REMOVAL OF RESIDUAL CROSS-TALK COMPONENTS IN BLIND SOURCE SEPARATION USING TIME-DELAYED SPECTRAL SUBTRACTION

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

This paper describes a post processing method to refine output signals obtained by Blind Source Separation (BSS). The performance of BSS using Independent Component Analysis (ICA) declines significantly in a reverberant environment. The degradation is mainly caused by the cross-talk components derived from the reverberation of the jammer signal. Utilizing this knowledge, we propose a new method, time-delayed non-stationary spectral subtraction, which removes the residual components from the separated signals precisely. The proposed method compensates for the weakness of BSS in a reverberant environment. Experimental results using speech signals show that the proposed method improves the signal-to-noise ratio by 3 to 5 dB.

1. INTRODUCTION

Blind Source Separation (BSS) is a technique that estimates original source signals using only observed mixtures of signals. BSS using Independent Component Analysis (ICA) is quite effective for instantaneous (non-convolutive) mixtures [1][2]. However, the separation performance declines significantly in a reverberant environment [3][4]. In recent research [5], we analyzed the separation and dereverberation performance of an unmixing 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 cause of the performance deterioration.

In this paper, we propose a method to improve the separation performance of BSS by subtracting residual cross-talk components precisely. The proposed method compensate the weakness of BSS in a reverberant environment. The effect of the proposed method is shown by experimental results using speech signals.

Figure 1 shows a block diagram of the proposed method for one output channel. In contrast to the original spectral subtraction [6], which assumes stationary noise and periods with no target signal to estimate the noise spectrum, our method requires neither assumption, because we use BSS in the first stage. Utilizing the nature of BSS that residual cross-talk components are derived from reverberation, we introduce two parameters, *i.e.*, time delay and leakage coefficient, in order to model residual cross-talk components. Our model differs from a simple attenuation-and-delay model [7] because the parameters are estimated for every frequency bin and combination of channels.

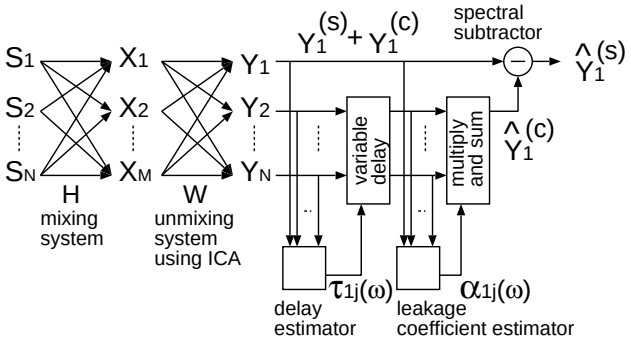

Fig. 1. Block diagram of proposed system (for $i = 1$)

2. FREQUENCY DOMAIN BSS OF CONVOLUTIVE MIXTURES

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 unmixed signals are $y_i(t)(1 \leq i \leq N)$, the BSS model can be described by the following equations:

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

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

where $\boldsymbol{h}_{ji}$ is the impulse response from source i to microphone j, $\boldsymbol{w}_{ij}$ is the coefficient when the unmixing system is assumed as an FIR filter, and $*$ denotes the convolution operator. To simplify the problem, we assume that the permutation problem is solved so that the i-th source signal $s_i(t)$ is separated as the i-th output signal $y_i(t)$.

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

$$\boldsymbol{X}(\omega, t) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega, t). \tag{3}$$

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

$$\boldsymbol{Y}(\omega, t) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, t) \tag{4}$$

$$= \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)\boldsymbol{S}(\omega, t) \tag{5}$$

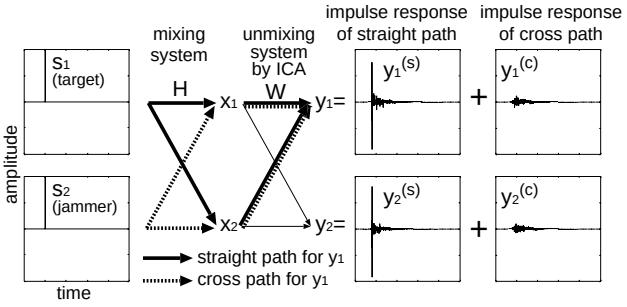

Fig. 2. Impulse responses of straight path and cross path

where $\boldsymbol{S}(\omega,t) = [S_1(\omega,t),...,S_N(\omega,t)]^T$ is the source signal in frequency bin ω, $\boldsymbol{X}(\omega,t) = [X_1(\omega,t),...,X_M(\omega,t)]^T$ denotes the observed signals, $\boldsymbol{Y}(\omega,t) = [Y_1(\omega,t),...,Y_N(\omega,t)]^T$ is the estimated source signal, and $\boldsymbol{W}(\omega)$ represents the unmixing matrix. $\boldsymbol{W}(\omega)$ is determined so that $Y_i(\omega,t)$ and $Y_j(\omega,t)$ become mutually independent. The above calculations are carried out for each frequency independently.

