REMOVAL OF RESIDUAL CROSSTALK COMPONENTS IN BLIND SOURCE SEPARATION USING LMS FILTERS

Ryo Mukai Shoko Araki Hiroshi Sawada Shoji Makino
NTT Communication Science Laboratories, NTT Corporation,
2–4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619–0237, Japan
{ryo,shoko,sawada,maki}@cslab.kecl.ntt.co.jp

Abstract. The performance of Blind Source Separation (BSS) using Independent Component Analysis (ICA) declines significantly in a reverberant environment. The degradation is mainly caused by the residual crosstalk components derived from the reverberation of the jammer signal. This paper describes a post-processing method designed to refine output signals obtained by BSS.

We propose a new method which uses LMS filters in the frequency domain to estimate the residual crosstalk components in separated signals. The estimated components are removed by non-stational spectral subtraction. The proposed method removes the residual components precisely, thus it compensates for the weakness of BSS in a reverberant environment.

Experimental results using speech signals show that the proposed method improves the signal-to-interference ratio by 3 to 5 dB.

INTRODUCTION

Blind Source Separation (BSS) is a technique for estimating original source signals using only observed mixtures of signals. Independent Component Analysis (ICA) is a typical BSS method that is effective for instantaneous (non-convolutive) mixtures [3, 5, 7]. However, the performance of BSS using ICA declines significantly in a reverberant environment [2, 9]. In our recent research [11], we analyzed the separation and dereverberation performance of a separating system obtained by ICA using impulse responses, and revealed that, although the system can completely remove the direct sound of jammer signals, it cannot remove the reverberation, and this is one of the main causes of the deterioration in performance.

We have also shown that when we use a long filter to cover the reverberation, the performance becomes poor with frequency domain BSS [2]. This is because the number of data in each frequency bin becomes small, when we

0-7803-7616-1/02/$17.00 (c) 2002 IEEE

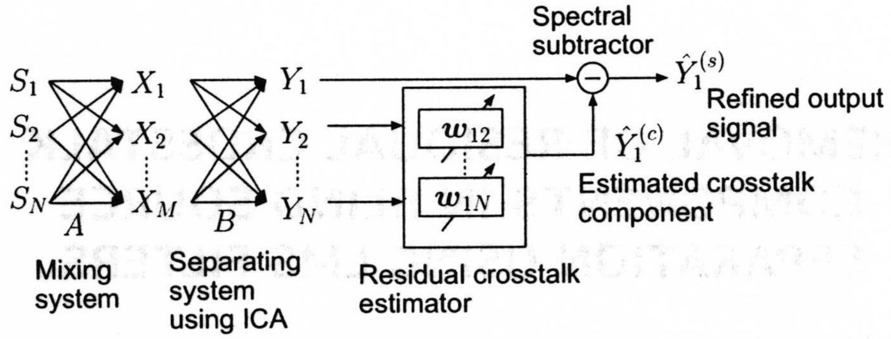

Figure 1: Block diagram of proposed system (for $i = 1$)

use a longer frame.

Previously, we proposed a post-processing method for BSS using time delay and attenuation parameters to estimate and remove residual crosstalk components [12]. The method utilized the nature of BSS in which residual crosstalk components are derived from reverberation.

In this paper, we propose a new method for refining output signals obtained by BSS. We introduce LMS filters in the frequency domain to model and estimate the residual crosstalk components. The filters are prepared for every frequency bin and combination of channels. The estimated residual crosstalk components are subtracted by non-stational spectral subtraction. The new method is a generalized version of our previous method.

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 [4], 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.

Our method compensates for the weakness of BSS in a reverberant environment. We show the effect of the proposed method with experimental results obtained using speech signals.

BLIND SOURCE SEPARATION OF CONVOLUTIVE MIXTURES USING FREQUENCY DOMAIN ICA

In this section, we briefly review the algorithm of BSS using frequency domain ICA.

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

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

$$y_i(t) = \sum_{j=1}^{M} (\boldsymbol{b}_{ij} * x_j)(t) \tag{2}$$

where $\boldsymbol{a}_{ji}$ is the impulse response from source i to microphone j, $\boldsymbol{b}_{ij}$ is the coefficient when we assume that a separating system is used as an FIR filter, and $*$ denotes the convolution operator.

