Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Speech separation based on law of causality

Kensaku Fujii[a], Hirofumi Nakano[a] and Mitsuji Muneyasu[b]

[a]University of Hyogo, 2167 Shosha, 671-2280 Himeji, Japan
[b]Kansai University, 3-3-35 Yamate-cho, 564-8680 Suita, Japan
fujiken@eng.u-hyogo.ac.jp

This paper proposes a microphone system separating speech signals, based on a different principle from independent component analysis (ICA). This system applies linear prediction error filters to microphone outputs, and using the prediction errors, adjusts the coefficients of adaptive filters. In this case, only the prediction errors satisfying the law of causality become available for the adjustment; consequently, this system can steer a null toward a direction satisfying it. The permutation problem discussed in ICA can be thereby avoided. This system also can compensate the separated speech signals by using adaptive filters, and can provide high quality speech signals. This paper finally verified the performance of the proposed system by using the speech signal data measured in an ordinary room. This result shows that the proposed system works well even in reverberation environment.

1 Introduction

A target speech signal is generally detected together with annoying noises; therefore various methods reducing the noises to enhance the speech signal have been studied and proposed. Actually, the various methods can be classified into two types, according to the number of microphones used in the systems. One is a type applied to single microphone systems, which is represented by the spectral subtraction method [1]. Another is a type used for microphone array systems [2]. They each have suitable application fields.

The microphone array systems can be moreover classified into two types. One is the delay-and-sum array, which reduces noises by steering a main lobe toward the incident direction of the target speech signal [2]. Another is the adaptive beam former [3] eliminating a noise by steering a null. They both, however, require the detection of the incident direction of the speech signal [4,5].

We have hence proposed a latter type of method not requiring the detection of the incident direction [6]. The method uses the assumption that the noise is stationary in comparison with the speech signal. Ingeniously using this assumption, the method can adaptively steer a null toward the noise source under the incidence of the speech signal.

On the other hand, the assumption states that the method does not work well when the noise is unstationary, such as speech signal. Blind source separation using independent component analysis (ICA) can successfully enhance the speech signal deteriorated by such unstationary noise. Most of the latest studies on the speech enhancement are hence based on ICA; on the contrary, the conventional adaptive beam former seems to have become past topics.

In this paper, we propose a new microphone system using the law of causality for separating two speech signals. The proposed system is characterized by applying linear prediction error filters to two microphone outputs. Using the prediction errors, the proposed system adjusts the coefficients of adaptive filters. The proposed system can then steer a null only toward the speech source satisfying the law of causality. Another speech signal works as a disturbance. This means that the permutation problem discussed in ICA is avoidable in the proposed system.

This paper finally verifies the performance of the proposed system by computer simulations using impulse responses measured placing at an interval of 100 mm on a table in an experimental room and two speech signal sets: male and female (Japanese), Identical male (English). The results demonstrate that the proposed system can separate two speech signals with the difference of more than 25 dB. The proposed system can be also applied to the noise reduction enhancing a speech signal. In this paper, the noise reduction effect obtained by the proposed system is moreover shown.

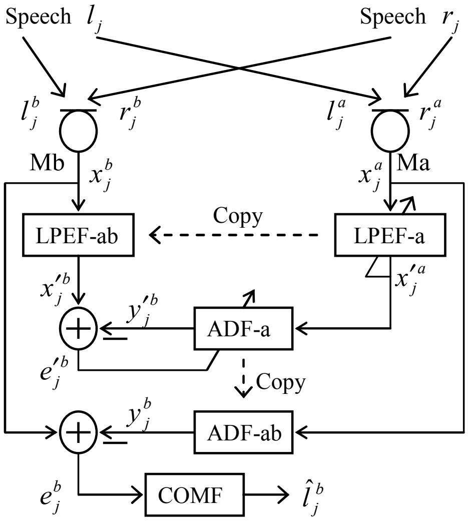


Fig.1 Configuration of proposed system extracting speech signal arriving from left hand direction.

2 Basic Configuration

Figure 1 shows a basic configuration of the proposed system extracting speech signal l_j^b detected by microphone Mb, where j denotes sample time index. In this configuration, two speech signals l_j and r_j are incident from the left and the right hand directions, which are detected as l_j^a , r_j^a , l_j^b , and r_j^b by microphones Ma and Mb, respectively.

