RAPID FILTER ADAPTATION FOR FREQUENCY-DOMAIN INDEPENDENT COMPONENT ANALYSIS IN VARIOUS CAR ENVIRONMENTS

Atsunobu Kaminuma[†], *Daisuke Saitoh*[†]
Hiroshi Saruwatari[‡], *Tsuyoki Nishikawa*[‡], *and Akinobu Lee*[‡]

[†]Nissan Research Center, NISSAN MOTOR Co., LTD.
1, Natsushima-cho, Yokosuka-shi, Kanagawa 237-8523, Japan
[‡]Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara 630-0192, Japan

ABSTRACT

A computational complexity reduction method in a noise reduction algorithm using ICA [1] in car components is described. We examined a noise suppression system and a speech enhancement system that use frequency-domain independent component analysis (hereafter "FDICA") in car compartments with speech input systems [2, 3]. To achieve real-time processing in a car, we must reduce the computational complexity. We solved this problem with real-time processing by controlling the adaptive timing.

1. INTRODUCTION

Car compartments with speech input systems have been used for mobile phones and speech recognition systems. Because various noise sources exist inside car compartments, the system needs to eliminate noises in equipment that uses a speech input system. We have to consider two points, performance and compatibility, to create a successful car compartment system.

The ability to maintain the tone quality of mobile phones and that of improving the accuracy of speech recognition are related to performance. Customers require noise suppression while tone quality is maintained. In addition, an improvement in speech recognition accuracy resulting from better noise suppression is necessary for these systems. Therefore, we need good algorithms for both noise rejection and tone quality for real-time use.

Compatibility requires that special adjustments are not needed at the time of system importation of a unit bought from a supplier, a microphone, a speech recognition system, or a mobile phone. For instance, a nonlinear processing noise reduction method such as spectral subtraction (SS)[4] is often buried in a speech recognition system. In this case, a system that is compatible with the SS of other parts is needed. To solve these two problems, we examined a noise reduction system located in the front part of a mobile phone and a speech recognition system (Fig. 1).

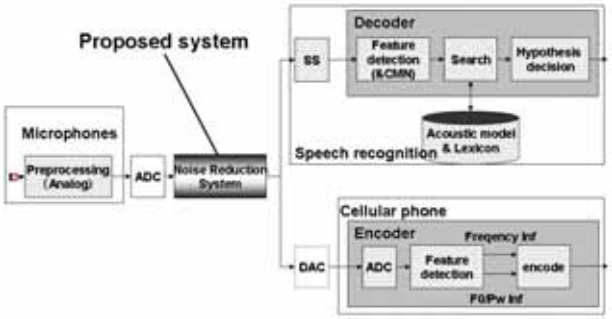

Fig. 1: **Speech input system in automobile**

In addition, frequency domain independent component analysis (FDICA) was applied to create an algorithm at the center of the noise reduction system. The performance of the frequency-domain ICA, which separates the sound source, is extremely high. Moreover, since it is a linear system, it has high compatibility with non-linear systems like those for spectral subtraction.

However, operating it in real time is difficult because of the enormous computational complexity. Therefore, we think that reducing computational complexity is one of the targets for operating frequency-domain ICA in real time. Consequently, we have been working on the following problems.

1. Removing calculation redundancy in the time domain
2. Removing calculation redundancy in the frequency domain
3. Optimizing parameters
4. Optimizing software

In this paper, we examine an algorithm that removes the calculation redundancy in the time domain.

In Section 2, we explain the FDICA algorithm used in our work.

In Section 3, we explain the procedure of the system for executing filter adaptation processing, which is done only

when the characteristics of the car compartment noise change: we call this a rough adaptive filter concept (RAF).

In Section 4, we examine the conditions of continuous noise that cause speech recognition rate changes.

In Section 5, we evaluate the speech recognition rate under the condition of executing filter adaptation for utterances and the condition of executing filter adaptation only when the characteristics of car compartment noise change, and compare the results.