A convolutive mixture in the time domain corresponds to an instantaneous mixture 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 for (1), we obtain

$$\boldsymbol{X}(\omega,n) = \boldsymbol{A}(\omega)\boldsymbol{S}(\omega,n). \tag{3}$$

The separating process can be formulated in each frequency bin ω as:

$$\begin{aligned} \boldsymbol{Y}(\omega,n) &= \boldsymbol{B}(\omega)\boldsymbol{X}(\omega,n) & (4)\\ &= \boldsymbol{B}(\omega)\boldsymbol{A}(\omega)\boldsymbol{S}(\omega,n), & (5) \end{aligned}$$

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

For the calculation of separating matrix $\boldsymbol{B}$, we use an optimization algorithm based on the minimization of the mutual information of $\boldsymbol{Y}$. The optimal $\boldsymbol{B}$ is obtained by using the following iterative equation:

$$\boldsymbol{B}_{i+1} = \boldsymbol{B}_i + \mu[\boldsymbol{I} - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^H\rangle]\boldsymbol{B}_i, \tag{6}$$

where i is an index for the iteration, $\boldsymbol{I}$ is an identity matrix, μ is a step size parameter, $\langle\cdot\rangle$ denotes the averaging operator, and $\Phi(\cdot)$ is a non-linear function. Because the signals are complex valued in the frequency domain, we use a polar-coordinated based non-linear function [14]:

$$\Phi(\boldsymbol{Y}) = \tanh(g \cdot \mathrm{abs}(\boldsymbol{Y}))e^{\jmath \arg(\boldsymbol{Y})}, \tag{7}$$

where g is a gain parameter to control the nonlinearity.

The above calculations are carried out separately for each frequency. After ICA is solved in all frequency, we need to solve the permutation and scaling problem. We use the method in [10] to solve permutation and scaling (phase) problem, and the method in [8] to solve scaling (amplitude) problem.

RESIDUAL CROSSTALK ESTIMATION USING LMS FILTERS IN THE FREQUENCY DOMAIN

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.

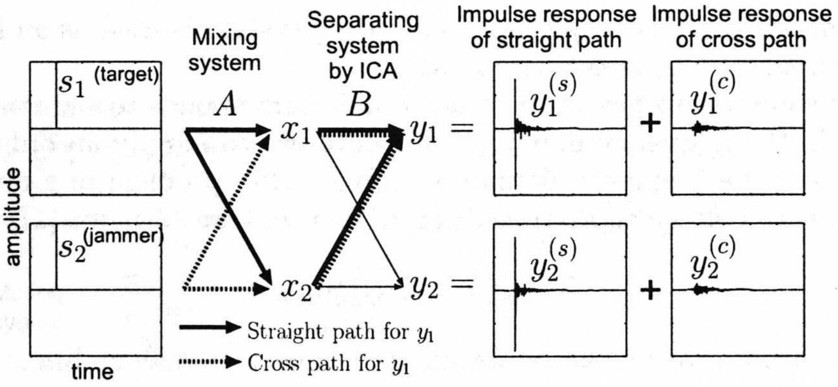

Figure 2: Impulse responses of straight path and cross path

Straight and crosstalk components of BSS

When we denote the concatenation of a mixing system and a separating system as $\boldsymbol{G}$, *i.e.*, $\boldsymbol{G} = \boldsymbol{BA}$, 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

$$\begin{aligned} Y_i(\omega, n) &= Y_i^{(s)}(\omega, n) + Y_i^{(c)}(\omega, n) && (9) \\ Y_i^{(s)}(\omega, n) &= G_{ii}(\omega) S_i(\omega, n) && (10) \\ Y_i^{(c)}(\omega, n) &= \sum_{j \neq i} G_{ij}(\omega) S_j(\omega, n). && (11) \end{aligned}$$

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 [11], 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. 2). We utilize these characteristics of separated signals to estimate the crosstalk components.