This system works so as to cancel the speech signal, r_j^b , by subtracting the output of adaptive filter ADF-ab, y_j^b, from the microphone output,

$$x_j^b = l_j^b + r_j^b . \tag{1}$$

By the subtraction, the target speech signal, l_j^b , is extracted as the estimation error, e_j^b . This subtraction, however, distorts the speech signal, l_j^b . COMF is a filter used for compensating the distortion. Using the compensation filter, COMF, the proposed system provides the speech signal, $\hat{l}_j^b$, more approximate to l_j^b .

The other components shown in Fig. 1: adaptive filter ADF-a, linear prediction error filters LPEF-a and LPEF-ab, are used for estimating the coefficients of ADF-ab. Their functions are detailed in Chapter 4.

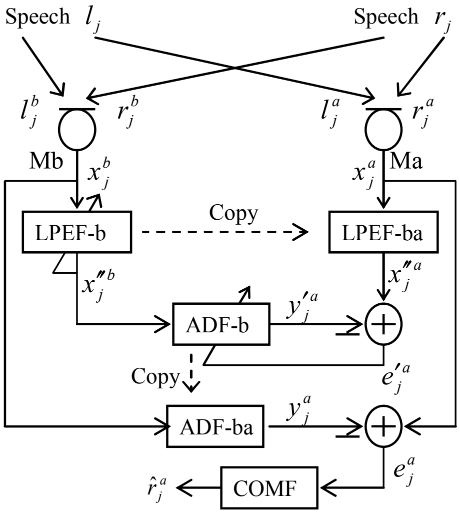


Fig. 2 Configuration of proposed system extracting speech signal arriving from right hand direction.

Figure 2 shows another configuration of the proposed system, which is used for extracting another speech signal, r_j^a, arriving from the right hand direction. As seen from the configuration, this system subtracts the output of adaptive filter ADF-ba, y_j^a, from the microphone output,

$$x_j^a = l_j^a + r_j^a, \tag{2}$$

and thereby extracts the speech signal, r_j^a, as the estimation error, e_j^a. The error, e_j^a, is similarly compensated by COMF; consequently, the speech signal more approximate to r_j^a is provided as $\hat{r}_j^a$. The functions of the other components are detailed in Chapter 4

3 Extraction Matrix

The extraction operations mentioned above can be expressed by the following equation,

$$\begin{bmatrix} \hat{r}^a(z) \\ \hat{l}^b(z) \end{bmatrix} = c(z) \begin{bmatrix} 1 & -H^{ba}(z) \\ -H^{ab}(z) & 1 \end{bmatrix} \begin{bmatrix} x^a(z) \\ x^b(z) \end{bmatrix}, \tag{3}$$

where $\hat{r}^a(z)$, $\hat{l}^b(z)$, $x^a(z)$ and $x^b(z)$ are z-transforms of the outputs of the compensation filters and the microphones, and $c(z)$, $H^{ba}(z)$ and $H^{ab}(z)$ are the transfer functions of filters COMF, ADF-ba and ADF-ab, respectively. Here, it should be noted that the purpose of the proposed system is to extract the microphone outputs,

$$r^a(z) = h^{ra}(z)r(z) \tag{4}$$

and

$$l^b(z) = h^{lb}(z)l(z), \tag{5}$$

by canceling

$$r^b(z) = h^{rb}(z)r(z) \tag{6}$$

and

$$l^a(z) = h^{la}(z)l(z), \tag{7}$$

not original speech signals $r(z)$ and $l(z)$, where $h^{ra}(z)$, $h^{rb}(z)$, $h^{la}(z)$ and $h^{lb}(z)$ are the transfer functions of the acoustic paths from the speech sources to the microphones. We hence define the transfer functions of the acoustic paths between Microphones Ma and Mb as

$$h^{ab}(z) = h^{rb}(z) / h^{ra}(z) \tag{8}$$

and

$$h^{ba}(z) = h^{la}(z) / h^{lb}(z). \tag{9}$$

By using the definition, the microphone outputs are rewritten as

$$\begin{bmatrix} x^a(z) \\ x^b(z) \end{bmatrix} = \begin{bmatrix} 1 & h^{ba}(z) \\ h^{ab}(z) & 1 \end{bmatrix} \begin{bmatrix} r^a(z) \\ l^b(z) \end{bmatrix}. \tag{10}$$

From (10) and (3), we can see that the speech signals, $r^a(z)$ and $l^b(z)$, can be extracted when the relation,

$$c(z) \begin{bmatrix} 1 & -H^{ba}(z) \\ -H^{ab}(z) & 1 \end{bmatrix} \begin{bmatrix} 1 & h^{ba}(z) \\ h^{ab}(z) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{11}$$

is satisfied, and moreover, it can be obtained when we give the coefficients satisfying the relations,

$$H^{ab}(z) = h^{ab}(z), \tag{12}$$

$$H^{ba}(z) = h^{ba}(z), \tag{13}$$

and

$$c(z) = 1/1 - H^{ab}(z)H^{ba}(z), \tag{14}$$

to the adaptive filters and the compensation filter, respectively.

