4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

SPEECH ENHANCEMENT AND RECOGNITION IN CAR ENVIRONMENT USING BLIND SOURCE SEPARATION AND SUBBAND ELIMINATION PROCESSING

Hiroshi SARUWATARI, Katsuyuki SAWAI, Akinobu LEE, Kiyohiro SHIKANO, Atsunobu KAMINUMA†*, and Masao SAKATA*†

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN (e-mail: sawatari@is.aist-nara.ac.jp)
†Nissan Research Center, NISSAN MOTOR CO., LTD.
1 Natsushima-cho, Yokosuka-shi, Kanagawa 237-8523, JAPAN

ABSTRACT

We propose a new algorithm for blind source separation (BSS), in which independent component analysis (ICA) and beamforming are combined to resolve the low-convergence problem through optimization in ICA. The proposed method consists of the following four parts: (1) frequency-domain ICA with direction-of-arrival (DOA) estimation, (2) null beamforming based on the estimated DOA, (3) diversity of (1) and (2) in both iteration and frequency domain, and (4) subband elimination (SBE) based on the independence among the separated signals. The temporal alternation between ICA and beamforming can realize fast- and high-convergence optimization. Also SBE enforcedly eliminates the subband components in which the separation could not be performed well. The experiment in a real car environment reveals that the proposed method can improve the qualities of the separated speech and word recognition rates for both directional and diffusive noises.

1. INTRODUCTION

Blind source separation (BSS) is the approach taken to estimate original source signals using only the information of the mixed signals observed in each input channel. This technique is applicable to the realization of noise-robust speech recognition and high-quality hands-free telecommunication systems. In the recent works for the BSS based on the independent component analysis (ICA) [1], several methods, in which the inverse of the complex mixing matrices are calculated in the frequency domain, have been proposed to deal with the arrival lags among each of the elements of the microphone array system [2, 3, 4]. However, this ICA-based approach has the disadvantage that there is difficulty with the low convergence of nonlinear optimization [5].

In this paper, we describe a new algorithm for BSS in which ICA and subband elimination processing are combined. The proposed method consists of the following four parts: (1) frequency-domain ICA with estimation of the direction of arrival (DOA) of the sound source, (2) null beamforming based on the estimated DOA, and (3) integration of (1) and (2) based on the algorithm diversity in both iteration and frequency domain, and (4) subband elimination (SBE) based on the independence among the separated signals. The temporal utilization of null beamforming through ICA iterations can realize fast- and high-convergence optimization. Also the SBE can work so as to find the specific subbands in which the bad separation was performed, and to eliminate them enforcedly.

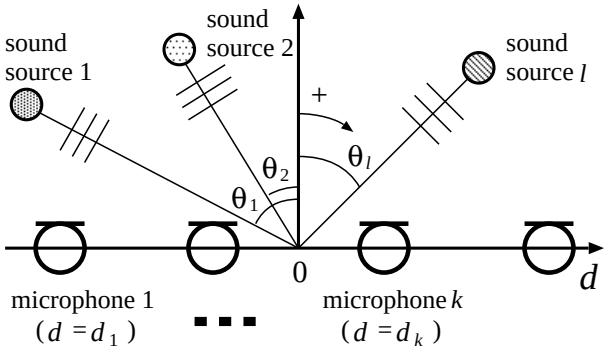

Fig. 1. Configuration of a microphone array and signals.

The experiment in a real car environment reveals that the proposed BSS with SBE is remarkably effective to improve the qualities of the separated speech and word recognition rates for both directional and diffusive noises.

2. DATA MODEL AND CONVENTIONAL BSS METHOD

In this study, a straight-line array is assumed. The coordinates of the elements are designated as d_k $(k = 1, \cdots, K)$,

and the directions of arrival of multiple sound sources are designated as θ_l ($l = 1, \cdots, L$) (see Fig. 1), where we deal with the case of $K = L = 2$.