2. FREQUENCY-DOMAIN ICA ALGORITHM

In this study, a straight-line microphone array was assumed. The number of microphones was K, and the number of multiple sound sources was L. In FDICA, a short-time analysis of the observed signals was first conducted in a frame-by-frame discrete Fourier transform (DFT). By plotting frame by frame, we determined the spectral values of each microphone input in a frequency bin, namely, sub-bands, as a time series. We designate the time series here as $\mathbf{X}(f,t)=[X_1(f,t), \ldots, X_K(f,t)]^T$.

Next, we performed signal separation using the complex-valued inverse of the mixing matrix $\mathbf{W}^{(f)}(f)$ so that the L time-series output $\mathbf{Y}^{(f)}(f,t)=[Y^{(f)}_1(f,t),\ldots, Y^{(f)}_L(f,t)]^T$ became mutually independent. This procedure can be given as

$$\mathbf{Y}^{(f)}(f,t)=\mathbf{W}^{(f)}(f)\mathbf{X}(f,t) \qquad (1)$$

We performed this procedure with respect to all sub-bands, f. Finally, by applying the inverse DFT and the overlap-add technique to the separated time series, $\mathbf{Y}^{(f)}(f,t)$, we reconstructed the resultant source signals in the time domain, $\mathbf{Y}^{(f)}(t)$.

In conventional FDICA, the optimal $\mathbf{W}^{(f)}(f)$ can be obtained using the following iterative equation.

$$\mathbf{W}^{(f)}_{i+1}(f)=\eta\Big[diag\Big(\big\langle \mathbf{\Phi}\big(\mathbf{Y}^{(f)}(f,t)\big)\mathbf{Y}^{(f)}(f,t)^H\big\rangle_t\Big) - \big\langle \mathbf{\Phi}\big(\mathbf{Y}^{(f)}(f,t)\big)\mathbf{Y}^{(f)}(f,t)^H\big\rangle_t\Big]\mathbf{W}^{(f)}_i(f)+\mathbf{W}^{(f)}_i(f), \qquad (2)$$

where $\langle\cdot\rangle$ denotes the averaging operator, i is used to express the value of the i-th step in the iterations, H is the Hermitian transpose, η is the step-size parameter, and $\mathbf{\Phi}$ is the nonlinear vector function such as a sigmoid function. Thus, the separation filter can calculate equations (1) and (2) repeatedly.

3. ROUGH ADAPTIVE FILTER CONCEPT (RAF)

In conventional FDICA, the system learns the sound source separation filter at the start, and separates the sound source afterwards using the adapted separation filter described in Fig. 1. In this system, the learning process is always begun immediately after the user inputs some speech command. To achieve real-time processing, both the additional time spent on learning processing and the time spent on filtering processing have to be performed in real time (Fig. 2).

