ROBUST SPATIAL SUBTRACTION ARRAY WITH INDEPENDENT COMPONENT ANALYSIS FOR SPEECH ENHANCEMENT

Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano

Nara Institute of Science and Technology
8916-5 Takaya-cho, Ikoma-shi, Nara, 630-0192 JAPAN

ABSTRACT

In this paper, we propose a new spatial subtraction array (SSA) structure which includes independent component analysis (ICA)-based noise estimator. Recently, SSA has been proposed to realize noise-robust hands-free speech recognition. In SSA, noise reduction is achieved by subtracting the estimated noise power spectrum from the noisy speech power spectrum. The conventional SSA uses null beamformer (NBF) as a noise estimator, but NBF suffers from the adverse effect of microphone-element errors and room reverberations in real environments. To improve the problem, we newly replace NBF with ICA which can adapt its own separation filters to the element error and the reverberation. The affections by the element error and the reverberation can be mitigated in the proposed ICA-based noise estimator. Experimental results reveal that the accuracy of noise estimation by ICA outperforms that of NBF, and speech recognition performance of the proposed method overtakes that of the conventional SSA.

1. INTRODUCTION

A hands-free speech recognition system is essential for realizing an intuitive and stress-free human-machine interface. However, the quality of the distant-talking speech is always inferior to that of using close-talking microphone, and this leads to degradations of speech recognition. One approach for establishing a noise-robust speech recognition system is to enhance the speech signals by introducing microphone array signal processing. In delay-and-Sum (DS) array, we compensates the time delay for each element to reinforce the target signal arriving from the look direction. On the other hand, null beamformer (NBF) [1] provides more efficient noise reduction in which we steer the directional null to the direction of the noise signal. Moreover, Griffith-Jim adaptive array (GJ) [2] can achieve a superior performance relative to others. However, GJ requires a huge amount of calculations for learning adaptive multichannel FIR filters of, e.g., thousands or millions taps in total.

Spatial subtraction array (SSA) [3] is a successful candidate for hands-free speech recognition, and SSA is specifically designed for a speech recognition application. In SSA, noise reduction is achieved by subtracting the estimated noise power spectrum by NBF from the power spectrum of noisy observations in mel-scale filter bank domain. Since a common speech recognizer is not so sensitive to phase information, SSA which is performing subtraction processing only in the power spectrum domain is more applicable to the speech recognition, and it is reported that the speech recognition performance of SSA outperforms those of DS and GJ [3]. In SSA, noise estimation is performed by NBF which has decent performance under ideal conditions. However, NBF sustains the negative affection by microphone-element error and room reverberations. Therefore, in the real environment where the element error and the reverberation are always included, the performance of SSA significantly decreases because the noise-estimation accuracy by NBF decreases.

The work was partly supported by MEXT e-Society leading project.

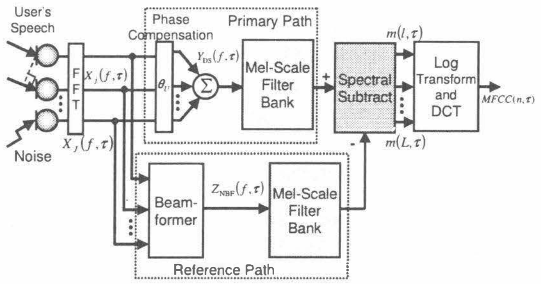

Fig. 1. Block diagram of conventional SSA.

In this paper, we propose a new SSA structure which replaces NBF-based noise estimator with independent component analysis (ICA)[4]-based noise estimator. ICA is a technique for source separation based on independence among multiple source signals. In acoustic source separation scenarios, ICA can also extract each source signal only using observed signals at the microphone array, and ICA does not require characteristics about sensor elements and the reverberation. Therefore, it is well expected that ICA can adapt its own separation filters to the element error and the reverberation. Accordingly the adverse effect by the element error and the reverberation can be mitigated in the proposed ICA-based noise estimator. Real-recording-based simulations are conducted, and we can indicate that the proposed method outperforms the conventional SSA on the basis of speech recognition performances.