In general, the observed signals in which multiple source signals are mixed linearly are given by the following equation in the frequency domain:

$$\boldsymbol{X}(f) = \boldsymbol{A}(f)\boldsymbol{S}(f), \tag{1}$$

where $\boldsymbol{X}(f)$ is the observed signal vector, $\boldsymbol{S}(f)$ is the source signal vector, and $\boldsymbol{A}(f)$ is the mixing matrix; these are given as

$$\boldsymbol{X}(f) = [X_1(f), \cdots, X_K(f)]^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{S}(f) = [S_1(f), \cdots, S_L(f)]^{\mathrm{T}}, \tag{3}$$

$$\boldsymbol{A}(f) = \begin{bmatrix} A_{11}(f) & \cdots & A_{1L}(f) \\ \vdots & & \vdots \\ A_{K1}(f) & \cdots & A_{KL}(f) \end{bmatrix}. \tag{4}$$

$\boldsymbol{A}(f)$ is assumed to be complex-valued because we introduce a model to deal with the arrival lags among each of the elements of the microphone array and room reverberations.

In the frequency-domain ICA, first, the short-time analysis of observed signals is conducted by frame-by-frame discrete Fourier transform (DFT) (see Fig. 2). By plotting the spectral values in a frequency bin of each microphone input frame by frame, we consider them as a time series. Hereafter, we designate the time series as

$$\boldsymbol{X}(f,t) = [X_1(f,t), \cdots, X_K(f,t)]^{\mathrm{T}}. \tag{5}$$

Next, we perform signal separation using the complex-valued inverse of the mixing matrix, $\boldsymbol{W}(f)$, so that the L time-series output $\boldsymbol{Y}(f,t)$ becomes mutually independent; this procedure can be given as

$$\boldsymbol{Y}(f,t) = \boldsymbol{W}(f)\boldsymbol{X}(f,t), \tag{6}$$

where

$$\boldsymbol{Y}(f,t) = [Y_1(f,t), \cdots, Y_L(f,t)]^{\mathrm{T}}, \tag{7}$$

$$\boldsymbol{W}(f) = \begin{bmatrix} W_{11}(f) & \cdots & W_{1K}(f) \\ \vdots & & \vdots \\ W_{L1}(f) & \cdots & W_{LK}(f) \end{bmatrix}. \tag{8}$$

We perform this procedure with respect to all frequency bins. Finally, by applying the inverse DFT and the overlap-add technique to the separated time series $\boldsymbol{Y}(f,t)$, we reconstruct the resultant source signals in the time domain.

In the conventional ICA-based BSS method, the optimal $\boldsymbol{W}(f)$ is obtained by the following iterative equation [2, 6]:

$$\boldsymbol{W}_{i+1}(f) = \eta\Big[\mathrm{diag}\Big(\big\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\big\rangle_t\Big) - \big\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\big\rangle_t\Big]\boldsymbol{W}_i(f) + \boldsymbol{W}_i(f), \tag{9}$$

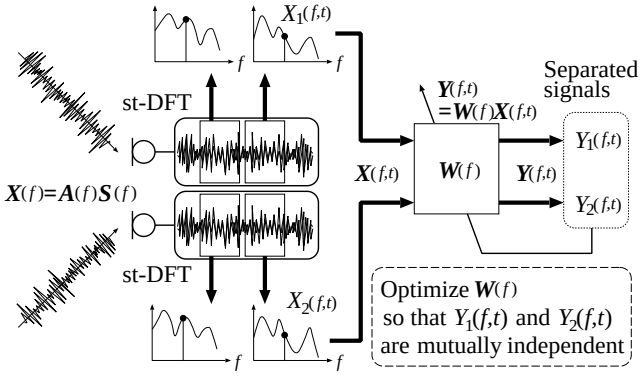

Fig. 2. BSS procedure based on frequency-domain ICA.

