SPEECH ENHANCEMENT USING NONLINEAR MICROPHONE ARRAY WITH COMPLEMENTARY BEAMFORMING

Hiroshi SARUWATARI †, *Shoji KAJITA*‡, *Kazuya TAKEDA*† *and Fumitada ITAKURA*‡

†Graduate School of Engineering, ‡Center for Information Media Studies, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, JAPAN
E-mail: sawatari@itakura.nuee.nagoya-u.ac.jp

ABSTRACT

This paper describes an improved spectral subtraction method by using the complementary beamforming microphone array to enhance noisy speech signals for speech recognition. The complementary beamforming is based on two types of beamformers designed to obtain complementary directivity patterns with respect to each other. In this paper, it is shown that the nonlinear subtraction processing with complementary beamforming can result in a kind of the spectral subtraction without the need for speech pause detection. In addition, the design of the optimization algorithm for the directivity pattern is also described. To evaluate the effectiveness, speech enhancement experiments and speech recognition experiments are performed based on computer simulations. In comparison with the optimized conventional delay-and-sum array, it is shown that the proposed array improves the signal-to-noise ratio of degraded speech by about 2 dB and performs about 10% better in word recognition rates under heavy noisy conditions.

1. INTRODUCTION

Speech enhancement in noisy environments is a typical and important approach to construct a robust man-machine interface, such as a speech recognition system used in the real world. Among various noise reduction methods, a microphone array is one of the most effective techniques. The Delay-and-Sum (DS) array[1] and the adaptive array[2] are the conventional and popular microphone arrays used for noise reduction. However, they must use a large number of microphones or much computational costs to achieve high performance, especially in low frequency regions. To achieve further improvement, several microphone arrays combined with nonlinear speech processing, such as the Spectral Subtraction (SS) method[3], have been proposed in the recent works[4, 5, 6]. In these methods, however, there exists other problems in terms of degradations of speech quality due to the speech pause detection error or the misestimation of noise directions.

In this paper, a new microphone array system based on nonlinear array signal processing is proposed. In this system, both complementary beamforming[7] and nonlinear subtraction processing are used to construct the SS without any speech pause detection and the estimation of noise directions. Using these techniques, a lower sidelobes can be achieved compared with the conventional DS array.

This paper is constructed as follows. In the following section, the nonlinear microphone array and its optimization algorithm for directivity patterns are described. In Sections 3 and 4, the directivity pattern design and some experiments based on computer simulations are performed. After discussions on the results of the experiments, we conclude this paper in Section 5.

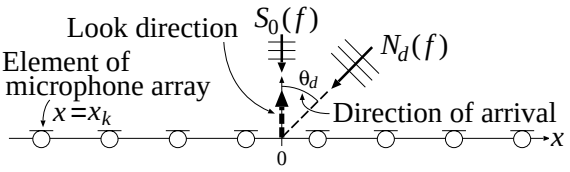

Figure 1: Configuration example of a microphone array and acoustic signals.

2. ALGORITHM

2.1. Nonlinear Microphone Array with Complementary Beamforming

In this study, a straight-line array is assumed. The coordinates of the elements are designated as x_k ($k = 1, \cdots, K$), and the directions of arrival of multiple signals are designated as θ_d ($d = 1, \cdots, D$) (see Fig. 1). Also, the look direction is set to be normal to the array ($\theta = 0$).

First, using two types of complementary weight vectors[7] of element $\boldsymbol{g} = [g_1, \cdots, g_K]$ and $\boldsymbol{h} = [h_1, \cdots, h_K]$, we construct the signal spectra $S^{(g)}(f)$ and $S^{(h)}(f)$. Here, "*complementary*" implies one of the following conditions: "directivity pattern gain $|\boldsymbol{g}\boldsymbol{a}_d(f)| \gg$ directivity pattern gain $|\boldsymbol{h}\boldsymbol{a}_d(f)|$" or "directivity pattern gain $|\boldsymbol{g}\boldsymbol{a}_d(f)| \ll$ directivity pattern gain $|\boldsymbol{h}\boldsymbol{a}_d(f)|$" for an arbitrary direction d (see, e.g., Fig. 3 (bottom)). The exception is that the gain of both directivity patterns is unity with respect to the look direction. Here, $\boldsymbol{a}_d(f)$ is a steering vector and is defined as follows

