Selective Microphone System using Blind Separation by Block Decorrelation of Output Signals

Masakazu Iwaki and Akio Ando

Three-dimensional Audio-visual Systems, NHK Science and Technical Research Laboratories
1-10-11 Kinuta, Setagaya, Tokyo 157-8510, Japan
{ iwaki.m-hy, andou.a-io }@nhk.or.jp

ABSTRACT

This paper describes a new microphone system which separates a target sound from other sounds using blind source separation, when those sounds travel from the same direction. This system divides the observation signals into blocks in the time domain and decorrelates the output signals in each block. The resultant separation process in the block is obtained as a set of matrices even identifying the permutations as the same one. As the number of blocks increases, the intersection of the sets converges to the matrix that can separate the source signals, based on the non-stationarity of the source signal. It was experimentally shown that there is an optimal length of block, in which the stationarity within a block and nonstationarity among blocks are thought to be balanced.

1. INTRODUCTION

There is a need to separate the target sound from background noise in broadcast program production. If the target sound travels from a different direction from that of the noise, it can be effectively separated by a microphone system with sharp directivity. However, if it travels from the same direction as one of the noise sources, signal processing technologies such as blind signal separation (BSS) are necessary to separate it.

In BSS, it is well-known that the separation can be achieved if there is statistical independence among output signals and therefore decorrelation is insufficient to achieve the separation [1]. In contrast, for nonstationary sources, Matsuoka, et al. [2] proposed a neural network which decorrelates the outputs of the network with each other at any instance of time. Since the network minimizes the product of the instantaneous values of output signals, instead of the ensemble average, in the practical implementation, the speed of convergence could be reduced.

This paper describes a new microphone system which separates a target sound from other sounds using the BSS. The system divides the observation signals into blocks in the time domain and decorrelates the output signals in each block. The block length is selected so as to best satisfy both the stationarity within a block and the non-stationarity among blocks. Since the correlation coefficient is estimated by the sample average within the block, rapid convergence is achieved in comparison with the case in which the instantaneous product is used as a substitute for the coefficient. This paper also shows that, under the framework of the system, (1) the decorrelation yields a set of separation matrices, which includes not only the matrix that can separate the sources, but also other matrices that cannot achieve the separation, and (2) because of the nonstationarity of the sources, the intersection of the sets converges to the target matrix, which can separate the source signals, as the number of blocks increases.

2. OVERVIEW OF SYSTEM

This paper assumes that

1) *The mixed signal (mixture of objective sound and noise) can be expressed as a linear combination of the source signals,*
2) *the mixing matrix is nonsingular and constant with time,*
3) *the source signals are statistically independent signals with zero mean,*
4) *the source signals are non-stationary.*

Figure 1 shows the signal sources and the arrangement of two microphones. This arrangement, for instance, can be used for microphones attached to a video camera recorder. In Fig. 1, the two microphones are on the same line as that of the objective sound and noise. To satisfy the assumption 1), we assume here that the sound is picked up in a dead room such as an anechoic room.

There is a time-delay (proportional to inter-microphone distance) in the output of microphone 1, so the two outputs from microphones 1 and 2 are in phase with each other. Further, the amplitudes of these two signal sources, which are picked up by microphones 1 and 2, can be expressed by the ratio of distance. If we define $s^{(i)}$(i=1,2) as the signal source and $x^{(j)}$(j=1,2) as the microphone output, then the following vector gives their relationship:

$$x(t) = As(t), \tag{1}$$

where, A is a mixing matrix.

$$s(t) = \begin{pmatrix} s^{(1)}(t) \\ s^{(2)}(t) \end{pmatrix}, \quad x(t) = \begin{pmatrix} x^{(1)}(t) \\ x^{(2)}(t) \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} . \tag{2}$$

Assuming that the amplitude of sound source 1 is 1 at microphone 1, the entry of the matrix can be expressed as a matrix without delay.

$$\begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & \dfrac{l_1 + l_2}{l_2} \\ \dfrac{l_1 + l_2}{l_1 + l_2 + l_3} & \dfrac{l_1 + l_2}{l_2 + l_3} \end{pmatrix} \begin{pmatrix} s^{(1)} \\ s^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & \alpha \\ k\beta & k \end{pmatrix} \begin{pmatrix} s^{(1)} \\ s^{(2)} \end{pmatrix} \tag{3}$$