where $\langle\cdot\rangle_t$ denotes the time-averaging operator, i is used to express the value of the i th step in the iterations, and η is the step-size parameter. Also, we define the nonlinear vector function $\boldsymbol{\Phi}(\cdot)$ as

$$\boldsymbol{\Phi}(\boldsymbol{Y}(f,t)) \equiv [\Phi(Y_1(f,t)), \cdots, \Phi(Y_L(f,t))]^{\mathrm{T}}, \tag{10}$$

$$\Phi(Y_l(f,t)) \equiv \big[1 + \exp(-Y_l^{(\mathrm{R})}(f,t))\big]^{-1} + j \cdot \big[1 + \exp(-Y_l^{(\mathrm{I})}(f,t))\big]^{-1}, \tag{11}$$

where $Y_l^{(\mathrm{R})}(f,t)$ and $Y_l^{(\mathrm{I})}(f,t)$ are the real and imaginary parts of $Y_l(f,t)$, respectively.

3. PROPOSED ALGORITHM

3.1. Fast-convergence algorithm [7]

The conventional ICA method inherently has a significant disadvantage which is due to low convergence through nonlinear optimization in ICA. In order to resolve the problem, we propose an algorithm based on the temporal alternation of learning between ICA and beamforming; the inverse of the mixing matrix, $\boldsymbol{W}(f)$, obtained through ICA is temporally substituted by the matrix based on null beamforming for a temporal initialization or acceleration of the iterative optimization. The proposed algorithm is conducted by the following steps with respect to all frequency bins in parallel (see Fig. 3).

[Step 1: Initialization] Set the initial $\boldsymbol{W}_i(f)$, i.e., $\boldsymbol{W}_0(f)$, to an arbitrary value, where the subscripts i is set to be 0.

[Step 2: 1-time ICA iteration] Optimize $\boldsymbol{W}_i(f)$ using the following 1-time ICA iteration:

$$\boldsymbol{W}_{i+1}^{(\mathrm{ICA})}(f) = \eta\Big[\mathrm{diag}\Big(\big\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\big\rangle_t\Big) - \big\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\big\rangle_t\Big]\boldsymbol{W}_i(f) + \boldsymbol{W}_i(f), \tag{12}$$

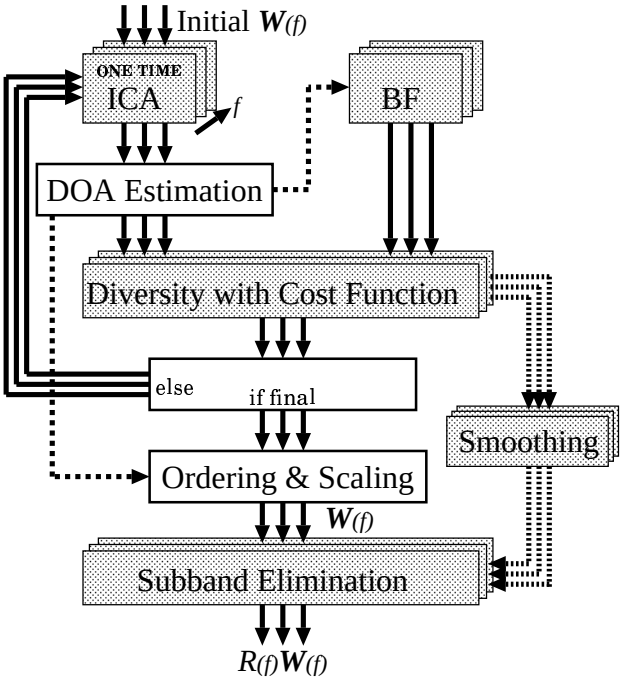

Fig. 3. Proposed algorithm combining frequency-domain ICA and beamforming with subband elimination.

where the superscript "(ICA)" is used to express that the inverse of the mixing matrix is obtained by ICA.