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

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

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 [8]:

$$\Phi(\boldsymbol{Y}) = \tanh(g \cdot \mathrm{abs}(\boldsymbol{Y}))e^{j\arg(\boldsymbol{Y})} \qquad (7)$$

where g is a gain parameter.

3. TIME DELAYED NON-STATIONARY SPECTRAL SUBTRACTION

3.1. Straight and cross-talk components of BSS

When we denote the concatenation of a mixing system and an unmixing system as $\boldsymbol{G}$, *i.e.*, $\boldsymbol{G} = \boldsymbol{WH}$, each of the separated signals Y_i obtained by BSS can be described as follows:

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

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

$$Y_i(\omega,t) = Y_i^{(s)}(\omega,t) + Y_i^{(c)}(\omega,t) \qquad (9)$$

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

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

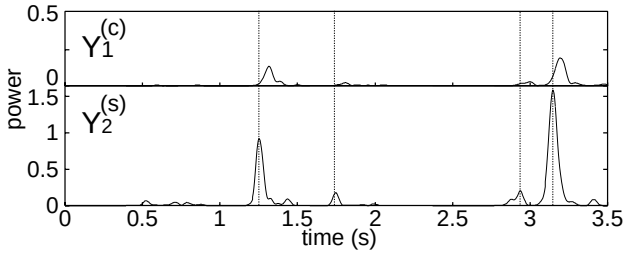

Fig. 3. Example of narrow band power spectrum of straight and cross-talk components (ω = 320Hz)

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 to obtain $\hat{Y}_i^{(s)}$ by subtracting $\hat{Y}_i^{(c)}$ from Y_i.

In previous research [5], we measured the impulse responses of the straight path and cross path 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 cross-talk components are derived from the reverberation (Fig. 2). We utilize these characteristics of separated signals to estimate cross-talk components.

3.2. Model of residual cross-talk component estimation

Figure 3 shows an example of a narrow band power spectrum of straight and cross-talk components in separated signals obtained by a 2-input 2-output BSS system. The cross-talk 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, and $Y_1^{(c)}$ 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 cross-talk component can be approximated by the straight component delayed and attenuated according to some delay parameter and leakage coefficient.

We extend this approximation to the case of multiple signals by introducing delay parameters $\tau_{ij}(\omega)$ and leakage coefficients $\alpha_{ij}(\omega)$ for each frequency bin and combination of channels. Furthermore, we use Y_j as an approximation of $Y_j^{(s)}$, because $Y_j^{(s)}$ is actually unknown. Therefore, the model of estimating residual cross-talk components is formulated as follows:

$$|\hat{Y}_i^{(c)}(\omega,t)|^\beta \approx \sum_{j\neq i} \alpha_{ij}(\omega)|Y_j(\omega,t-\tau_{ij}(\omega))|^\beta \qquad (12)$$

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

3.3. Parameter and spectrum estimation

Each of delay parameters $\tau_{ij}(\omega)$ is determined so that the correlation between $|Y_i^{(c)}(t)|$ and $|Y_j^{(s)}(t-\tau_{ij})|$ becomes maximum. In this procedure, unknown components $Y_i^{(c)}$ and $Y_j^{(s)}$ are substituted by Y_i and Y_j respectively. This

substitution is based on the assumption that narrow band signals $Y_i^{(s)}$ and $Y_j^{(s)}$ seldom have large power at the same time, and $|Y_i(t)||Y_j(t-\tau)|$ $(\tau > 0)$ can be approximated by $|Y_i^{(c)}(t)||Y_j^{(s)}(t-\tau)|$ almost anywhere, especially when the source signals are speech signals. The detailed analysis of overlapping frequency components of speech signals can be found in [9].

We calculate the correlation during the period in which $|Y_i(t)| \leq |Y_j(t-\tau)|$. This means that we determine the direction of the leakage according to the magnitudes of the signals. Therefore, the procedure to determine $\tau_{ij}(\omega)$ is formulated as follows:

$$\tau_{ij}(\omega) = \underset{0<\tau\leq\tau_{max}}{\operatorname{argmax}} \sum_{t\in U_{ij}(\omega)} |Y_i(\omega,t)||Y_j(\omega,t-\tau)| \quad (13)$$

where $U_{ij}(\omega) = \{t : |Y_i(\omega,t)| \leq |Y_j(\omega,t-\tau)|\}$ is a set of time and τ_{max} is a parameter that is determined according to the maximum reverberation time expected.