$$\boldsymbol{a}_d(f) \equiv [a_{1,d}(f), \cdots, a_{k,d}(f), \cdots, a_{K,d}(f)]^{\mathrm{T}}, \quad (1)$$

$$a_{k,d}(f) \equiv \exp[j2\pi f \cdot x_k \cdot \sin(\theta_d)/c], \quad (2)$$

where c is the velocity of sound. As for the signals $S^{(g)}(f)$ and $S^{(h)}(f)$, the following equations are applicable for the target speech signal arriving from the look direction, $S_0(f)$, and noise signals arriving from other directions, $N_d(f)$.

$$S^{(g)}(f) = S_0(f) + \sum_{d=1}^{D} \boldsymbol{g}\boldsymbol{a}_d(f) \cdot N_d(f) \quad (3)$$

$$S^{(h)}(f) = S_0(f) + \sum_{d=1}^{D} \boldsymbol{h}\boldsymbol{a}_d(f) \cdot N_d(f) \quad (4)$$

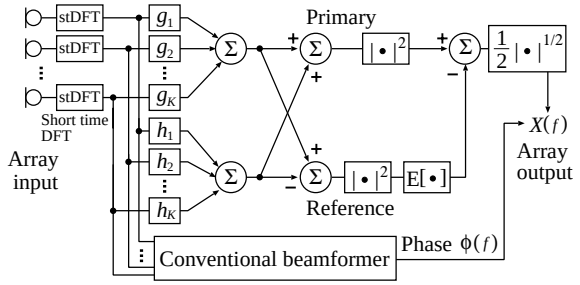

Figure 2: Block diagram of nonlinear microphone array with complementary beamforming.

The sum of Eqs. (3) and (4) is designated as the primary signal, $S^{(\mathrm{p})}(f)$, and the difference is designated as the reference signal, $S^{(\mathrm{r})}(f)$. These can be given as

$$S^{(\mathrm{p})}(f) = 2S_0(f) + \sum_{d=1}^{D} \{\boldsymbol{g}\boldsymbol{a}_d(f) + \boldsymbol{h}\boldsymbol{a}_d(f)\} \cdot N_d(f), \quad (5)$$

$$S^{(\mathrm{r})}(f) = \sum_{d=1}^{D} \{\boldsymbol{g}\boldsymbol{a}_d(f) - \boldsymbol{h}\boldsymbol{a}_d(f)\} \cdot N_d(f). \quad (6)$$

If the directivity patterns $|\boldsymbol{g}\boldsymbol{a}_d(f)|$ and $|\boldsymbol{h}\boldsymbol{a}_d(f)|$ are designed to be complementary, and if there are no correlations between arriving signals, the expectation value of the power spectrum of the noise component in the primary signal (the second term in Eq. (5)) can be approximated using the reference signal, i.e., it can be given as

$$\mathrm{E}\left[\left|\sum_{d=1}^{D} \{\boldsymbol{g}\boldsymbol{a}_d(f) + \boldsymbol{h}\boldsymbol{a}_d(f)\} \cdot N_d(f)\right|^2\right] \approx \mathrm{E}\left[\left|\sum_{d=1}^{D} \{\boldsymbol{g}\boldsymbol{a}_d(f) - \boldsymbol{h}\boldsymbol{a}_d(f)\} \cdot N_d(f)\right|^2\right]. \quad (7)$$

Using the primary and reference signals, without any speech pause detection, we can construct the SS method[3] as follows.

$$X(f) \equiv \frac{1}{2} \cdot \left||S^{(\mathrm{p})}(f)|^2 - \mathrm{E}\left[|S^{(\mathrm{r})}(f)|^2\right]\right|^{1/2} \cdot e^{j\phi(f)} \quad (8)$$

Here, $X(f)$ represents the complex spectrum of the speech signal recovered by the SS method. Also, $\phi(f)$ is an appropriate phase function; for example, the phase function obtained by a conventional DS beamformer is used. Figure 2 shows a block diagram of this array system.