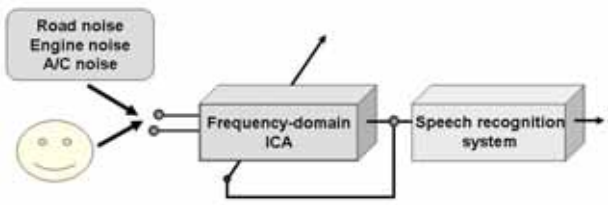

Fig. 2: **Block diagram of conventional method**

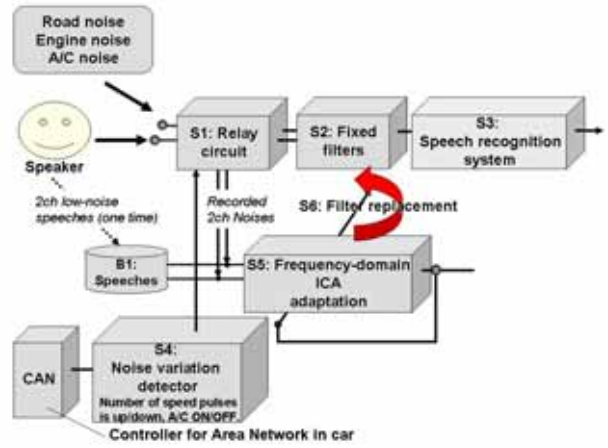

Fig. 3: **Block diagram of our method**

The noise separation capability of FDICA can achieve a speech gain recovery of about 5-15 dB with a two-microphone array system. However, the computing time spent on filter adjustment ranges from 10 seconds to a few minutes. Currently, most of the computing time is spent on processing for the filter adaptation, and little time is spent on the computing for filtering the speech signal. To make the system rapid, we thought that excluding the processing for the filter adaptation executed every time the user uttered something would be effective.

As one example, we showed a system that studied the sound source separation filter by background detection of the change in the noise from the controller area network (CAN) signal in Fig. 3.

First, the system records the noise when a change in the noise is detected with S4 (Fig. 3), and it adds to the clean speech signal maintained beforehand at B4 (Fig. 3).

Second, the filter adaptation to the background begins in S5. When the study ends, the system replaces sound source separation filter S2 with the newly adapted filter. Thus, if adapting the filter for each utterance is unnecessary, the filter can be adapted every time the noise changes, making processing in the background possible and enabling our target to be achieved. Therefore, we set up a hypothesis stating that changes in short-term noise in the car environment do not lead to changes in the

characteristics of the filter. We test this hypothesis in the next section.

4. CAR NOISE SELECTION

From the viewpoint of the time aspect and noise detection, the noise generated in the car compartment can be divided roughly into discrete noise and continuous noise. The state is changed into discrete noise by a switch operation such as for the air conditioner, wipers, and blinkers.

Constant-speed running noise and acceleration/deceleration running noise, on the other hand, exist as continuous noise. This can be used to predict changes in noise, for example, the car speed pulse.

In continuous noise, the load on the system in a running condition increases because the filter has to adapt to this whenever the noise changes. In this section, we describe our investigation of the performance variation in a speech recognition system on the market that is caused by the speed of a car, and we reveal our decision on the interval required for learning about continuous noise.

4.1. Experimental procedure and conditions

We investigated the speech recognition rate under a variety of running environments concerning car velocity and acceleration/deceleration. The work was performed according to the following procedures.

Step 1: Noise recording

Two microphones were set up in the car compartment (Fig. 4). The car was operating while repeating acceleration and deceleration during ordinary driving on a proving ground that assumed urban operation. In addition, vehicle interior noise N was collected intermittently in ten-second intervals. In addition, the level of the car speed and the degree of acceleration/deceleration were given as sound data. We preserved the data in a file. The sampling frequency was 48 kHz, and the quantization rate was 16 bits. After all the data was collected, we sampled it up to 11.025 kHz.

Step 2: The noise was plotted as based on speed and acceleration.

Step 3: Making speech test set

Three speech test sets for evaluating 300 utterances per set were made using noise N, which had been collected with the microphone nearest the driver of the 2-ch microphones.