Model of residual crosstalk component estimation

Figure 3 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

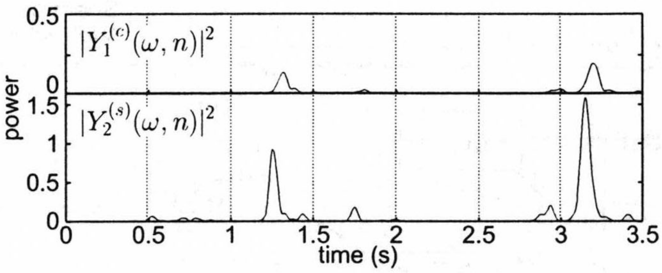

Figure 3: Example of narrow band power spectrum of straight and crosstalk components (ω = 320 Hz)

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 $Y_2^{(s)}$ 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 $\boldsymbol{w}_{ij}(\omega,n) = [w_{ij0}(\omega,n), ..., w_{ijL-1}(\omega,n)]^T$ 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:

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

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

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

Adaptive algorithm and spectrum estimation

Figure 4 shows a block diagram of the proposed method for one output channel. We estimate filters $\boldsymbol{w}_{ij}$ described in the previous section by using an adaptive algorithm based on the normalized LMS (NLMS) algorithm [6].

The filters $\hat{\boldsymbol{w}}_{ij}(\omega,n)$ are adapted so that the sum of the output signals becomes $|Y_i^{(c)}(\omega,n)|^\beta$ for input signals $|Y_j^{(s)}(\omega,n)|^\beta$. 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 de-

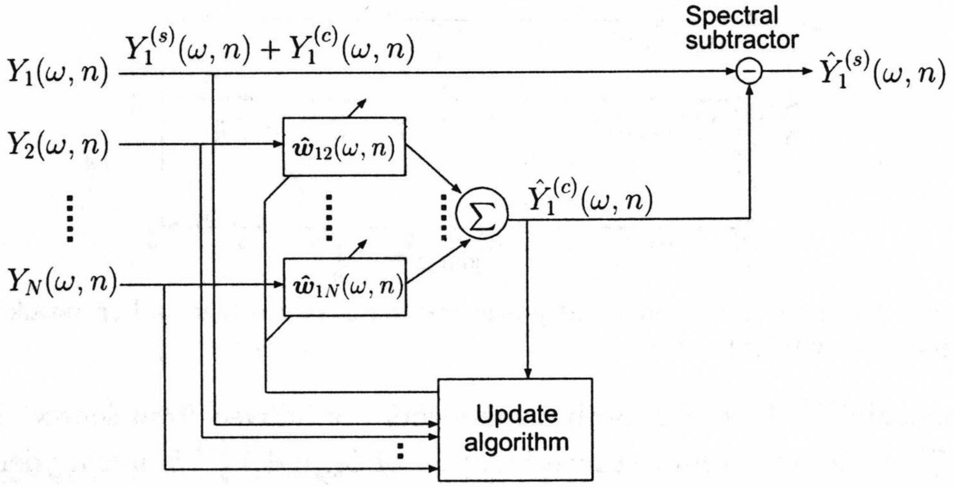
$Y_1(\omega,n)$
$Y_1^{(s)}(\omega,n)+Y_1^{(c)}(\omega,n)$
Spectral subtractor
$\hat{Y}_1^{(s)}(\omega,n)$
$Y_2(\omega,n)$
$\hat{\boldsymbol{w}}_{12}(\omega,n)$
Σ
$\hat{Y}_1^{(c)}(\omega,n)$
$Y_N(\omega,n)$
$\hat{\boldsymbol{w}}_{1N}(\omega,n)$
Update algorithm

Figure 4: Adaptive filters and spectral subtractor to estimate $Y_i^{(s)}$ (for $i=1$)

tailed analysis of overlapping frequency components of speech signals can be found in [1] and [13].

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

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

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

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

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

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

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

TABLE 1: EXPERIMENTAL CONDITIONS

Common	Sampling rate = 8 kHz Window = hanning Reverberation time t_{R}=150 ms, 300 ms Length of source signal = 6 s
ICA part	Frame length T_{ICA}= 32, 512 points (4, 64 ms) Frame shift = frame length / 4 $\mu = 0.1$, $g = \infty$ Number of iterations = 100
NLMS & Spectral subtraction part	Frame length T_{SS}= 1024 points (128 ms) Frame shift = 64 points (8 ms) Filter length L = 16 $\mu = 0.1$, $\delta = 0.01$, $\beta = 2$