2. CONVENTIONAL SPATIAL SUBTRACTION ARRAY

2.1. Overview

The conventional SSA [3] consists of a DS-based primary path and a reference path via the NBF-based noise estimation (see Fig. 1). The estimated noise component by NBF is efficiently subtracted from the primary path in the power spectrum domain without phase information. In SSA, we assume that the target speech direction and speech break interval are known in advance. Detailed signal processing is shown below.

2.2. Partial speech enhancement in primary path

First, the short-time analysis of observed signals is conducted by a frame-by-frame discrete Fourier transform (DFT). By plotting the spectral values in a frequency bin for each microphone input frame by frame, we consider these values as a time series. Hereafter, we designate the time series as

$$X(f,\tau) = [X_1(f,\tau),\ldots,X_J(f,\tau)]^{\mathrm{T}}, \tag{1}$$

1-4244-0779-6/07/$20.00 ©2007 IEEE

where J is the number of microphones, f is the frequency bin and τ is the frame number. Also, $X(f,\tau)$ can be rewritten as

$$X(f,\tau) = A(f)\left(S(f,\tau) + N(f,\tau)\right), \qquad (2)$$

$$S(f,\tau) = [\underbrace{0,\ldots,0}_{U-1}, S_U(f,\tau), \underbrace{0,\ldots,0}_{K-U}]^{\mathrm{T}}, \qquad (3)$$

$$N(f,\tau) = [N_1(f,\tau),\ldots,N_{U-1}(f,\tau),0,N_{U+1}(f,\tau),\ldots,N_K(f,\tau)]^{\mathrm{T}}, \qquad (4)$$

where $A(f)$ is a mixing matrix, $S(f,\tau)$ is a target speech signal vector, $N(f,\tau)$ is a noise signal vector, U expresses the target speech number, and K is the number of sound sources.

Next, the target speech signal is partly enhanced in advance by DS. This procedure can be given as

$$\begin{aligned} Y_{\mathrm{DS}}(f,\tau) &= W_{\mathrm{DS}}^{\mathrm{T}}(f)X(f,\tau) \\ &= W_{\mathrm{DS}}^{\mathrm{T}}(f)A(f)S(f,\tau) + W_{\mathrm{DS}}^{\mathrm{T}}(f)A(f)N(f,\tau), \qquad (5) \end{aligned}$$

$$W_{\mathrm{DS}}(f) = [W_1^{(\mathrm{DS})}(f),\ldots,W_J^{(\mathrm{DS})}(f)]^{\mathrm{T}}, \qquad (6)$$

$$W_j^{(\mathrm{DS})}(f) = \frac{1}{J}\exp\left(-i2\pi(f/M)f_s d_j \sin\theta_U/c\right), \qquad (7)$$

where $Y_{\mathrm{DS}}(f,\tau)$ is a primary-path output which slightly enhances the target speech, $W_{\mathrm{DS}}(f)$ is a filter coefficient vector of DS, M is the DFT size, f_s is sampling frequency, d_j is a microphone position, and c is sound velocity. Besides, θ_U is a known direction-of-arrival (DOA) of the target speech. In Eq. (5), the second term in the right-hand side expresses the remaining noise in the output of the primary path.