[Step 3: DOA estimation] Estimate DOAs of the sound sources by utilizing the directivity pattern of the array system, $F_l(f,\theta)$, which is given by

$$F_l(f,\theta) = \sum_{k=1}^{K} W_{lk}^{(\mathrm{ICA})}(f) \exp\left[j2\pi f d_k \sin\theta/c\right], \quad (13)$$

where $W_{lk}^{(\mathrm{ICA})}(f)$ is the element of $\boldsymbol{W}_{i+1}^{(\mathrm{ICA})}(f)$. In the directivity patterns, directional nulls exist in only two particular directions. Accordingly, by obtaining statistics with respect to the directions of nulls at all frequency bins, we can estimate the DOAs of the sound sources. The DOA of the l th sound source, $\hat{\theta}_l$, can be estimated as

$$\hat{\theta}_l = 2\sum_{m=1}^{N/2} \theta_l(f_m)/N, \quad (14)$$

where N is a total point of DFT, and $\theta_l(f_m)$ represents the DOA of the l th sound source at the m th frequency bin. These are given by

$$\theta_1(f_m) = \min\left[\operatorname*{argmin}_{\theta}|F_1(f_m,\theta)|, \operatorname*{argmin}_{\theta}|F_2(f_m,\theta)|\right], \quad (15)$$

$$\theta_2(f_m) = \max\left[\operatorname*{argmin}_{\theta}|F_1(f_m,\theta)|, \operatorname*{argmin}_{\theta}|F_2(f_m,\theta)|\right], \quad (16)$$

where $\min[x,y]$ ($\max[x,y]$) is defined as a function in order to obtain the smaller (larger) value among x and y.

[Step 4: Beamforming] Construct an alternative matrix for signal separation, $\boldsymbol{W}^{(\mathrm{BF})}(f)$, based on the null-beamforming technique where the DOA results obtained in the previous step is used. In the case that the look direction is $\hat{\theta}_1$ and the directional null is steered to $\hat{\theta}_2$, the elements of the matrix for signal separation are given as

$$W_{11}^{(\mathrm{BF})}(f_m) = -\exp\left[-j2\pi f_m d_1 \sin\hat{\theta}_2/c\right] \times \left\{-\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c\right] + \exp\left[j2\pi f_m d_2(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c\right]\right\}^{-1}, \quad (17)$$

$$W_{12}^{(\mathrm{BF})}(f_m) = \exp\left[-j2\pi f_m d_2 \sin\hat{\theta}_2/c\right] \times \left\{-\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c\right] + \exp\left[j2\pi f_m d_2(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c\right]\right\}^{-1}. \quad (18)$$

Also, in the case that the look direction is $\hat{\theta}_2$ and the directional null is steered to $\hat{\theta}_1$, the elements of the matrix are given as

$$W_{21}^{(\mathrm{BF})}(f_m) = \exp\left[-j2\pi f_m d_1 \sin\hat{\theta}_1/c\right] \times \left\{\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c\right] - \exp\left[j2\pi f_m d_2(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c\right]\right\}^{-1}, \quad (19)$$

$$W_{22}^{(\mathrm{BF})}(f_m) = -\exp\left[-j2\pi f_m d_2 \sin\hat{\theta}_1/c\right] \times \left\{\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c\right] - \exp\left[j2\pi f_m d_2(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c\right]\right\}^{-1}. \quad (20)$$

[Step 5: Diversity with cost function] Select the most suitable unmixing matrix in each frequency bin and each iteration point, i.e., algorithm diversity in both iteration and frequency domain. As a cost function used to achieve the diversity, we calculate two kinds of cosine distances between the separated signals which are obtained by ICA and beamforming. These are given by

$$J^{(\mathrm{ICA})}(f)$$

$$= \frac{\left|\left\langle Y_1^{(\mathrm{ICA})}(f,t)Y_2^{(\mathrm{ICA})}(f,t)^*\right\rangle_t\right|}{\left\langle\left|Y_1^{(\mathrm{ICA})}(f,t)\right|^2\right\rangle_t^{\frac{1}{2}}\left\langle\left|Y_2^{(\mathrm{ICA})}(f,t)\right|^2\right\rangle_t^{\frac{1}{2}}}, \quad (21)$$