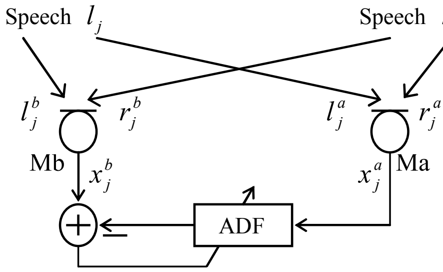


Fig. 3 Configuration of conventional beam former.

4 Estimation of Acoustic Paths

The proposed system is characterized by the insertion of the linear prediction error filters shown in Fig. 1 and 2. In this chapter, we explain the principle that the insertion enables the estimation of the acoustic paths, $h^{ab}(z)$ and $h^{ba}(z)$ even when the two speech signals are simultaneously incident to the proposed system.

Figure 3 shows a configuration of conventional beam formers. In this configuration, adaptive filter ADF tries to predict l_j^b preceding l_j^a when r_j^a and r_j^b are absent. This try, however, is useless when l_j has no auto-correlation. On the other hand, the system identification using r_j^a and r_j^b is valid even when the auto-correlation of r_j is low. This duality enables the estimation of the acoustic paths.

The proposed system inserts the linear prediction error filters, LPEF-a and LPEF-a, shown in Fig. 1 and 2, so as to use this duality. By the linear prediction error filters, the microphone outputs, x_j^a and x_j^b, are changed to low auto-correlation signals,

$$x_j'^a = l_j'^a + r_j'^a \tag{15}$$

and

$$x_j'^b = l_j'^b + r_j'^b, \tag{16}$$

where $l_j'^a$, $r_j'^a$, $l_j'^b$ and $r_j'^b$ are linear prediction errors. As mentioned above, adaptive filter ADF-a then cannot form the linear prediction circuit using $l_j'^a$ preceding $l_j'^b$. The prediction errors, $l_j'^a$ and $l_j'^b$, act on ADF-a only as disturbances.

On the other hand, r_j^a preceding r_j^b satisfies the law of causality. In this case, the acoustic path from microphones Ma to Mb,

$$\boldsymbol{h}^{ab} = \begin{bmatrix} h^{ab}(0) & h^{ab}(1) & \cdots & h^{ab}(l-1) \end{bmatrix} \tag{17}$$

can be identified by the adaptive filter, ADF-a. The proposed system copies the coefficient vector of ADF-a, $\boldsymbol{H}_j^{ab}$, to ADF-ab, and yields the estimation error,

$$\begin{aligned} e_j^b &= l_j^b - \boldsymbol{H}_j^{ab^T} \boldsymbol{l}_j^a + (\boldsymbol{h}^{ab^T} - \boldsymbol{H}_j^{ab^T})\boldsymbol{r}_j^a \\ &\approx l_j^b - \boldsymbol{H}_j^{ab^T} \boldsymbol{l}_j^a \end{aligned} \tag{18}$$

when $\boldsymbol{H}_j^{ab} \approx \boldsymbol{h}^{ab}$ is satisfied by the identification operation. This relation also can be expressed as

$$e^b(z) \approx l^b(z) - H^{ab}(z) l^a(z) \tag{19}$$

by using z-transform expression, which is rewritten as

$$\begin{aligned} e^b(z) &\approx l^b(z) - H^{ab}(z) h^{ba}(z) l^b(z) \\ &= [1 - H^{ab}(z) h^{ba}(z)] l^b(z) \end{aligned}. \tag{20}$$

Moreover, applying the compensation filter, COMF, whose transfer function is $c(z)$, to the estimation error, $e^b(z)$, gives

$$l^b(z) = e^b(z)c(z) \approx \frac{1 - H^{ab}(z)h^{ba}(z)}{1 - H^{ab}(z)H^{ba}(z)} l^b(z) \approx l^b(z). \tag{21}$$