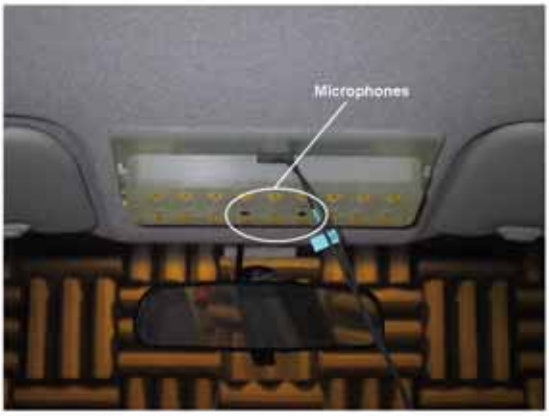

Fig. 4: **Arrangement of microphones**

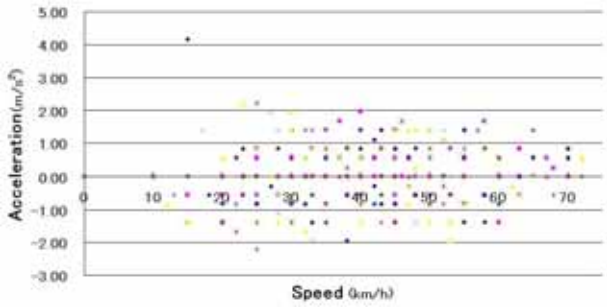

Fig. 5: **Classification of noise**

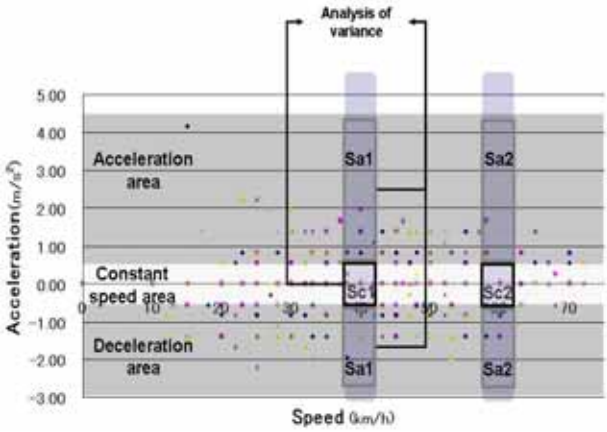

Fig. 6: **Decision on evaluated clusters**

The utterances that were collected in a soundproof chamber beforehand, as well as a transfer function, were used from the driver's utterance position to the microphones.

Three hundred utterances were selected at random from the utterances made by 17 men and 5 women.

Step 4: Calculation of speech recognition rate

The speech recognition rate was measured for the speech test sets of each noisy environment condition made in Step 3. The recognition decoder used VORERO Ver.4.3 [5], which has spectral subtraction. The system has a language dictionary that waits for 69 isolated words at the same time.

Step 5: The speech recognition rate evaluations of constant speed operation and acceleration operation were compared.

Step 6: The recognition accuracy was compared at different vehicular speeds.

4.2. Classification of noise and setting of experimental conditions

Using the car velocity and acceleration/deceleration degree (Fig. 5), 545 noise files were classified. The x-axis shows the velocity from 0 to 70 km/h, and the y-axis shows the acceleration from -3 to 5 m/s^2. Here, a negative value shows deceleration. It should be noted that most noise files were included within the range from -2 to 2 m/s^2.

Figure 6 shows the noise file used for the evaluation. We selected the acceleration range from -0.56 to 0.56 m/s^2 as the constant speed area, and chose another range (under -0.56 m/s^2 and over 0.56 m/s^2) as the acceleration area. In addition, we chose four speech recognition evaluation areas to investigate steps 5 and 6. We compared a 40 km/h constant speed area (Sc1: 22 speech test sets) and a 40 km/h acceleration area (Sa1: 23 speech test sets). Then, we compared a 60 km/h (Sc2: 13 speech test sets) constant speed area and a 60 km/h acceleration area (Sa2: 8 speech test sets) as experimental conditions, all for Step 5. Similarly, we compared a 40 km/h constant speed area (Sc1: 22 speech test sets) and a 60 km/h constant speed area (Sc2: 13 speech test sets) as the experimental conditions of Step 6.

4.3. Speech recognition rate

Figure 7 shows four recognition results in the speech recognition evaluation area. The vertical axis shows the speech recognition rate, and the four-bar chart shows the average speech recognition rate of three speech test-sets in each evaluation area. The beard that accompanies the bar chart shows the maximum and minimum values of the voice recognition rate of the object area.