$$J^{(\mathrm{BF})}(f) = \frac{\left|\left\langle Y_1^{(\mathrm{BF})}(f,t)Y_2^{(\mathrm{BF})}(f,t)^*\right\rangle_t\right|}{\left\langle\left|Y_1^{(\mathrm{BF})}(f,t)\right|^2\right\rangle_t^{\frac{1}{2}}\left\langle\left|Y_2^{(\mathrm{BF})}(f,t)\right|^2\right\rangle_t^{\frac{1}{2}}}, \quad (22)$$

where $Y_l^{(\mathrm{ICA})}(f,t)$ is the separated signal by ICA, and $Y_l^{(\mathrm{BF})}(f,t)$ is the separated signal by beamforming. If the separation performance of beamforming is superior to that of ICA, we obtain the condition, $J^{(\mathrm{ICA})}(f) > J^{(\mathrm{BF})}(f)$; otherwise $J^{(\mathrm{ICA})}(f) \leq J^{(\mathrm{BF})}(f)$. Thus, an observation of the conditions yields the following algorithm:

$$\boldsymbol{W}(f) = \begin{cases} \boldsymbol{W}_{i+1}^{(\mathrm{ICA})}(f), & \left(J^{(\mathrm{ICA})}(f) \leq J^{(\mathrm{BF})}(f)\right) \\ \boldsymbol{W}^{(\mathrm{BF})}(f), & \left(J^{(\mathrm{ICA})}(f) > J^{(\mathrm{BF})}(f)\right). \end{cases} \quad (23)$$

If the $(i+1)$th iteration was the final iteration, go to **step 6**; otherwise go beck to **step 2** and repeat the ICA iteration inserting the $\boldsymbol{W}(f)$ given by Eq. (23) into $\boldsymbol{W}_i(f)$ in Eq. (12) with an increment of i.

[Step 6: Ordering and scaling] Using the DOA information obtained in **step 3**, we detect and correct the source permutation and the gain inconsistency [8].

3.2. Subband elimination

Even in the proposed fast-convergence algorithm, there are some subbands in which the separation performances are not so well especially when the interference is diffusive noise. In order to resolve the problem, subband elimination (SBE) processing is introduced. The SBE can work so as to (1) find the specific subbands in which the bad separation was performed, and (2) eliminate them enforcedly (see Fig. 4).

In SBE, first, we calculate the smoothed cost function $J^{(\mathrm{s})}(f)$ as

$$J^{(\mathrm{s})}(f) = \frac{1}{f_\mathrm{m}}\int_{f-\frac{f_\mathrm{m}}{2}}^{f+\frac{f_\mathrm{m}}{2}} \min\left[J_{\mathrm{final}}^{(\mathrm{ICA})}(f'), J_{\mathrm{final}}^{(\mathrm{BF})}(f')\right] df', \quad (24)$$

where $J_{\mathrm{final}}^{(\mathrm{ICA})}(f)$ and $J_{\mathrm{final}}^{(\mathrm{BF})}(f)$ are the cosine distances obtained in the final step 5 described in Sect. 3.1. f_m is the frequency bandwidth for smoothing to decrease the discontinuity in the frequency characteristics of $J^{(\mathrm{s})}(f)$.

Next, we decide the reduction gain for each subband, $R(f)$, as

$$R(f) = \begin{cases} 1, & \left(J^{(\mathrm{s})}(f) \leq J_\mathrm{T}\right) \\ \epsilon, & \left(J^{(\mathrm{s})}(f) > J_\mathrm{T}\right) \end{cases}, \quad (25)$$

where J_T is the threshold for the decision of the elimination and ϵ is the small value less than 1. To use $R(f)$, we can finally obtain the separated signals as follows:

$$\hat{\boldsymbol{Y}}(f,t) = R(f)\boldsymbol{W}(f)\boldsymbol{X}(f,t). \quad (26)$$