Similarly, adaptive filter ADF-b shown in Fig. 2 can estimate the acoustic path from microphones Mb to Ma,

$$\boldsymbol{h}^{ba} = [h^{ba}(0) \quad h^{ba}(1) \quad \cdots \quad h^{ba}(l-1)]^T \tag{22}$$

by using the linear prediction errors, $x_j''^a$ and $x_j''^b$. The coefficient vector of ADF-b, $\boldsymbol{H}_j^{ba}$, obtained by the estimation, is copied to ADF-ab, which provides the estimation error,

$$e_j^a \approx r_j^a - \boldsymbol{H}_j^{ba^T} \boldsymbol{r}_j^b. \tag{23}$$

This estimation error is moreover compensated by COMF; thereby, the proposed system can provide

$$\hat{r}^b(z) = e^a(z)c(z) = \frac{1 - H^{ba}(z)h^{ab}(z)}{1 - H^{ab}(z)H^{ba}(z)} r^a(z) \approx r^a(z). \tag{24}$$

The proposed system thus separates the speech signals.

5 Adaptive Operation

As seen from Figs. 1 and 2, since the configurations are symmetric, the same adaptive operation is applied to the estimation of the coefficient vectors, $\boldsymbol{H}_j^{ab}$ and $\boldsymbol{H}_j^{ba}$. We hence explain the adaptive operation, referring only the configuration shown in Fig. 1.

In this system, since the power of speech signals extremely change, we use the following block implementation algorithm,

$$\boldsymbol{H}_{n+1}^{ab} = \boldsymbol{H}_n^{ab} + \mu_n^{ab} \frac{\sum_{j=nJ+1}^{(n+1)J} e_j'^b \boldsymbol{x}_j'^a}{\sum_{j=nJ+1}^{(n+1)J} \boldsymbol{x}_j'^{a^T} \boldsymbol{x}_j'^a}, \tag{25}$$

to which applying the block length and the step size controls, where n is block number, μ_n^{ab} is a variable called step size, and J denotes block length [7]. In this algorithm, the estimation error, $e_j'^b$, is controlled by the step size, the block length, and the powers of $r_j'^a$ and $l_j'^b$.

A problem is that the power of $r_j'^a$ extremely changes. To absorb the change, the adaptive operation extends the block length J until the estimation error power,

$$P_n^a = \sum_{j=nJ+1}^{(n+1)J} \boldsymbol{e}_j'^{a^T} \boldsymbol{e}_j'^a, \tag{26}$$

satisfies the relation,

$$P_n^a > P_0^a \times I, \tag{27}$$

and then updates the coefficient vector, $\boldsymbol{H}_n^{ab}$, where P_0^a is the average power of $r_j'^a$, and I is the number of taps of adaptive filter ADF-a.

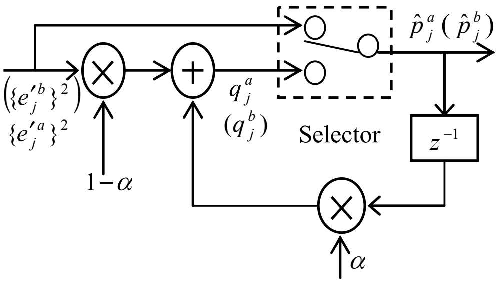


Fig. 4 Power detection filter.

Another problem is that the power of $l_j'^b$ disturbing the estimation of $\boldsymbol{H}_n^{ab}$ also changes. To prevent this disturbance, we apply the following variable step size [7],

$$\mu_n^{ab} = \frac{2p_n'^a C_0}{p_n'^b + p_n'^a C_0}, \tag{28}$$

to (25), where C_0 is a target level of the estimation error, $p_n'^a$ and $p_n'^b$ are the powers of $r_j'^a$ and $l_j'^b$ measured at nth block, respectively.

