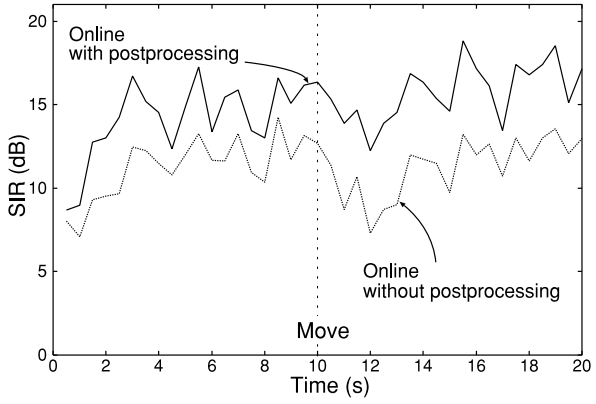

Fig. 12 Performance of online algorithm with and without postprocessing. Jammer signal moved from C to D at 10 sec. (T_b = 1.0 sec.)

separating system obtained by ICA. The solution of frequency domain BSS works in the same way as an adaptive beamformer, which forms a spatial null towards a jammer signal (Fig. 10). Because of this characteristic, BSS using ICA is robust as regards a moving target signal but fragile with respect to a moving jammer signal.

4.4 Performance of Blockwise Batch Algorithm with Postprocessing

The most important factor when estimating the crosstalk component $Y_1^{(c)}$ using (14) and (15) is Y_2, and Y_2 is estimated robustly even when S_2 moves, because S_2 is a target signal for Y_2. Therefore, postprocessing works robustly even when the jammer signal S_2 moves.

Figure 11 shows the SIR of blockwise batch algorithm with postprocessing when the jammer signal moves (solid line). We can see that the SIR is improved by the postprocessing, and the drop of the SIR when the jammer moves is reduced. This result shows that our postprocessing method can compensate the fragility of the blockwise batch algorithm when a jammer signal moves. Although crosstalk components still remaining in the postprocessed output signal sometimes make a musical noise, the power is much smaller than ordinary spectral subtraction.

4.5 Performance of Online Algorithm

Figure 12 shows the SIR of online algorithm with and without postprocessing. The online algorithm is more stable than blockwise algorithm, however the performance is worse when the sources are stationary, as we described in Sect. 4.2. The postprocessing is also effective for this case, thus we may choose the algorithm in the first stage according to requirements of the application.

5. Conclusion

We proposed a robust real-time BSS method for moving source signals. The combination of the blockwise batch and the postprocessing realizes a robust low-delay real-time BSS. We can solve a permutation problem quickly by using analytical calculation of null directions, and this technique is useful for solving convolutive BSS problems in realtime. Postprocessing using crosstalk component estimation and non-stationary spectral subtraction improves the separation performance and reduces the performance deterioration when a jammer signal moves. Experimental results using speech signals recorded in a room showed the effec-

tiveness of the proposed method. Some sound examples can be found on our web site [23].

Acknowledgement

We thank Dr. Shigeru Katagiri for his continuous encouragement.