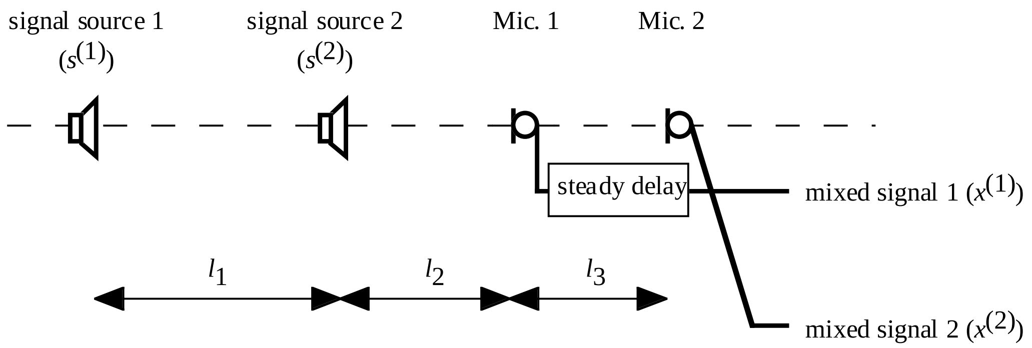

Fig. 1 Signal sources and the arrangement of two microphones. They are on the same line.

where

$$\alpha = \frac{l_1 + l_2}{l_2}, \quad \beta = \frac{l_2 + l_3}{l_1 + l_2 + l_3}, \quad k = \frac{l_1 + l_2}{l_2 + l_3}. \tag{4}$$

The output signal *Y* can be expressed as follows:

$$y(t) = CX, \tag{5}$$

where

$$y = \begin{pmatrix} y^{(1)}(t) \\ y^{(2)}(t) \end{pmatrix}, \quad C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}. \tag{6}$$

3. ALGORITHM

The proposed system divides the observation signal *X* into blocks in the time domain. Thus we introduce the detailed notations. Let $s_t(t)$, $x_t(t)$ and $y_t(t)$ be *S*, *X* and *Y* at the *t*-th sample in the *t*-th block, respectively:

$$s_\tau(t) = \begin{cases} s(\tau + tT); & 0 \le \tau < T, \\ 0; & \text{elsewhere}, \end{cases} \tag{7}$$

$$x_\tau(t) = \begin{cases} x(\tau + tT); & 0 \le \tau < T, \\ 0; & \text{elsewhere}, \end{cases} \tag{8}$$

$$y_\tau(t) = \begin{cases} y(\tau + tT); & 0 \le \tau < T, \\ 0; & \text{elsewhere}, \end{cases} \tag{9}$$

Then, the mixing process and the separation process in each block are described as follows:

$$x_t(\tau) = As_t(\tau) \tag{10}$$

$$y_t(\tau) = Cx_t(\tau), \tag{11}$$

where *A* and *C* denote the mixing and separation matrix, respectively. Let $s_t^{(i)}(t)$ and $y_t^{(i)}(t)$ denote the *i*-th element of the vectors $s_t(t)$ and $y_t(t)$, respectively. Suppose the correlation function of $s_t^{(i)}(t)$ and $s_t^{(j)}(t)$ and the correlation function of $y_t^{(i)}(t)$ and $y_t^{(i)}(t)$ can be estimated by

$$r_s^{(ij)}(t) = \frac{1}{T}\sum_{\tau=0}^{T-1} s_t^{(i)}(\tau) s_t^{(j)}(\tau) \tag{12}$$

and

$$r_y^{(ij)}(t) = \frac{1}{T}\sum_{\tau=0}^{T-1} y_t^{(i)}(\tau) y_t^{(j)}(\tau), \tag{13}$$

respectively. This paper briefly writes $r_s^{(ii)}(t)$ and $r_y^{(ii)}(t)$ as $r_s^{(i)}(t)$ and $r_y^{(i)}(t)$, respectively.

The proposed method finds in each block a set of separation matrices which decorrelates the output signal (i.e. it makes $r_y^{(ij)}(t)$ (i,j=1,...,N; $i \neq j$) equal to 0). Let Γ_t be the resultant set of separation matrices in the *t*-th block. Figure 2 shows an experimental result of the correlation of output signals as a function of c_{12} and c_{21} which are off-diagonal elements of the separation matrix, when *N* is set to 2. As can be seen in Fig. 2, the correlation function takes the value 0 on two hyperbolas in the c_{12}- c_{21} plane. Thus, the decorrelation of output signals yields a (infinite) set of separation matrices, instead of one separation matrix, even identifying the permutations as the same one. It can be concluded, therefore, that the decorrelation generally results in a set of separation matrices.