2.2. Optimization of Directivity Patterns

To optimize the processing of Eq. (8), we design directivity patterns so that the noise component in the expectation value of the power spectrum in the primary signal decreases. Here, using the $\mathrm{E}[|S^{(\mathrm{p})}(f)|^2]$ instead of the $|S^{(\mathrm{p})}(f)|^2$ in Eq. (8), the estimated power spectrum of Eq. (8) is given as

$$\begin{aligned} |\hat{X}(f)|^2 &= (1/4) \cdot \left|\mathrm{E}\left[|S^{(\mathrm{p})}(f)|^2\right] - \mathrm{E}\left[|S^{(\mathrm{r})}(f)|^2\right]\right| \\ &= \left|\mathrm{E}\left[|S_0(f)|^2\right] + \sum_{d=1}^{D} \mathrm{Re}\left[\boldsymbol{g}\boldsymbol{a}_d(f) \cdot (\boldsymbol{h}\boldsymbol{a}_d(f))^*\right] \cdot \mathrm{E}\left[|N_d(f)|^2\right]\right| \\ &\le \mathrm{E}\left[|S_0(f)|^2\right] + \sum_{d=1}^{D} |\boldsymbol{g}\boldsymbol{a}_d(f) \cdot \boldsymbol{h}\boldsymbol{a}_d(f)| \cdot \mathrm{E}\left[|N_d(f)|^2\right]. \end{aligned} \quad (9)$$

Eq (9) indicates that the gain for the target signal is one, and the gain for the noise is $|\boldsymbol{g}\boldsymbol{a}_d(f) \cdot \boldsymbol{h}\boldsymbol{a}_d(f)|$. Accordingly, to reduce the noise component in Eq. (9), it is not necessary to produce sidelobes which have small $|\boldsymbol{g}\boldsymbol{a}_d(f)|$ and $|\boldsymbol{h}\boldsymbol{a}_d(f)|$ individually, but to design them so as to obtain a small $|\boldsymbol{g}\boldsymbol{a}_d(f) \cdot \boldsymbol{h}\boldsymbol{a}_d(f)|$ in the directivity patterns.

The optimization method of directivity patterns is explained as follows. First, let us define the following vector as a function which represents the product of directivity pattern values.

$$\boldsymbol{f}(\boldsymbol{g}, \boldsymbol{h}) \equiv [\boldsymbol{g}\boldsymbol{a}_1(f)\boldsymbol{h}\boldsymbol{a}_1(f), \cdots, \boldsymbol{g}\boldsymbol{a}_D(f)\boldsymbol{h}\boldsymbol{a}_D(f)]^{\mathrm{T}} \quad (10)$$

Next, we define a vector with the desired directivity pattern in each direction, $\boldsymbol{q} \equiv [q_1, \cdots, q_d, \cdots, q_D]^{\mathrm{T}}$. In this study, the value of q_{d_0} which corresponds to the look direction, θ_d=0 ($\equiv\theta_{d_0}$), is set to be large, and the other values are set to be small. Using these vectors, the second term on the right hand side of Eq. (9) is optimized using the criterion of the weighted square norm minimum. More practically, the constrained least squares problem shown in Eqs. (11) and (12) is solved.

$$\min_{\boldsymbol{g},\boldsymbol{h}} \quad \|\boldsymbol{W}_{\mathrm{d}} \cdot \boldsymbol{q} - \boldsymbol{W}_{\mathrm{d}} \cdot \boldsymbol{f}(\boldsymbol{g}, \boldsymbol{h})\|_2 \quad (11)$$

$$\text{subject to} \quad \boldsymbol{g}\boldsymbol{a}_{d_0}(f) = \boldsymbol{h}\boldsymbol{a}_{d_0}(f) = (q_{d_0})^{1/2} \quad (12)$$
$$\left(\boldsymbol{a}_{d_0}(f) = [1, \cdots, 1]^{\mathrm{T}}\right)$$

Here, $\boldsymbol{W}_{\mathrm{d}}$ represents the following weighting matrix for each direction.

$$\boldsymbol{W}_{\mathrm{d}} \equiv \mathrm{diag}(w_1, \cdots, w_d, \cdots, w_D) \quad (13)$$

Since Eq. (11) is a problem of nonlinear minimization, it can be minimized using the iterative method; for example, the Gauss-Newton method is used on this work.