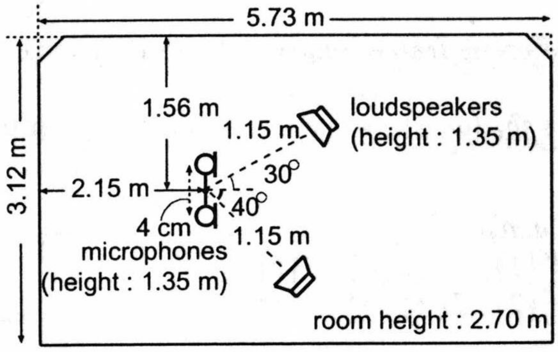

Figure 5: Layout of room used in experiments

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

EXPERIMENTS

To examine the effectiveness of the proposed method, we carried out experiments for $N = M = 2$ using speech signals convolved with impulse responses measured in a room.

Experimental conditions

The layout of the room we used to measure the impulse responses of the mixing system $\boldsymbol{A}$ is shown in Fig. 5. We used a two-element microphone array with an inter-element spacing of 4 cm. The directions of the source signals were −30° and 40°. Other conditions are summarized in Table 1. To investigate the influence of the ICA performance on the effectiveness of the proposed method, we used two different reverberation times (t_{R}= 150 ms and

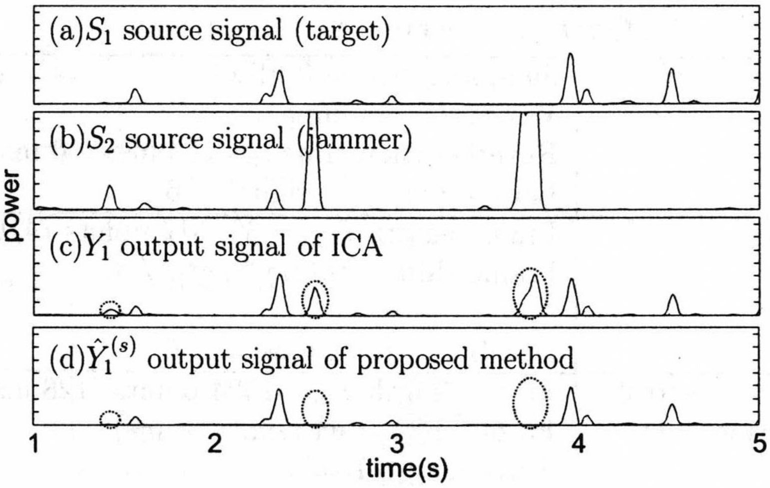

Figure 6: Example of narrow band power spectrum of source and output signals (ω=440 Hz, t_R=150 ms, T_{ICA}= 32, T_{SS}= 1024)

300 ms) and two different frame lengths of the ICA part (T_{ICA}= 32 and 512 points).

To update filters $\hat{\boldsymbol{w}}_{ij}(\omega, n)$, 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 as a signal, and the difference between the output signal and the straight component as a noise. We defined the output signal-to-interference ratio (SIR_O) in the time domain as follows:

$$\mathrm{SIR}_{Oi} \equiv 10 \log \frac{\sum_t |y_i^{(s)}(t)|^2}{\sum_t |\hat{y}_i^{(s)}(t) - y_i^{(s)}(t)|^2} \ (\mathrm{dB}). \qquad (17)$$

We used the average of SIR_{O1} and SIR_{O2} as a performance measure in order to cancel out the input SIR. This measurement is consistent with the performance evaluation of BSS in which the crosstalk components are assumed as noise.

For each $t_R$and T_{ICA}, we measured SIRs with 42 combinations of source signals using two male and two female speakers.

Experimental results

Before viewing the results of the SIR evaluation, let us investigate one example of a narrow band power spectrum of source and output signals; Figs. 6(a) and (b) show the source signals, Fig. 6(c) the output signal of ICA, and Fig. 6(d) the output signal of the proposed method. By comparing Figs. 6(c) and (d), we can see that the residual crosstalk component indicated by the dashed circles (which is derived from S_2) is properly removed and also that there is no over-subtraction or under-subtraction.