The proposed method then takes the intersection of Γ_0, Γ_1,..., Γ_t at block t. Let $\hat{\Gamma}$ be the target set of separation matrices that make the output signals statistically independent and therefore can completely separate the source signals. Since a separation matrix, which makes the output signals statistically independent, decorrelates the output signals, we have for each t

$$\Gamma_t \supseteq \hat{\Gamma}. \tag{14}$$

We define the intersection $\tilde{\Gamma}_t$ as

$$\Gamma_t = \bigcap_{i=0}^{t} \Gamma_t. \tag{15}$$

Since the sequence of sets $\tilde{\Gamma}_t$ decreases monotonously, it converges to the limit $\tilde{\Gamma}_\infty$. If

$$\tilde{\Gamma}_\infty = \hat{\Gamma}, \tag{16}$$

the proposed method can separate the source signals (see Fig. 3). This can be shown by proving the following:

PROPOSITION *Under the assumptions*

1) $s_t^{(j)}(t)$ *(i=1,...,N) are statistically independent,*
2) *Mixing matrix A is nonsingular and constant with time,*
3) $r_s^{(i)}(t)/r_s^{(j)}(t)$ $(i,j=1,...,N;\ i\neq j)$ *are not constant with time,*

if $r_y^{(ii)}(t)$ $(i,j=1,...,N;\ i\neq j)$ *are all zero at any block t, then there exists a diagonal matrix D and a permutation matrix P such that* $C=DPA^{-1}$.

This proposition can be proved based on the idea shown in reference [2]. The proof is shown in [3].

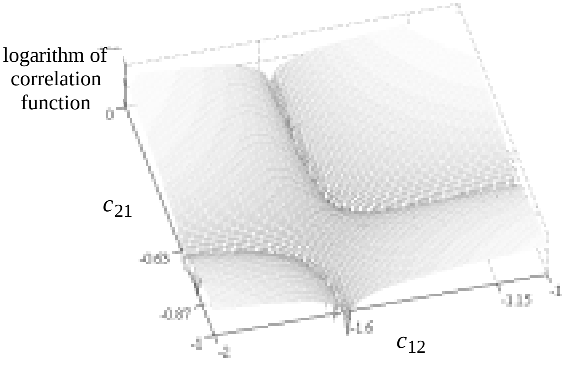

Fig. 2 Example of correlation function of output signals

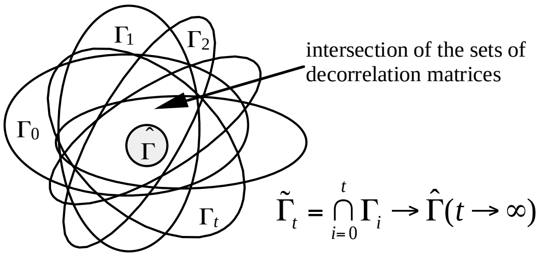
Γ_1
Γ_2
Γ_0
$\hat{\Gamma}$
Γ_t
intersection of the sets of decorrelation matrices
$\tilde{\Gamma}_t = \bigcap_{i=0}^{t} \Gamma_i \to \hat{\Gamma}(t \to \infty)$

Fig. 3 Intersection of the sets of decorrelation matrices converges to the target set.

$\hat{\Gamma}$ is a set of matrices which provide statistical independency among output signals.

4. EXPERIMENT

In the proposed method, the performance of source separation depends on the length of blocks. Longer blocks lead to more accurate estimation of correlation because of the large amount of data (see Eq. (13)). On the other hand, it is well-known in short-time spectral analysis that signals like speech have short-time stationarity [4]. Thus, we can estimate the correlation of the signal from the empirical average (Eq. (13)) within a block in which the stationarity of the signal is maintained. However, if the length of the block becomes longer such that stationarity cannot be assumed in the block, Eq. (13) picks up a kind of global characteristic of the signal and thus nonstationarity among blocks is no longer maintained. Thus longer blocks not only improve the accuracy of estimation, but also reduce the nonstationarity among blocks, which could damage the speed of convergence. Because of this trade-off relation, an optimal length of block is thought to exist, which depends on the nature of source signals. We examined this trade-off through 15 experiments, all combinations of two of six signals, which were white noise, white noise modulated by a 200 Hz sinusoidal function, English male and female speech and Japanese male and female speech. In the experiments we minimized the following objective function instead of decorrelating among output signals:

$$f_t\ (c_{12}, c_{21}) = \left(\sum_{\tau=0}^{T-1} y_t^{(1)}(\tau) y_t^{(2)}(\tau)) \right)^2$$
$$= \left(\sum_{\tau=0}^{T-1} (x_t^{(1)}(\tau) + c_{12} x_t^{(2)}(\tau))(c_{21} x_t^{(1)}(\tau) + x_t^{(2)}(\tau) \right)^2 \qquad (17)$$