The leakage coefficient $\alpha_{ij}(\omega)$ is calculated using the correlation between $|Y_i^{(c)}(t)|^\beta$ and $|Y_j^{(s)}(t-\tau_{ij})|^\beta$, and the ratio of the total power (or magnitude) of these signals. In this procedure, unknown components $Y_i^{(c)}$ and $Y_j^{(s)}$ are substituted by Y_i and Y_j again. Therefore, $\alpha_{ij}(\omega)$ is estimated by

$$\alpha_{ij}(\omega) = \frac{\sum |Y_i(\omega,t)|^\beta |Y_j'(\omega,t)|^\beta}{\sqrt{\sum |Y_i(\omega,t)|^{2\beta}}\sqrt{\sum |Y_j'(\omega,t)|^{2\beta}}} \cdot \frac{\sum |Y_i(\omega,t)|^\beta}{\sum |Y_j'(\omega,t)|^\beta} \quad (14)$$

where $Y_j'(\omega,t) = Y_j(\omega, t-\tau_{ij}(\omega))$ is a delayed signal, and the summation $\sum$ is executed at time $t \in U_{ij}'(\omega)$, *i.e.*, $\sum$ means $\sum_{t\in U_{ij}'(\omega)}$, where $U_{ij}'(\omega) = \{t : |Y_i(\omega,t)| \leq |Y_j'(\omega,t)|\}$. The first factor of (14) is the correlation of Y_i and Y_j', and the second factor is the ratio of these signals. When the correlation equals to 1, (14) gives such α_{ij} that exactly eliminates Y_i at the time $t \in U_{ij}'$.

Finally, we estimate the spectrum of $Y_i^{(c)}$ using (12), (13), and (14), and obtain the estimation of straight component as $\hat{Y}_i^{(s)}$ by the following spectral subtraction procedure:

$$\hat{Y}_i^{(s)}(\omega,t) = \begin{cases} (|Y_i(\omega,t)|^\beta - |\hat{Y}_i^{(c)}(\omega,t)|^\beta)^{1/\beta} \dfrac{Y_i(\omega,t)}{|Y_i(\omega,t)|} \\ \qquad\qquad \text{(if } |Y_i(\omega,t)| > |\hat{Y}_i^{(c)}(\omega,t)|\text{)} \\ 0 \qquad\qquad \text{(otherwise)} \end{cases} \quad (15)$$

4. EXPERIMENTS

In order to examine the effectiveness of the proposed method, we carried out experiments in the case of $N = M = 2$ using speech signals convolved with impulse responses.

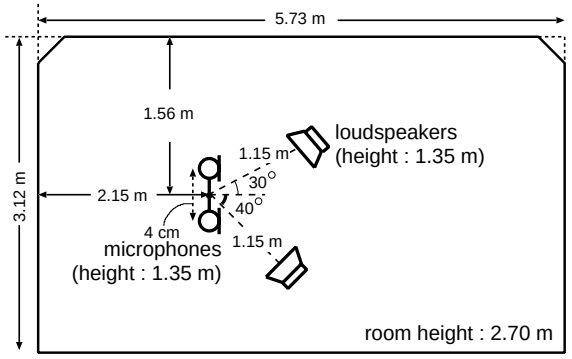

Fig. 4. Layout of room used in experiments

Table 1. Experimental conditions

Common	Sampling rate = 8kHz
	Window = hanning
	Reverberation time T_{rev}=150ms, 300ms
	Length of source signal = 6 s
ICA part	Frame length T_{ICA} = 32, 512 points (4, 64ms)
	Frame shift = frame length / 4
	μ = 0.1, g = 100.0
	Number of iterations = 100
Spectral subtraction part	Frame length T_{SS}=1024 points (128ms)
	Frame shift = 64 points (8ms)
	β = 2

4.1. Conditions for the experiments

The layout of the room we used to measure the impulse responses of the mixing system $\boldsymbol{H}$ is shown in Fig. 4. We used a two-element microphone array with an inter-element spacing of 4 cm. The directions of source signals were -30° and 40°. Other conditions are summarized in Table 1. In order to investigate the influence of the ICA performance on the performance of the proposed method, we used two different reverberation time T_{rev} and two different frame length of the ICA part T_{ICA}.

We assumed the straight component as a signal, and the difference between the output signal and the straight component as a noise, and defined the output signal-to-noise ratio (SNR$_O$) as follows:

$$\text{SNR}_{Oi} \equiv 10\log \frac{|y_i^{(s)}|^2}{|\hat{y}_i^{(s)} - y_i^{(s)}|^2} \text{ (dB)}. \quad (16)$$

We used the average of SNR$_{O1}$ and SNR$_{O2}$ as a performance measure in order to cancel the input SNR. This measurement is consistent with the performance evaluation of BSS in which the cross-talk components are assumed as noise.