2.3. Noise estimation in reference path

In the reference path, we estimate the noise signal by using NBF. This procedure is given as

$$Z_{\mathrm{NBF}}(f,\tau) = W_{\mathrm{NBF}}^{\mathrm{T}}(f)X(f,\tau), \qquad (8)$$

$$W_{\mathrm{NBF}}(f) = \left\{[1,0]\cdot[a(f,\theta_O), a(f,\theta_U)]^{+}\right\}^{\mathrm{T}}, \qquad (9)$$

$$a(f,\theta) = [a_1(f,\theta),\ldots,a_J(f,\theta)]^{\mathrm{T}}, \qquad (10)$$

$$a_j(f,\theta) = \exp\left(i2\pi(f/M)f_s d_j \sin\theta/c\right), \qquad (11)$$

where $Z_{\mathrm{NBF}}(f,\tau)$ is the estimated noise by NBF, $W_{\mathrm{NBF}}(f)$ is a NBF-filter coefficient vector which steers the directional null in the direction of the DOA of the target speech, θ_U, and steers unit gain in the arbitrary direction $\theta_O(\neq \theta_U)$. $a(f,\theta)$ is a steering vector which expresses phase information of the sound source arriving from the direction θ. Besides, M^{+} denotes Moore-Penrose pseudo inverse matrix of M. This processing can suppress the target speech arriving from θ_U, which is equal to an extraction of noises from sound mixtures if we take into account affections of sensor errors and reverberations. Thus we can estimate the noise signals by NBF under ideal conditions. Note that $Z_{\mathrm{NBF}}(f,\tau)$ is the function of the frame number τ, unlike the constant noise prototype estimated in the traditional spectral subtraction method [5]. Therefore, SSA can deal with a *non-stationary* noise.

2.4. Mel-scale filter bank analysis

SSA includes mel-scale filter bank analysis, and outputs mel-frequency cepstrum coefficient (MFCC) [6]. The triangular window $W_{\mathrm{mel}}(k;l)$ $(l = 1,\cdots,L)$ to perform mel-scale filter bank analysis is designated as follows:

$$W_{\mathrm{mel}}(f,l) = \begin{cases} \dfrac{f - f_{\mathrm{lo}}(l)}{f_{\mathrm{c}}(l) - f_{\mathrm{lo}}(l)} & (f_{\mathrm{lo}}(l) \le f \le f_{\mathrm{c}}(l)), \\ \dfrac{f_{\mathrm{hi}}(l) - f}{f_{\mathrm{hi}}(l) - f_{\mathrm{c}}(l)} & (f_{\mathrm{c}}(l) \le f \le f_{\mathrm{hi}}(l)), \end{cases} \qquad (12)$$

where $f_{\mathrm{lo}}(l)$, $f_{\mathrm{c}}(l)$, and $f_{\mathrm{hi}}(l)$ are the lower, center, and higher frequency bins of each triangle window, respectively. They satisfy the relation among adjacent windows as

$$f_{\mathrm{c}}(l) = f_{\mathrm{hi}}(l-1) = f_{\mathrm{lo}}(l+1). \qquad (13)$$

Moreover, $f_{\mathrm{c}}(l)$ is arranged in regular intervals on mel-frequency domain. Mel-scale frequency $Mel_{f_{\mathrm{c}}(l)}$ for $f_{\mathrm{c}}(l)$ is calculated as

$$Mel_{f_{\mathrm{c}}(l)} = 2595\log_{10}\left\{1 + \frac{f_{\mathrm{c}}(l)f_s}{700\cdot M}\right\}. \qquad (14)$$

2.5. Noise reduction processing

In SSA, noise reduction is carried out by subtracting the estimated noise power spectrum from the partly enhanced target speech power spectrum in the mel-scale filter bank domain as

$$m(l,\tau) = \begin{cases} \displaystyle\sum_{f=f_{lo}(l)}^{f_{hi}(l)} W_{\mathrm{mel}}(f;l)\{|Y_{\mathrm{DS}}(f,\tau)|^2 - \alpha(l)\cdot\beta\cdot|Z_{\mathrm{NBF}}(f,\tau)|^2\}^{\frac{1}{2}} \\ \qquad (\text{ if } |Y_{\mathrm{DS}}(f,\tau)|^2 - \alpha(l)\cdot\beta\cdot|Z_{\mathrm{NBF}}(f,\tau)|^2 \ge 0\ ), \\ \displaystyle\sum_{f=f_{lo}(l)}^{f_{hi}(l)} W_{\mathrm{mel}}(f;l)\{\gamma\cdot|Y_{\mathrm{DS}}(f,\tau)|\} \ (\text{otherwise}), \end{cases} \qquad (15)$$