References

[1] A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," Neural Comput., vol.7, no.6, pp.1129–1159, 1995.
[2] S. Haykin, ed., Unsupervised Adaptive Filtering, John Wiley & Sons, 2000.
[3] T.W. Lee, Independent Component Analysis, Kluwer Academic Publishers, 1998.
[4] A. Cichocki and S. Amari, Adaptive Blind Signal and Image Processing, John Wiley & Sons, 2002.
[5] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
[6] J. Anemüller and T. Gramss, "On-line blind separation of moving sound sources," Proc. Intl. Conf. on Independent Component Analysis and Blind Source Separation (ICA'99), pp.331–334, 1999.
[7] A. Koutras, E. Dermatas, and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environment," Proc. ICASSP 2000, pp.1133–1136, 2000.
[8] I. Kopriva, Z. Devcic, and H. Szu, "An adaptive short-time frequency domain algorithm for blind separation of non-stationary convolved mixtures," Proc. IJCNN 2001, pp.424–429, 2001.
[9] K.E. Hild II, D. Erdogmus, and J.C. Principe, "Blind source separation of time-varying, instantaneous mixtures using an on-line algorithm," Proc. ICASSP 2002, pp.993–996, 2002.
[10] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," Proc. ISSPA 2003, vol.2, pp.411–414, 2003.
[11] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," Proc. Eurospeech 2001, pp.2595–2598, 2001.
[12] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113–120, April 1979.
[13] M. Mizumachi and M. Akagi, "Noise reduction by paired-microphones using spectral subtraction," Proc. ICASSP'98, pp.1001–1004, 1998.
[14] Q. Zou, X. Zou, M. Zhang, and Z. Lin, "A robust speech detection algorithm in a microphone array teleconferencing system," ICASSP 2001, pp.3025–3028, 2001.
[15] S. Amari, A. Cichocki, and H.H. Yang, "A new learning algorithm for blind signal separation," in Advances in Neural Information Processing Systems 8, pp.757–763, MIT Press, 1996.
[16] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," Proc. ICASSP 2002, pp.1001–1004, 2002.
[17] F. Asano and S. Ikeda, "Evaluation and real-time implementation of blind source separation system using time-delayed decorrelation," Proc. Intl. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'00), pp.411–415, 2000.
[18] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual crosstalk components in blind source separation using LMS filters," Proc. NNSP 2002, pp.435–444, 2002.
[19] R. Mukai, S. Araki, and S. Makino, "Separation and dereverberation performance of frequency domain blind source separation," Proc. Intl. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'01), pp.230–235, 2001.
[20] S. Haykin, Adaptive Filter Theory, Prentice Hall, 2002.
[21] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," Acoust. Sci. & Tech., vol.22, no.2, pp.149–157, Feb. 2001.
[22] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," Proc. ICASSP 2002, pp.529–532, 2002.
[23] http://www.kecl.ntt.co.jp/icl/signal/mukai/demo/ieice2004/

Ryo Mukai received the B.S. and the M.S. degrees in information science from the University of Tokyo, Japan, in 1990 and 1992, respectively. He joined NTT in 1992. From 1992 to 2000, he was engaged in research and development of processor architecture for network service systems and distributed network systems. Since 2000, he has been with NTT Communication Science Laboratories, where he is engaged in research of blind source separation. His current research interests include digital signal processing and its applications. He is a member of the IEEE, ACM, the ASJ, and the IPSJ.

Hiroshi Sawada received the B.E., M.E. and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1991, 1993 and 2001, respectively. In 1993, he joined NTT Communication Science Laboratories. From 1993 to 2000, he was engaged in research on the computer aided design of digital systems, logic synthesis, and computer architecture. Since 2000, he has been engaged in research on signal processing and blind source separation for convolutive mixtures using independent component analysis. He received the best paper award of the IEEE Circuit and System Society in 2000. He is a member of the IEEE and the ASJ.

Shoko Araki received the B.E. and the M.E. degrees in mathematical engineering and information physics from the University of Tokyo, Japan, in 1998 and 2000, respectively. Her research interests include array signal processing, blind source separation applied to speech signals, and auditory scene analysis. She is a member of the IEEE and the ASJ.

Shoji Makino received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981. He is now an Executive Manager at the NTT Communication Science Laboratories. His research interests include blind source separation of convolutive mixtures of speech, acoustic signal processing, and adaptive filtering and its applications. He received the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Best Paper Award of the IWAENC in 2003, the Paper Award of the IEICE in 2002, the Paper Award of the ASJ in 2002, the Achievement Award of the IEICE in 1997, and the Outstanding Technological Development Award of the ASJ in 1995. He is the author or co-author of more than 170 articles in journals and conference proceedings and has been responsible for more than 140 patents. He is a member of the Conference Board of the IEEE SP Society and an Associate Editor of the IEEE Transactions on Speech and Audio Processing. He is a member of the Technical Committee on Audio and Electroacoustics of the IEEE SP Society. Dr. Makino is a Fellow of the IEEE, a member of the ASJ.