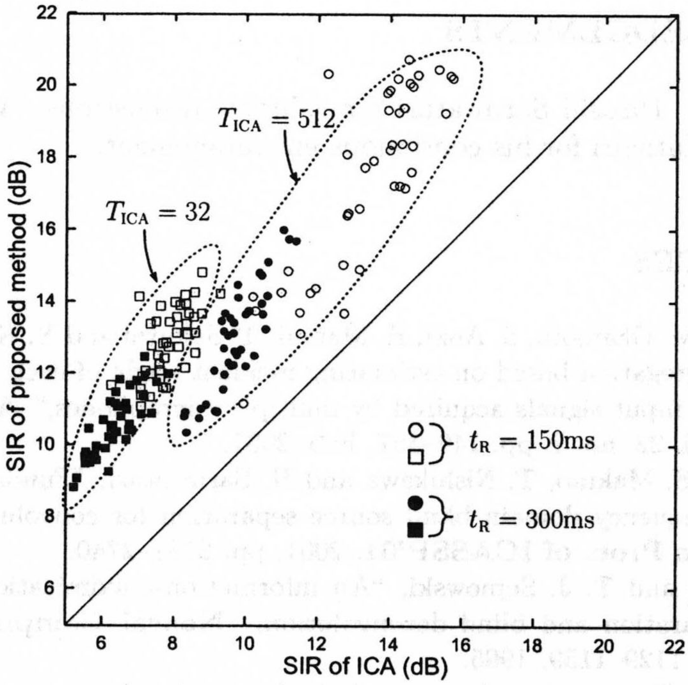

Figure 7: Comparison of SIR for ICA and proposed method

TABLE 2: AVERAGE SIR

t_{R}	T_{ICA}	ICA (dB)	Proposed method(dB)	Improvement (dB)
150 ms	32	7.9	12.9	5.0
	512	13.4	17.4	4.0
300 ms	32	6.4	10.6	4.2
	512	9.7	13.0	3.3

The results of the SIR evaluation are shown in Fig. 7. The horizontal axis denotes the SIR of ICA and the vertical axis denotes the SIR of the proposed method. Each point corresponds to one combination of source signals.

The proposed method achieved better performance than ICA for all combinations. The improvement of the SIR was stable even when the separation performance of ICA was bad. Table 2 shows the average SIR and improvement for each frame length and reverberation time.

CONCLUSION

We proposed a method for estimating and subtracting residual crosstalk components in separated signals obtained by BSS using ICA. The model and algorithm using NLMS filters in the frequency domain estimates residual crosstalk components accurately. This model is based on the nature of BSS in which the crosstalk components in the output signals are derived from reverberation. Experimental results using mixed speech signals proved the effectiveness of the method in compensating for weaknesses of BSS in a reverberant environment.

ACKNOWLEDGEMENTS

We thank Dr. Hiroshi Saruwatari for valuable discussions. We also thank Dr. Shigeru Katagiri for his continuous encouragement.

REFERENCES

[1] 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.

[2] S. Araki, S. Makino, T. Nishikawa and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," in **Proc. of ICASSP'01**, 2001, pp. 2737–2740.

[3] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," **Neural Computation**, vol. 7, no. 6, pp. 1129–1159, 1995.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," **IEEE Trans. Acoust., Speech, and Signal Processing**, vol. ASSP-27, no. 2, pp. 113–120, April 1979.

[5] S. Haykin (ed.), **Unsupervised Adaptive Filtering**, John Wiley & Sons, 2000.

[6] S. Haykin, **Adaptive Filter Theory**, Prentice Hall, 2002.

[7] A. Hyvärinen, J. Karhunen and E. Oja, **Independent Component Analysis**, John Wiley & Sons, 2001.

[8] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in **Proc. of ICA'99**, 1999, pp. 365–370.

[9] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in **Proc. of ICASSP'00**, 2000, pp. 1041–1044.

[10] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in **Proc. of ICASSP'00**, 2000, pp. 3140–3143.

[11] R. Mukai, S. Araki and S. Makino, "Separation and dereverberation performance of frequency domain blind source separation," in **Proc. of Intl. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2001)**, 2001, pp. 230–235.

[12] R. Mukai, S. Araki, H. Sawada and S. Makino, "Removal of residual cross-talk components in blind source separation using time-delayed spectral subtraction," in **Proc. of ICASSP'02**, 2002, accepted.

[13] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in **Proc. of ICASSP'02**, 2002.

[14] H. Sawada, R. Mukai, S. Araki and S. Makino, "A polar-coordinate based activation function for frequency domain blind source separation," in **Proc. of Intl. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2001)**, 2001, pp. 663–668.