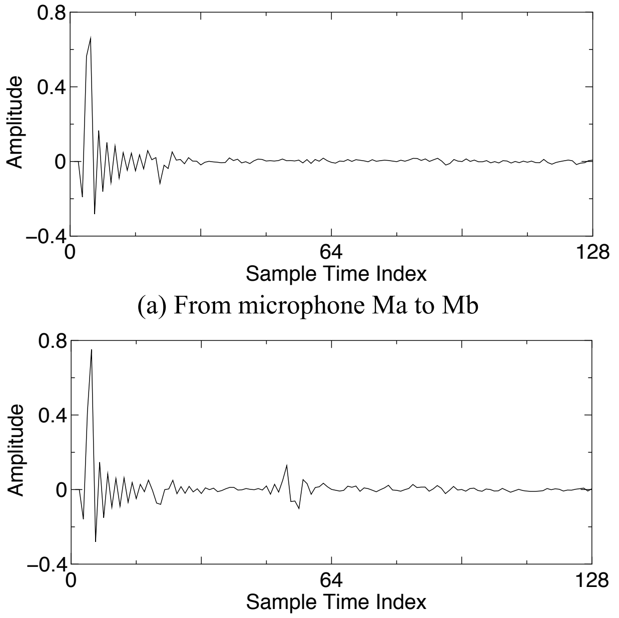
In this system, since the correct estimation of the powers is difficult, we approximate the powers by using two filters of the configuration shown in Fig. 4, where α is a constant smaller than unity. The powers are estimated to be $\hat{p}_j^a$ and $\hat{p}_j^b$ obtained by feeding $\{e_j'^a\}^2$ and $\{e_j'^a\}^2$ into the filters, respectively. The selector introduced into the filter is inserted for quickly tracking the increase of the disturbance, $l_j'^b$ included in $e_j'^b$, and the decrease of the reference signal, $r_j'^a$, involved in $e_j'^a$.

In the estimation of $\hat{p}_j^b$, the selector compares $\{e_j'^b\}^2$ with

$$q_j^b = (1-\alpha)\{e_j'^b\}^2 + \alpha\hat{p}_{j-1}^b \tag{29}$$

and then selects the larger one of them. The filter can thereby track the quick increase of the disturbance power.

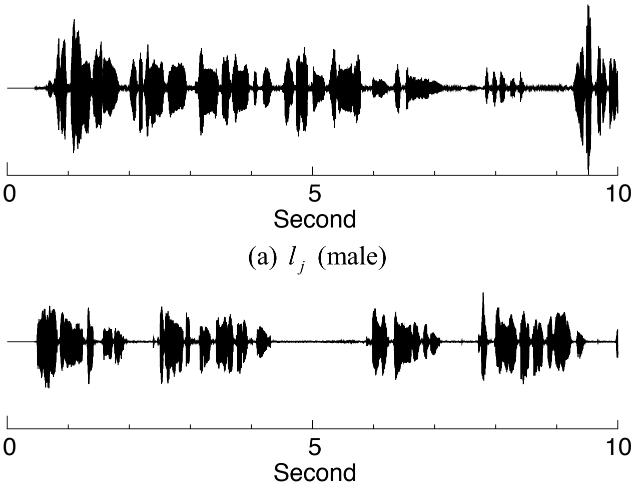
Similarly, the selector inserted in the filter estimating $\hat{p}_j^a$ compares $(e_j'^a)^2$ with the average power of $r_j'^a$ involved in $e_j'^a$, that is P_0^a, and then provides the smaller one of them as the filter output. The proposed system can steadily separate the speech signals by using the step size calculated substituting $\hat{p}_j^a$ and $\hat{p}_j^b$ for p_n^a and p_n^b in (28).



(a) From microphone Ma to Mb

(b) From microphone Mb to Ma

Fig. 5 Impulse response used for computer simulations.



(a) l_j (male)

(b) r_j (female)

Fig. 6 Speech signals used for computer simulations.

6 Computer Simulation

Finally, we verify the performance of the proposed system by computer simulations. Figure 5 shows the impulse responses between two Microphones, Ma and Mb, which are measured placing at an interval of 100 mm on a table (1.8 m long, 1.8 m wide and 0.7 m high) in an experimental room (8.4 m long, 6.0 m wide and 2.7 m high). In addition, loudspeakers are placed 500 mm apart from the midpoint of the microphones and at angles of 45 and 135 degrees.