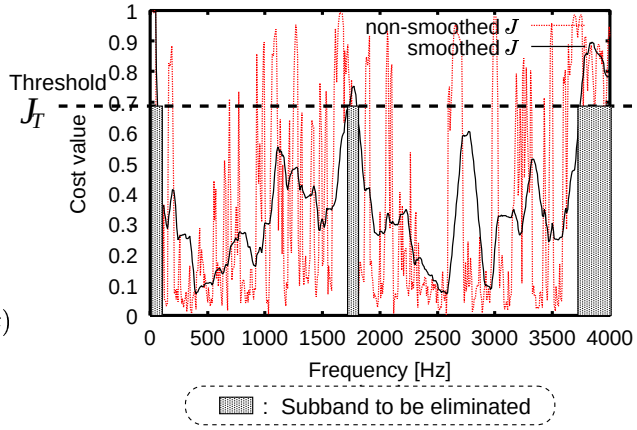

Fig. 4. Subband elimination procedure after ICA.

4. EXPERIMENTS IN CAR ENVIRONMENT

4.1. Conditions for experiments

A two-element array with the interelement spacing of 4 cm is used to record the sounds in a real car environment as shown in Fig. 5. The target signal is a driver's speech which arrives from $-54°$. As for the typical noise in car environment, we use the six kinds of noises as follows: (1) the speaker in the assistant seat which arrives from $58°$ **(assist)**, (2) engine noise **(eng)**, (3) road noise from the car tires at a speed of 30 km/h **(r30)**, (4) noise from air conditioner **(acd)**, (5) winker sound **(wnk)**, and (6) wiper sound **(wip)**. The analytical conditions of these experiments are as follows: the sampling frequency is 16 kHz, the frame length is 128 msec, the frame shift is 2 msec, the step-size parameter η is set to be 1.0×10^{-5}. In the SBE, J_T is set to be 0.7, and ϵ is 0.

4.2. Objective evaluation of separated signals

In order to evaluate the performance of the proposed algorithm, the *noise reduction rate* (NRR), defined as the output signal-to-noise ratio (SNR) in dB minus input SNR in dB, is shown in Fig. 6.

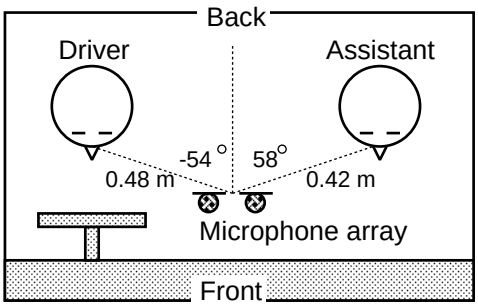

Fig. 5. Layout of array in car cabin used in experiment.

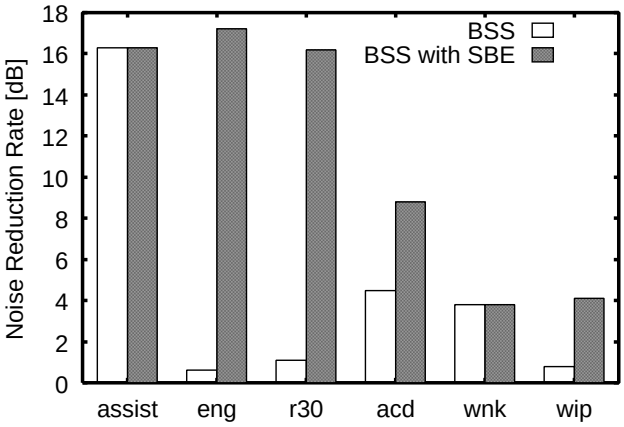

Fig. 6. Noise reduction rates for different noises in car environment.

Figure 6 shows NRR results of the proposed BSS without SBE (see the white bars), and those with SBE (see the gray bars). From this figure, it is evident that the separation performance of the BSS with/without SBE for assistant speech is superior to those for the other noises This is because the assistant speech is considered as the directional noise, and BSS can separate such kind of noise easily [5].