where $m(l,\tau)$ is the output from the mel-scale filter bank. The system switches in two equations depending on the conditions in Eq. (15). $m(l,\tau)$ is a function of the over-subtraction parameter β and the parameter $\alpha(l)$ which is determined during a speech break so that the resultant output $m(l,\tau)$ is zero. On the other hand, if the power spectrum takes a negative value, $m(l,\tau)$ is obtained by using flooring processing, where γ is the flooring coefficient.

Since a common speech recognition is not so sensitive to phase information, SSA which is performing subtraction processing in the power domain is more applicable to the speech recognition. Moreover, in general, the order of the filter bank l is set to 24, and consequently SSA optimizes only 24 parameters. On the other hand, GJ requires the adaptive learning of FIR-filters of thousands or millions of taps. Finally, we perform mel-scale filter bank analysis, log transform and discrete cosine transform to obtain MFCC for speech recognizer.

3. PROPOSED METHOD

3.1. Error robustness analysis for noise estimation by NBF

In this section, we discuss the problem of the conventional SSA. The NBF-based noise estimator is used in the conventional SSA, but NBF suffers from the adverse effect of the microphone element error and the room reverberation. NBF is a technique to suppress an interference source signal by generating a null against the direction of the interference source signal. If the interference source signal arrives from the same direction as the null, we can suppress the interference source signal perfectly. In a reverberant environment, however, the interference source signal arrives from not only the null's direction but also outside of the direction. Therefore, in the reverberant room, we cannot suppress the interference source signal sufficiently. In addition, a microphone element usually involves gain and phase errors. NBF is designed under the ideal assumption that all elements have the same characteristics. In the real environment, however, the characteristics of each element are different. From the above-mentioned fact, the directivity pattern shaped by NBF in the ideal environment is apart from that of in the real environment.

Figure 2 illustrates directivity patterns which are shaped by two-element NBF in the ideal (solid line) and the real (dotted line) environment where the reverberation time is 200 ms. In this figure, the null direction is set to zero degree. We can see that the depth of the null in the real environment which contains the element error and the reverberation shallows. Therefore, we cannot suppress the interference source signal completely in the real environment by using NBF. Indeed, in SSA, we perform noise estimation via NBF which steers null against the target speech signal, but we cannot suppress the target speech signal sufficiently. In fact, NBF cannot estimate noise signal completely.

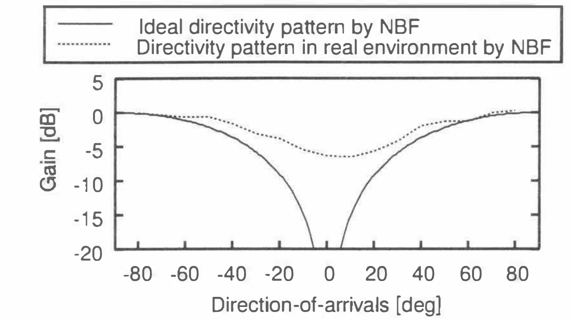

Fig. 2. Directivity patterns shaped by NBF in ideal environment and real environment which contains element error and reverberation.

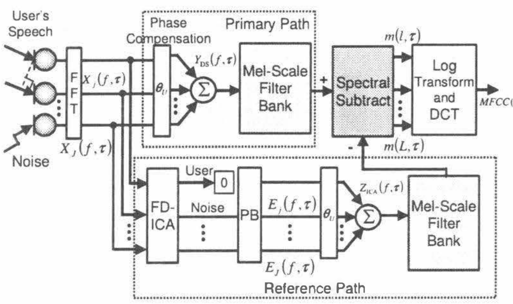

Fig. 3. Block diagram of proposed method.