Figure 6 shows the speech signals (Japanese) used for computer simulations, where the sampling frequency is 8 kHz, the word length is 16 bit. Using the speech signals and the impulse responses, we verify the performance of the proposed system under the following conditions:

(1) The number of taps of adaptive filters (ADF-a, ADF-ab, ADF-b and ADF-ba) is 128.
(2) The maximum step size is limited to 1.0.
(3) The maximum block length is 16.
(4) The average power of $r_j'^a$ is $p_0^a = 0.0005329$.
(5) The average power of $l_j'^b$ is $p_0^b = 0.001$.
(6) The calculation of (25) is skipped when the block length is more than 16.
(7) The number of taps of linear prediction error filters (LPEF-a, LPEF-ab, LPEF-b, LPEF-ba) is 128.
(8) The adaptive algorithm applied to the linear prediction error filters is the normalized least mean square (NLMS) algorithm whose step size is 0.01.
(9) $C_0 = 0.01$, $\alpha = 1 - 1/256$.
(10) The number of taps of COMF is 512.
(11) The coefficients of COMF are estimated using the equation error method shown in Fig. 7, where $N = 64$.

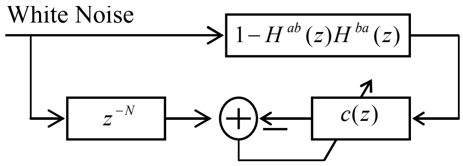
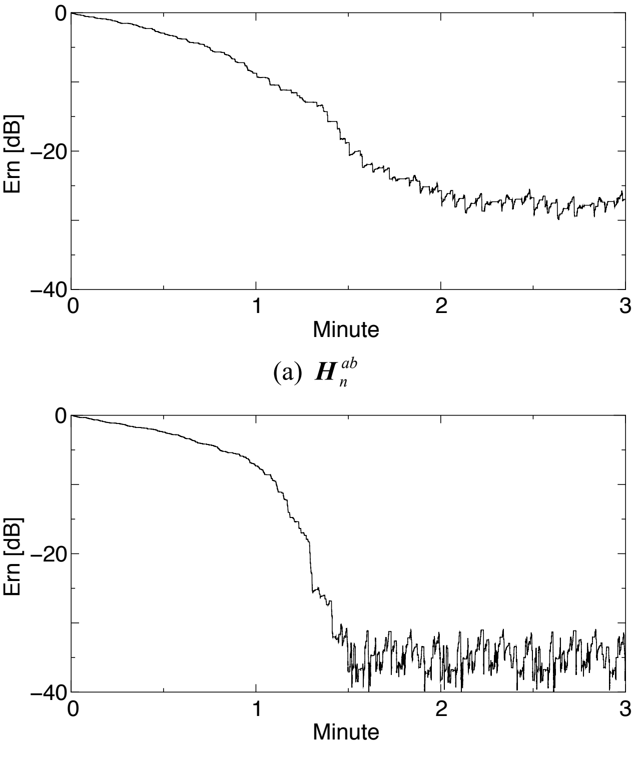


Fig. 7 Estimation system for COMF.



(a) $\boldsymbol{H}_n^{ab}$

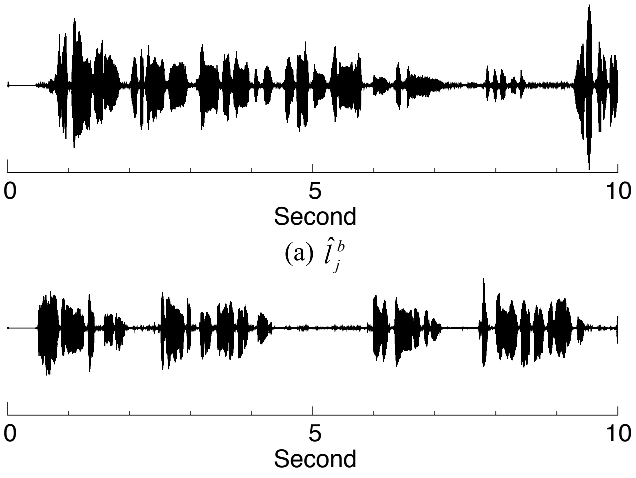
(b) $\boldsymbol{H}_n^{ba}$

Fig. 8 Convergence properties of coefficient vector

Figure 8 shows the convergence properties of $\boldsymbol{H}_n^{ba}$ and $\boldsymbol{H}_n^{ab}$ estimated under the above conditions, where the estimation errors are calculated using

$$Er_n = 10\log_{10}\frac{\sum_{i=1}^{I}\{H_n^{ab}(i) - h^{ab}(i)\}^2}{\sum_{i=1}^{I}\{h^{ab}(i)\}^2} \qquad (30)$$

The results show that the proposed system can reduce the estimation error to about -30 dB. In addition, Fig. 9 shows the speech signals separated by the proposed system. In the voiceless terms, we can see residual speech signals slightly.