3. DIRECTIVITY PATTERN DESIGN

3.1. Design Condition

An eight-element array with an interelement spacing of 5 cm is assumed in the design and the weight vectors are calculated using Eq. (11) for each frequency independently.

As common design conditions for each frequency, the desired directivity pattern, q_d, is obtained by setting the value of 1 for the look direction (q_{d_0}=1) and 0 for other directions. As for the weighting matrix $\boldsymbol{W}_{\mathrm{d}}$, w_d for from -90 to -14° and from 14 to 90° are set to be 100 and those for other directions are set to be 1. For the purpose of stabilization, the step size parameter for iterations in the Gauss-Newton method is set to be 0.1.

3.2. Initial Condition and Design Example

In the above approach, appropriate initial values must be selected before the iterative designing. In this study, we design two weight vectors which have different directivity patterns for each frequency, and the iteration is started using these initial values. As examples

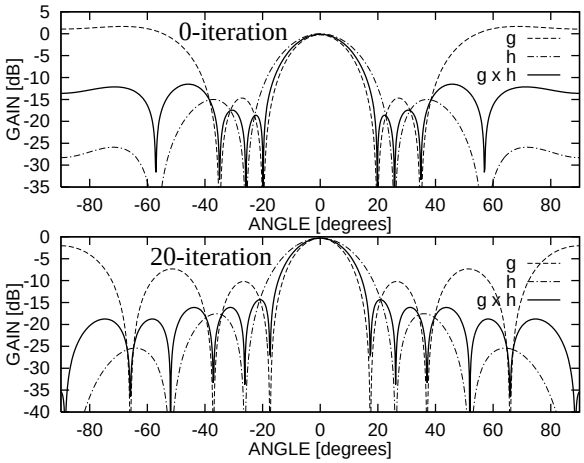

Figure 3: Directivity patterns at initial condition (top), and at 20th iteration (bottom).

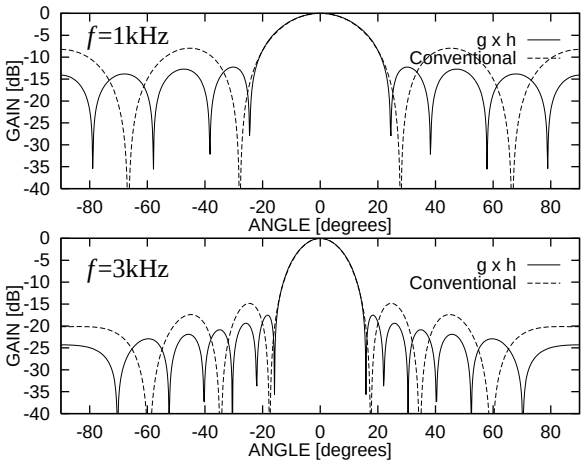

Figure 4: Resultant directivity patterns at 1 kHz (top), and at 3 kHz (bottom).

of directivity patterns at the initial and 20th iteration, the directivity patterns of $\boldsymbol{g}$ (broken line), the directivity patterns of $\boldsymbol{h}$ (dash-dotted line) and these product characteristics, $|\boldsymbol{g}\boldsymbol{a}_d(f) \cdot \boldsymbol{h}\boldsymbol{a}_d(f)|$ (solid line) are shown in Fig. 3 (frequency f is set to be 2 kHz). As shown in Fig. 3, the iterative improvement works towards reducing the magnitude of sidelobes in $|\boldsymbol{g}\boldsymbol{a}_d(f) \cdot \boldsymbol{h}\boldsymbol{a}_d(f)|$ as the number of iterations increases.

The directivity patterns obtained at 40 iterations are designated as the resultant directivity patterns because the squared error converged at 40 iterations for each frequency. The solid lines in Fig. 4 show the resultant directivity patterns $|\boldsymbol{g}\boldsymbol{a}_d(f) \cdot \boldsymbol{h}\boldsymbol{a}_d(f)|$ for 1 and 3 kHz as typical frequencies. For comparison, we also plot the optimized directivity patterns for a conventional DS array based on a single weight vector. It is evident from Fig. 4 that the ability of the sidelobe reduction in the proposed array is improved by about 5 dB for each frequency region.