However, regarding the diffusive noise like the engine noise or the road noise, the separation performance of the BSS without SBE remarkably degrade. Figures 7–9 show the typical examples of cosine distance in the final iteration of ICA, $\min[J_{\text{final}}^{(\text{ICA})}(f), J_{\text{final}}^{(\text{BF})}(f)]$, for the assistant speech, the engine noise, and the road noise. As shown in these figures, the BSS without SBE can separate the sound sources in almost all frequency regions when the noise is the assistant speech. However, as for the engine noise and the road noise, the BSS without SBE can not separate the sources especially in the low-frequency region ($f < 200$ Hz), mid-frequency region ($1500 < f < 2000$ Hz), and high-frequency region (3500 Hz $< f$), compared with those in the case of the assistant speech.

On the other hand, the separation performance of the BSS with SBE can be improved even for the diffusive noise like the engine noise or the road noise. These results indicate that the performance of the simple BSS is insufficient in the car environment, however the combination with SBE is effective to improve the separated speech quality.

Regarding the air-conditioner noise, the BSS without SBE can reduce the noise to a certain extent, and the BSS with SBE can achieve a more better performance because this noise has both properties of directional and diffusive noises.

As for the winker and wiper sounds, the BSS with/without SBE cannot reduce the noises.

4.3. Speech Recognition Test

The HMM continuous speech recognition (CSR) experiment is performed in a speaker-independent (gender-dependent) manner, where we use the Japanese dictation system with *Julius* and typical models provided by IPA Japanese dictation toolkit [9]. For the CSR experiment, the PTM model is trained using clean sentences.

Figure 10 shows the results in terms of word accuracy under different noise conditions. In this figure, the white bars represent the speech recognition results for the observed signals at the microphone, the shaded bars represent the results by the BSS without SBE, and the gray bars represent the results by the BSS with SBE, respectively. These results indicate that the BSS with SBE is applicable to the speech recognition system, particularly when confronted with the assistant speech, engine noise, and road noise.

5. CONCLUSION

In this paper, we described a fast-convergence algorithm for BSS where null beamforming is used for temporal algorithm diversity through ICA iterations. Also we newly introduced the subband elimination technique based on the independence among the separated sources. The experiment in a real car environment reveals that the proposed method is remarkably effective to improve the qualities of the separated speech and word recognition rates for both directional and diffusive noises.

6. ACKNOWLEDGEMENT

This work was partly supported by NISSAN MOTOR CO., LTD. The authors are grateful to Mr. Masaru Yamazaki of NISSAN MOTOR CO., LTD. for his help on the measurement of car noises.

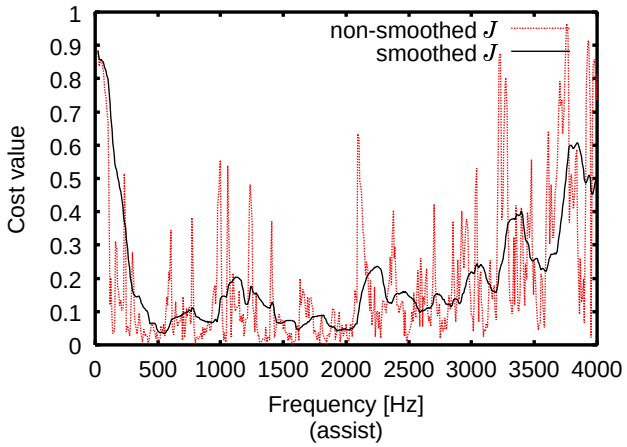

Fig. 7. Example of cosine distance in the final iteration of ICA for the assistant speech (dotted line), and its smoothed value $J^{(s)}(f)$ given in Eq. (24) (solid line).