(a) $\hat{l}_j^b$

(b) $\hat{r}_j^a$

Fig. 9 Compensated speech signals.

Here, to verify the performance of the proposed system numerically, we calculate the speech quality and the separation efficiency defined by

$$SQ = 10\log_{10}\frac{\sum_{j=0}^{K-1}\{r_{j-N}^a\}^2}{\sum_{j=0}^{K-1}\{r_{j-N}^a - \hat{r}_j^a\}^2} \qquad (31)$$

and

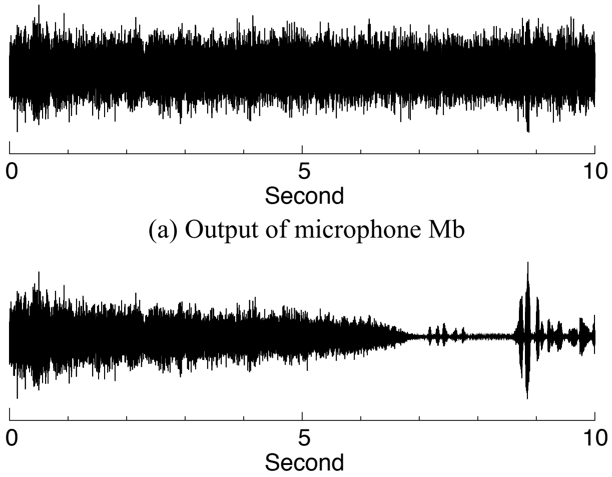
$$SN = 10\log_{10}\frac{\sum_{j=0}^{K-1}\{\hat{r}_j^a\}^2}{\sum_{j=0}^{K-1}\{\hat{l}_j^a\}^2}, \qquad (32)$$

respectively, which are summarized in Table 1, where K is the number of samples of the speech signals: $K = 200{,}000$. In addition, the performances obtained using the speech signals of another person (English) are shown in the same table. The results show that the proposed system can successfully extract speech signals with high fidelity even if they are identical person's speech signals.

	Male & Female (Japanese)		Identical Male (English)	
	$\hat{l}_j^b$	$\hat{r}_j^a$	$\hat{l}_j^b$	$\hat{r}_j^a$
SQ(dB)	19.9	23.5	21.4	21.2
SN(dB)	28.5	25.3	26.5	26.3

Table 1 Performance of proposed system

The proposed system can be also applied to the field of noise reduction. This paper hence presents the noise reduction effect obtained by the proposed system. Figure 10 shows a microphone output buried in a jet fan noise and a speech signal enhanced by the proposed system. In this simulation, a jet fan noise is substituted to for speech signal r_j. As seen from the simulation results, speech signal l_j^b completely buried in the noise is rapidly and successfully enhanced.



(a) Output of microphone Mb

(b) Enhanced speech signal

Fig. 10 Noise reduction effect obtained by the proposed system.

7 Conclusion

In this paper, we have proposed a new speech separation system. This system uses the law of causality for the separation; accordingly, the permutation problem discussed in ICA is avoidable. We moreover verify the performance of the proposed system by the computer simulations and show that the proposed system can successfully extracts the speech signal with high fidelity.

In the near future, we will further study on the step size control method to improve the performance of the proposed system. Moreover, we will verify the performance placing the experimental system in practical rooms.

References

[1] S. F. Boll, "Suppression of acoustic noise in speech spectral subtraction," *IEEE Trans. Acoust., Speech & Signal Process.*, vol. ASSP-27, pp. 113-120 (1979).

[2] J. L. Flanagan, J. D. Johnston, R. Zahn and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, pp. 1508-1518 (1985).

[3] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust., Speech & Signal Process.*, vol. ASSP-34, pp. 1391-1400 (1986).

[4] P. M. Zurek and J. E. Greenberg, "Sensitivity to design parameters in an adaptive-beamforming," In *Proc. ICASSP1990*, pp. 1129-1132.

[5] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926-935 (1972).

[6] K. Fujii, T. Amitani, S. Miyata, N. Sasaoka and Y. Itoh, "Two-microphone system using linear prediction and noise reconstruction," *Acoust. Sci. & Tech.*, vol. 28, pp. 115-123 (2007).

[7] K. Fujii and J. Ohga, "A method to keep the estimation error at a desired level for acoustic echo canceller systems," *IEICE Trans. Fundamentals*, vol. J83-A, pp. 141-151 (2000).