Thus the improvement of robustness in the noise estimator part is a problem demanding prompt attention.

3.2. Strategy of proposed method

We propose an improved SSA which includes ICA-based noise estimator instead of NBF-based noise estimator to address the problems which are discussed in the previous section. In the proposed method, the primary path and noise reduction processing are the same as the conventional SSA. As for the reference path, we newly introduce ICA as a robust noise estimator for adapting the filters to the element error and the reverberation (see Fig. 3). In ICA, an unmixing matrix is optimized so that output signals become mutually independent only using observed signals, and a priori information about the sensors and the room acoustics is not required. Therefore the proposed method can reduce these adverse effects because ICA can estimate noise signals which involve whole characteristics of the microphone elements and the reverberation. Detailed signal processing is shown below.

3.3. ICA-based noise estimation in reference path

The proposed method includes ICA-based noise estimation. In ICA part, we perform signal separation using the complex valued unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$, so that the output signals $\boldsymbol{O}(f,\tau)=[O_1(f,\tau),\ldots,O_J(f,\tau)]^{\mathrm{T}}$ become mutually independent; this procedure can be represented by

$$\boldsymbol{O}(f,\tau) = \boldsymbol{W}(f)\boldsymbol{X}(f,\tau), \tag{16}$$

$$\boldsymbol{W}(f) = \boldsymbol{P}(f)\boldsymbol{W}_{\mathrm{ICA}}(f), \tag{17}$$

where $\boldsymbol{P}(f)$ is a permutation matrix and $\boldsymbol{W}(f)$ is a new unmixing matrix which resolves the permutation problem. The permutation matrix $\boldsymbol{P}(f)$ is determined by looking at null directions in the directivity pattern which is shaped by $\boldsymbol{W}_{\mathrm{ICA}}(f)$ [1], so that the U-th output $O_U(f,\tau)$ is set to the target speech signal. The optimal $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is obtained by the following iterative updating equation [7]:

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) = \mu\left[\boldsymbol{I} - \langle \boldsymbol{\Phi}(\boldsymbol{O}(f,\tau))\,\boldsymbol{O}^{\mathrm{H}}(f,\tau)\rangle_\tau\right]\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f), \tag{18}$$

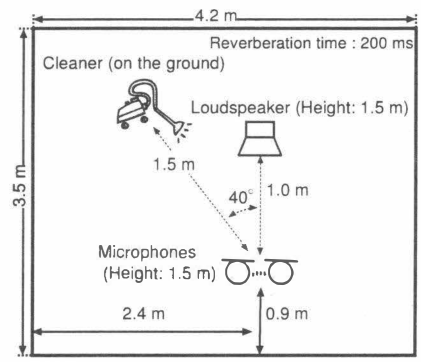

Fig. 4. Layout of reverberant room used in our experiment.

where μ is the step-size parameter, $[p]$ is used to express the value of the p-th step in the iterations, and $\boldsymbol{I}$ is an identity matrix. Besides, $\langle\cdot\rangle_\tau$ denotes a time-averaging operator, $\boldsymbol{M}^{\mathrm{H}}$ denotes conjugate transpose of matrix $\boldsymbol{M}$, and $\boldsymbol{\Phi}(\cdot)$ is the appropriate nonlinear vector function [1]. In the reference path, the target signal is not required because we want to estimate only the noise component. Accordingly we remove the separated speech component $O_U(f,\tau)$ from ICA outputs $\boldsymbol{O}(f,\tau)$, and construct the following "noise-only vector," $\boldsymbol{Q}(f,\tau)$;

$$\boldsymbol{Q}(f,\tau) = [O_1(f,\tau),\ldots,O_{U-1}(f,\tau),0,O_{U+1}(f,\tau),\ldots,O_J(f,\tau)]^{\mathrm{T}}. \tag{19}$$