For each T_{rev} and T_{ICA}, we measured SNRs with 24 combinations of source signals using two male and two female speakers.

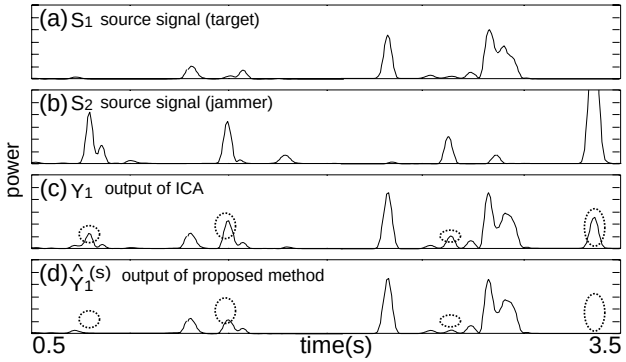

Fig. 5. Example of narrow band power spectrum of source and output signals (ω=1950 Hz, T_{rev}=150ms, T_{ICA}=32, T_{SS}=1024)

4.2. Experimental results

Before viewing the results of the SNR evaluation, let us investigate one example of a narrow band power spectrum of source and output signals; Figs. 5(a) and (b) show the source signals, Fig. 5(c) the output signal of ICA, and Fig. 5(d) the output signal of the proposed method. By comparing Fig. 5(c) and (d), we can see that the residual cross-talk 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.

The results of the SNR evaluation are shown in Fig. 6. The horizontal axis denotes the SNR of ICA, and the vertical axis denotes the SNR of the proposed method. Each point corresponds to one combination of source signals.

The proposed method achieved better performance than ICA for all combinations almost uniformly. For the same frame length T_{ICA}, the improvement of the performance is better, when the SNR of ICA is better. One of the reasons for this can be attributed to the validity of the substitution of $Y_j^{(s)}$ with Y_j in (12). Table 2 shows the average SNR and improvement for each frame length and reverberation time.

5. CONCLUSION

We proposed a method to estimate and subtract residual cross-talk components from separated signals obtained by BSS using ICA. The model, which uses the time delay and leakage coefficient, estimates residual cross-talk components accurately. This model is based on the nature that the cross-talk components in the output signals of BSS are derived from the reverberation. Experimental results using mixed speech signals proved the effectiveness of the method to compensate for weaknesses of BSS in a reverberant environment.

ACKNOWLEDGEMENTS

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

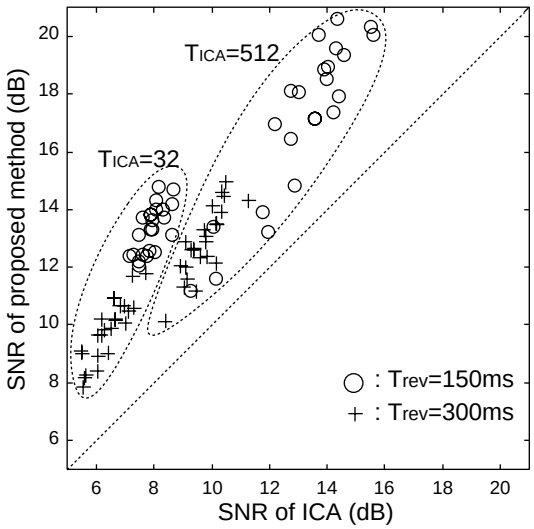

Fig. 6. Comparison of SNR for ICA and proposed method

Table 2. Average SNR

T_{rev}	T_{ICA}	ICA (dB)	Proposed method(dB)	Improvement (dB)
150ms	32	7.9	13.3	5.4
	512	13.2	17.1	3.9
300ms	32	6.4	9.9	3.5
	512	9.8	12.9	3.1

6. REFERENCES

[1] 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.

[2] S. Haykin, Ed., *Unsupervised adaptive filtering*, John Wiley & Sons, 2000.

[3] 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.

[4] 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 ICASSP2001*, 2001, pp. 2737–2740.

[5] R. Mukai, S. Araki, and S. Makino, "Separation and dereverberation performance of frequency domain blind source separation for speech in a reverberant environment," in *Proc. of Eurospeech2001*, 2001, pp. 2599–2602.

[6] 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.

[7] J. Laroche, "Removing preechos from audio recordings," in *Proc. of WASPAA'95*, 1995, pp. 147–150.

[8] 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*, 2001, (accepted).

[9] 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.