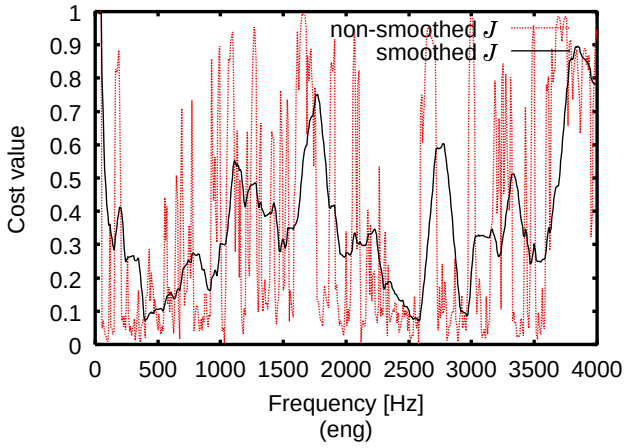

Fig. 8. Example of cosine distance in the final iteration of ICA for the engine noise (dotted line), and its smoothed value $J^{(s)}(f)$ given in Eq. (24) (solid line).

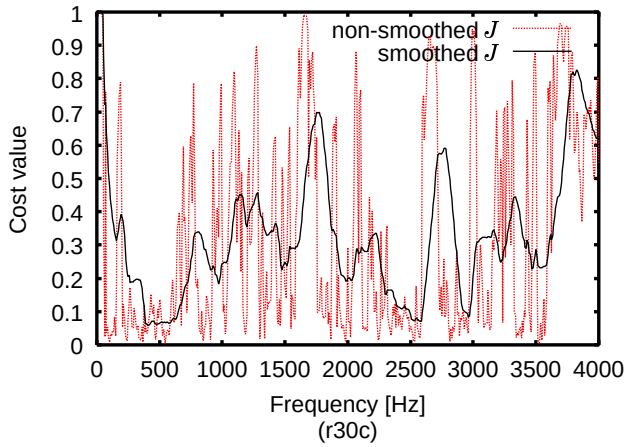

Fig. 9. Example of cosine distance in the final iteration of ICA for the road noise (dotted line), and its smoothed value $J^{(s)}(f)$ given in Eq. (24) (solid line).

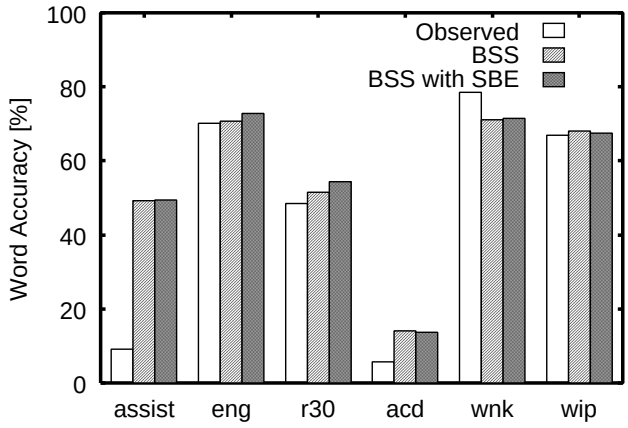

Fig. 10. Word accuracy for different noises in car environment.

7. REFERENCES

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

[2] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," Proc. NOLTA'98, vol.3, pp.923–926, 1998.

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

[4] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," IEEE Trans. Speech & Audio Process., vol.8, pp.320–327, 2000.

[5] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, and K. Shikano, "Blind source separation based on subband ICA and beamforming," Proc.ICSLP2000, vol.3, pp.94–97, 2000.

[6] A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. Circuits and Systems I*, vol.43, no.11, pp.894–906, 1996.

[7] H. Saruwatari, T. Kawamura, and K. Shikano "Blind source separation for speech based on fast-convergence algorithm with ICA and beamforming," Proc. EUROSPEECH2001, pp.2603–2606, 2001.

[8] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," Proc. ICASSP2000 vol.5, pp.3140–3143, 2000.

[9] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Ito, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Free software toolkit for Japanese large vocabulary continuous speech recognition," Proc.ICSLP2000, vol.4, pp.476–479, 2000.