Next, we apply the projection back (PB) [8] method to remove the ambiguity of amplitude. This procedure can be written as

$$\boldsymbol{E}(f,\tau) = \boldsymbol{W}^{+}(f)\boldsymbol{Q}(f,\tau). \tag{20}$$

Here, $\boldsymbol{Q}(f,\tau)$ is composed of only noise components. Therefore, $\boldsymbol{E}(f,\tau)$ is a good estimation of the received noise signals at the microphone positions;

$$\boldsymbol{E}(f,\tau) \simeq \boldsymbol{A}(f)\boldsymbol{N}(f,\tau). \tag{21}$$

Finally, we obtain the estimated noise signal $Z_{\mathrm{ICA}}(f,\tau)$ by performing DS as follows:

$$Z_{\mathrm{ICA}}(f,\tau) = \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{E}(f,\tau) \simeq \boldsymbol{W}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{N}(f,\tau). \tag{22}$$

Equation (22) is expected to be equal to the noise term of Eq. (5) in the primary path. Of course, Eq. (22) contains estimation errors to some extent. Even though the level of the noise estimation error is not negligible, we can still enhance the target speech via over-subtraction [5] in the power spectrum domain.

4. EXPERIMENTS AND RESULT

4.1. Experimental setup

Figure 4 shows a layout of the reverberant room used in our experiments. We use the following 16 kHz sampled signals as test data; the original speech convoluted with the impulse responses recorded in the real environment, and added with a cleaner noise which was recored in the real environment. The cleaner noise is not a point source but consists of several non-stationary noises emitted from, e.g., a motor, air duct and nozzle. Moreover the cleaner noise includes background noise. The input signal-to-noise ratio (SNR) is set to 5, 10, or 15 dB at the array. A four-element array with the interelement spacing of 2 cm is used, and DFT size is 512. Over-subtraction parameter β is 1.4 and flooring coefficient γ is 0.2.

4.2. Accuracy of estimated noise signal

First, we analyze the directivity pattern shaped by ICA in the real environment. Figure 5 depicts the directivity pattern of ICA (broken line) in the real environment. From this result, we can confirm that the null shaped by ICA becomes deep compared with that of the NBF-based conventional SSA. Therefore, it is

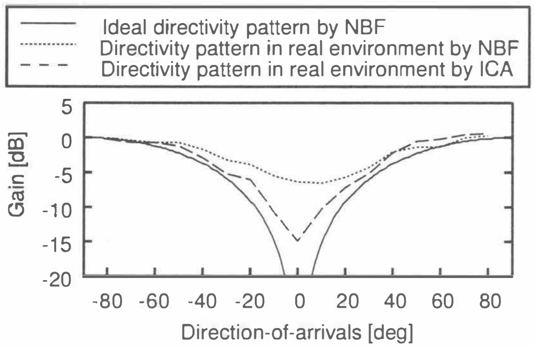

Fig. 5. Directivity patterns shaped by NBF and ICA in ideal environment and real environment which contains element error and reverberation.

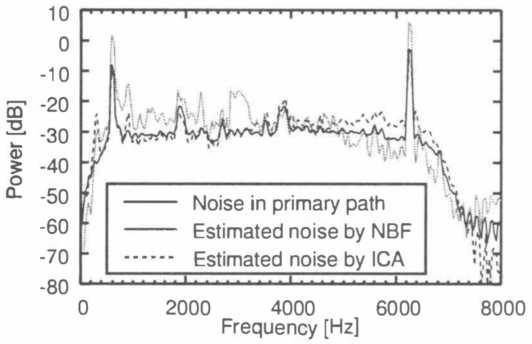

Fig. 6. Accuracy of estimated noise signal by NBF and ICA.