4. EXPERIMENTS AND RESULTS

Some computer simulations are performed to examine the availability of the proposed method. In this section, the proposed array

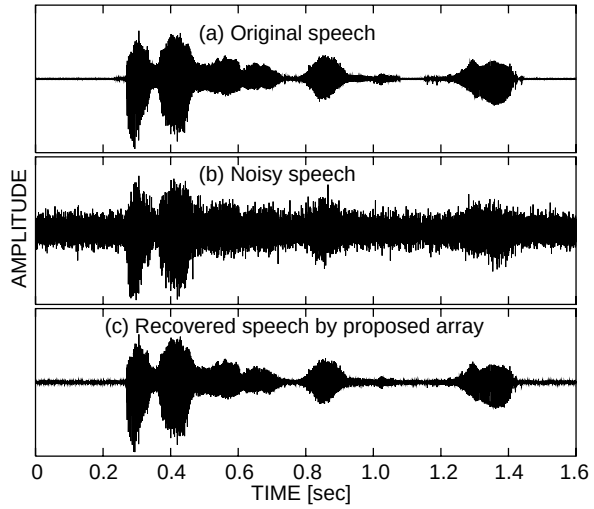

Figure 5: Examples of waveforms, (a) original speech, (b) noisy speech at a microphone (input SNR of 0 dB), (c) recovered speech by proposed array.

shown in Fig. 4 is compared with the conventional DS array shown in Fig. 4 from two standpoints, an objective evaluation of recovered speech quality and a word recognition test.

4.1. Analysis Conditions for Experiments

All sound data prepared in this experiments are sampled at 12 kHz with 16 bit accuracy. To remove the noise components in lower frequency regions, which cannot be reduced by the conventional DS or proposed arrays, all sound data received by microphones are filtered by the highpass filter which has a non-steep spectral envelope as follows: the cut off frequency is set to be 500 Hz and the transient characteristic is -14 dB/oct. Noise reduction processing is conducted frame by frame under the following conditions: the frame length is 21.3 msec, the frame shift is a half of the frame length, and the window function is rectangular. The expectation value of $|S^{(\mathrm{r})}(f)|^2$ in Eq. (8) is calculated by averaging the power spectra of reference signals over 10 frames.

4.2. Objective Evaluation

We generate noisy signals by artificially adding white noises to clean speech signals with different signal-to-noise ratios (SNRs) from -10 to 10 dB. The noises are assumed to arrive from a single direction selected from between 20 and 80°. As for a clean speech material, a Japanese sentence (/arayuru geNjitsu o subete jibuN no ho:e nejimagetanoda/) of female speaker is used.

First, as an example of waveforms, the original speech, the noisy speech at a single microphone (the input SNR of 0 dB and the noise direction of 50°) and the resultant speech of proposed array are shown in Fig. 5, respectively. From these figures, it is clear that the proposed array has a great ability of noise reduction.

Secondly, the output SNRs with different noise directions are shown in Figure 6, where the input SNR is 0 dB. Also, to illustrate the behavior of the proposed array between the different input SNRs, the *noise reduction rate*, defined as output SNR in dB minus input SNR in dB, with the noise direction of 50° is shown in Figure 7. In both Figs. 6 and 7, the solid lines show the results of

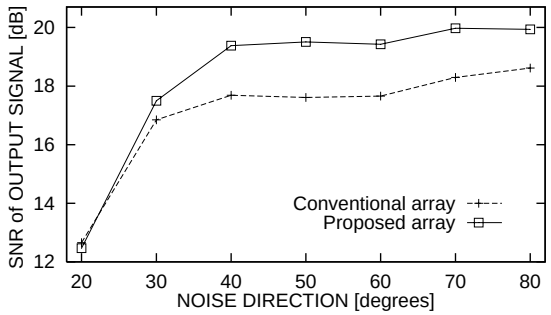

Figure 6: Output SNR with different noise directions (input SNR of 0 dB).

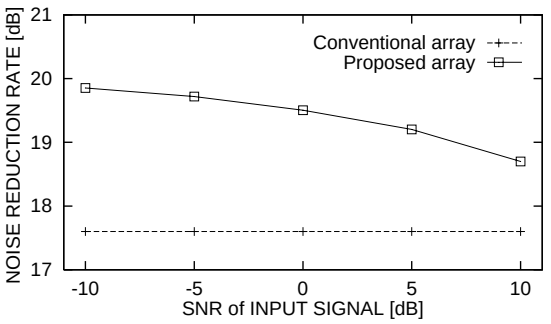

Figure 7: Noise reduction rate with different input SNRs (noise direction of 50°).