where the diagonal elements of C are set to 1. Mixing matrix A was given as

$$A = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \qquad (18)$$

The length of all signals was set to 10 seconds. Sampling frequency and quantization accuracy were 44.1 kHz and 16 bits, respectively. For evaluation, the Noise Reduction Ratio (NRR), which is the difference between output and input S/N ratios (dB), was calculated for each of nine kinds of block length from 1 samples(1/44100 ms) to 32,768 samples (743 ms). Let L be the block length. Figure 4 shows the results where NRR took the maximum value at L=128 samples, and decreased at both shorter and longer L. L=1 was the case in which the instantaneous product is used as the objective function. Therefore, Fig. 4 shows the effectiveness of block processing. Figure 5 shows correlation coefficients among source signals. The correlation monotonously decreased at longer L. In longer L, short-time stationarity is no longer kept and some kind of global characteristic of the signal is extracted. Thus the correlation among the source signals is thought to be decreased. Figure 6 shows the variance among source signals in adjacent blocks. It shows nonstationarity among blocks was not maintained in longer L.

5. CONCLUSION

A blind signal separation method, which divides the observation signals into blocks and decorrelates the output signals in each block, was proposed. We showed in this framework that, with identifying all permutations of the matrices as the same one,

(1) decorrelation of output signals in each block does not yield a definite separation matrix, but a set of separation matrices,

(2) the intersection of the sets of matrices converges to the target separation matrix, which can separate the sources, for nonstationary source signals.

We experimentally showed there is an optimal length of block, in which the stationarity within a block and the nonstationarity among blocks are thought to be balanced.

In practice, covariances $R_s(t)$ and $R_y(t)$ possibly have off-diagonal elements. We are planning to investigate the errors by relaxing the assumptions of this paper.

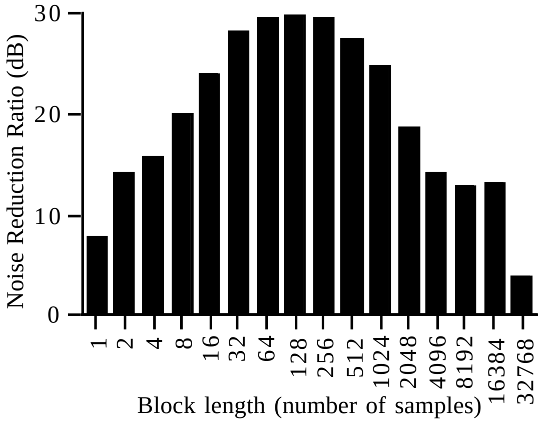

Fig. 4 Results of experiments.

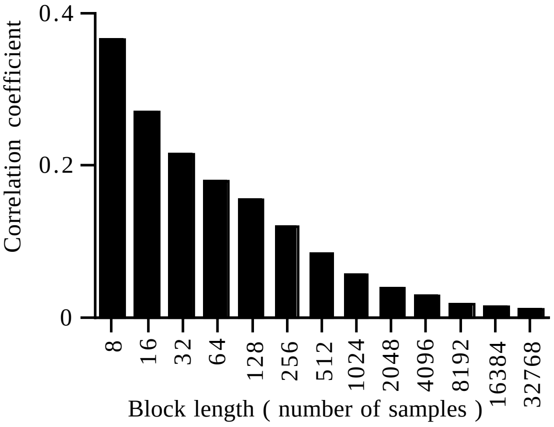

Fig. 5 Correlation Coefficients of source signals.

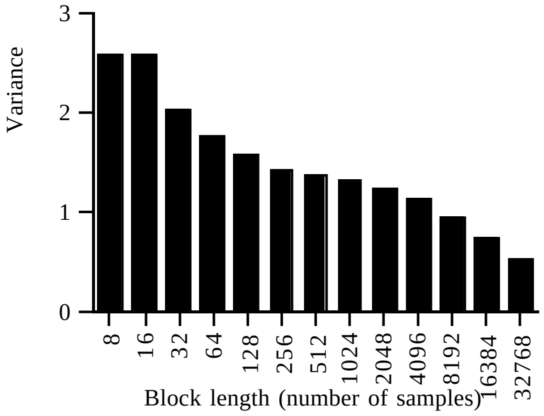

Fig. 6 Variance of power ratio among source signals in adjacent blocks.

6. REFERENCES

[1] T-W Lee: "Independent Component Analysis," Kluwer, 1998

[2] K. Matsuoka, M. Ohya and M. Kawamoto: "A Neural Net for Blind Separation of Nonstationary Signals," Neural Networks, Vol. 8, No. 3, pp. 411-419, 1995.

[3] A. Ando and M Iwaki: "Blind Separation of Nonstationary Sources by Block Decorrelation of Output Signal," Technical Report of IEICE, EA2002-57, 2002 (in Japanese)

[4] J.L. Flanagan: "Speech Analysis Synthesis and Perception, 2nd Edition," Springer, 1972.