expected that the target speech suppression performance of ICA (equals the accuracy of the noise estimation) outperforms that of NBF. Next, we compare the conventional SSA and the proposed method in the accuracy of the estimated noise signal. Figure 6 shows the long-term-averaged power spectra of the estimated noise signals by NBF and ICA. The black solid line indicates the power spectrum of the noise signal in the primary path, and this power spectrum is needed to be estimated. The gray solid line represents the power spectrum of the estimated noise signal by NBF, and the dotted line shows the power spectrum of the estimated noise signal by ICA. We can see that the power spectrum of the estimated noise signal by NBF is not accurate. This is due to that the target speech component still remains in the output of NBF because the null shaped by NBF is shallow. On the other hand, we can see that the power spectrum of the estimated noise signal by ICA is a good estimation because the depth of the null shaped by ICA is enough for suppressing the target speech. This result points out that ICA-based noise estimator is a more accurate noise estimator than NBF-based one. This gives propriety in which we use ICA as a noise estimator.

4.3. Results of speech recognition performance

We compare DS, the conventional SSA, and the proposed method on the basis of word accuracy scores. Table 1 describes the conditions for speech recognition, and we use 46 speakers (200 sentences) as original speech. Figure 7 shows the word accuracy in each method. Here, "Unprocessed" refers to the result without any noise reduction processing. From this result, we can see that the word accuracy of the proposed method is obviously superior to those of the conventional methods. This is a promising evidence that the proposed method has an applicability to noise-robust speech recognition rather than the conventional SSA.

Table 1. Conditions for speech recognition

Database	JNAS [9], 306 speakers (150 sentences / 1 speaker)
Task	20 k newspaper dictation
Acoustic model	phonetic tied mixture (PTM) [9], clean model
Number of training speakers for acoustic model	260 speakers (150 sentences / 1 speaker)
Decoder	JULIUS [9] ver 3.5.1

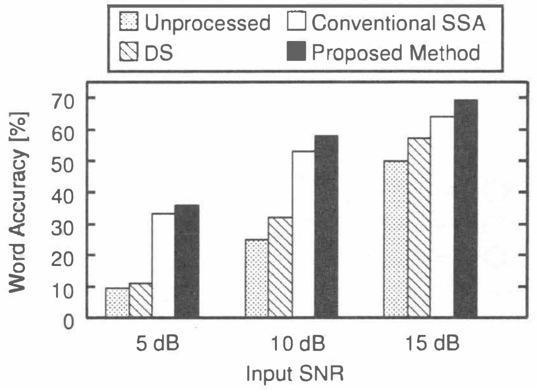

Fig. 7. Results of word accuracy in each method.

5. CONCLUSIONS

In this paper, we proposed a new SSA which involves ICA-based noise estimation to realize a robust hands-free speech recognition in noisy environments. First, we pointed out NBF suffers from the adverse effect of the element error and the reverberation in the real environment. Secondly, based on the above-mentioned fact, we proposed a new SSA structure which replaces NBF-based noise estimator in the conventional SSA with ICA-based noise estimator. Finally, it was confirmed that the word accuracy of the proposed method overtook that of the conventional SSA in the experiment.

6. REFERENCES

[1] H. Saruwatari, et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Applied Signal Proc.*, vol.2003, no.11, pp.1135–1146, 2003.

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

[3] Y. Ohashi, et al., "Noise robust speech recognition based on spatial subtraction array," *Proc. NSIP*, pp.324–327, 2005.

[4] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol.36, pp.287–314, 1994.

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

[6] S. B. Davis, et al., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol.ASSP-28, no.4, pp.357–366, 1982.

[7] P. Smaragdis, "Blind separation of convoluted mixtures in the frequency domain," *Neurocomputing*, vol.22, pp.21–34, 1998.

[8] S. Ikeda and N. Murata, "A method of ICA in the frequency domain," *Proc. International Workshop on ICA and BSS*, pp.365–371, 1999.

[9] A. Lee, et al., "Julius – an open source real-time large vocabulary recognition engine," *Proc. EUROSPEECH*, pp.1691–1694, 2001.