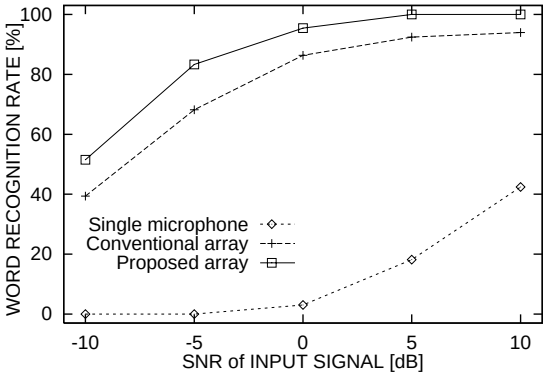

Figure 8: Word recognition rate with different input SNRs.

the proposed array and the broken lines show those of the conventional DS array. From these figures, it is evident that the abilities of noise reduction in the proposed array is the same as that in the optimized conventional array when the noise arrives from the direction at the mainlobe. However, when the noise exists in the direction at the sidelobes, 40–80°, an improvement of about 2 dB in output SNR can be obtained using the proposed array, and the improvement increases as the input SNR decreases.

4.3. Word Recognition Test

The HMM continuous speech recognition (CSR) experiment is performed in a speaker dependent manner. For the CSR experiment, 10 sentences of one female speaker are used as test data, and the monophone HMM model is trained by 140 phonetically balanced sentences. Both test and training set are selected from the ASJ continuous speech corpus for research. The rest of conditions are summarized in Table 1.

Table 1: Analysis Conditions for CSR Experiments

Frame Length	25 msec
Frame Shift	10 msec
Feature Vector	12 MFCC + ΔMFCC + $\Delta\Delta$MFCC ΔPOWER + $\Delta\Delta$POWER
Vocabulary	68
Grammar	no grammar

Figure 8 shows the results of the word recognition rates with different input SNRs. As shown in this figure, the recognition rate using a only single microphone is quite low, and both conventional DS and proposed arrays are effective to improve the word recognition rate. As compared with the results of the conventional DS array, by applying the proposed method, an improvement in recognition rate of more than 10% is obtained in -5 and -10 dB conditions. This indicates that the proposed array is applicable to a robust speech recognition system, especially in low speech quality conditions.

5. CONCLUSION

In this paper, a new nonlinear microphone array and its optimization algorithm for directivity patterns are proposed. From extensive computer simulations, compared with an optimized conventional delay-and-sum array, it is shown that: (1) the proposed array can improve the output SNR by about 2 dB, (2) the proposed array can improve a word recognition rate by about 10% where input SNR condition is -5 or -10 dB.

6. ACKNOWLEDGMENT

The authors are grateful to Dr. Mitsuo Komura in SECOM. CO., LTD. for his suggestions and discussions on this work.

7. REFERENCES

[1] 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, no.5, pp.1508–1518 (1985).

[2] L. J. Griffiths and C. W. Jim: "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. AP*, vol.30, no.1, pp.27–34 (1982).

[3] S. F. Boll: "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol.27, no.2, pp.113-120 (1979).

[4] H. Y. Kim, F. Asano, Y. Suzuki and T. Sone: "Speech enhancement based on short-time spectral amplitude estimation with two-channel beamformer," *IEICE Trans. Fundamentals*, vol.E79-A, no.12, pp.2151–2158 (1996).

[5] J. Meyer and U. Simmer: "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," *Proc. ICASSP 97*, vol.2, pp.1167–1170 (1997).

[6] M. Mizumachi and M. Akagi: "Noise reduction by paired-microphones using spectral subtraction," *Proc. ICASSP 98*, vol.2, pp.1001–1004 (1998).

[7] H. Saruwatari and M. Komura: "Synthetic aperture sonar in air medium using a nonlinear sidelobe canceller," *IEICE Trans. A*, vol.J81-A, no.5, pp.815–826 (1998) (in Japanese).