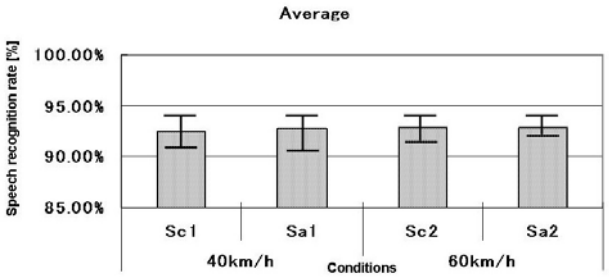

Fig. 7: **Speech recognition rate**

4.4 Comparison

4.4.1. Results of Step 5

We tested the hypotheses "The mean value of the recognition rate under a constant speed running condition (Sc1, Sc2) and the acceleration/deceleration running (Sa1, Sa2) were equivalent" by using a t-test. Expressions (3) and (4) are the results of calculating the p value by using the t-test of MS Excel.

$$p(Sc1, Sa1) = 0.13 > 0.05 \quad (3)$$

$$p(Sc2, Sa2) = 0.49 > 0.05 \quad (4)$$

In both cases, the hypothesis could not be dismissed according to the results of expressions (3) and (4). Therefore, we judged that the speech recognition rate under the constant speed running and the speech recognition rate under the acceleration/deceleration running are the same.

4.4.2. Results of Step 6

Similarly, we tested the hypothesis "The mean value of the recognition rate under the 40 km/h constant speed running condition (Sc1) and the 60 km/h constant speed running condition (Sc2) were equivalent" by using another t-test.

$$p(Sc1, Sc2) = 0.11 > 0.05 \quad (5)$$

In this case, the hypothesis also could not be dismissed according to the results of expression (5). Therefore, the speech recognition rate under the 40 km/h constant speed running and the speech recognition rate under the 60 km/h constant running speed were judged to be the same.

In addition, we double-checked this investigation for the speed of the car at several points. However, a significant difference was not detected between 20 and 60 km/h under an urban running environment.

4.5 Discussion

From the investigation in Section 4, we understood that the recognition performance is equal under an urban running environment in a car using a speech recognition

system with spectral subtraction. However, we estimate that similar results will not necessarily be obtained for an idling state and high-speed transit-time.

Therefore, we examined three conditions of idling, 60 km/h constant running speed, the top speed of urban driving under Japan-domestic regulations, and 100 km/h constant running speed, which is the top speed for expressways under Japan-domestic regulations.

5. DETERMINATION OF ADJUSTMENT BEGINNING TIME

We evaluated the speech recognition rate of the speech after FDICA processing using three techniques: a) a conventional method of processing filter adaptation at each utterance, b) another conventional technique involving processing filter adaptation only when either the speaker or noise changes, and c) our method of processing filter adaptation only when the car compartment noise changes. We then compared the results. In this experiment, nine environments for which the noise would change were assumed.

5.1. Experimental conditions

5.1.1. Making test-sets

Two microphones were set up in the car compartment (Fig. 4), Three kinds of interior vehicle noises N1 were collected under the conditions of idling and 60 and 100 km/h constant running speeds. Similarly, two interior vehicle noises N2 at air conditioner level two and air conditioner level four were collected in a semi-anechoic chamber. Three speech test-sets for the evaluation of 300 utterances a set were made by using noises N1 and N2; the utterances were collected in a sound-proof chamber beforehand, and the transfer function (from the driver's utterance position into two microphones). Approximately 300 utterances were selected from the utterance data of 17 men and 5 women at random. The microphone interval was 4 cm, the sampling frequency when collecting was 48 kHz, and the quantization bit rate was 16 bit. All data was down sampled to 11.025 kHz after it was all collected.

5.1.2. Sound source separation processing

We separated the sound source of the test set made as described in Section 5.1.1 by using the algorithm shown in Section 2. The filter length was 1024 points, the frame shift length was 256 points, the learning iterative computation was 200 times, and the sound source direction initial value was 0 and -60 degrees. In a), the technique for processing filter adaptation at each utterance, the system adapted to the filter 200 times at each utterance and separated the speech by using the generated filter.

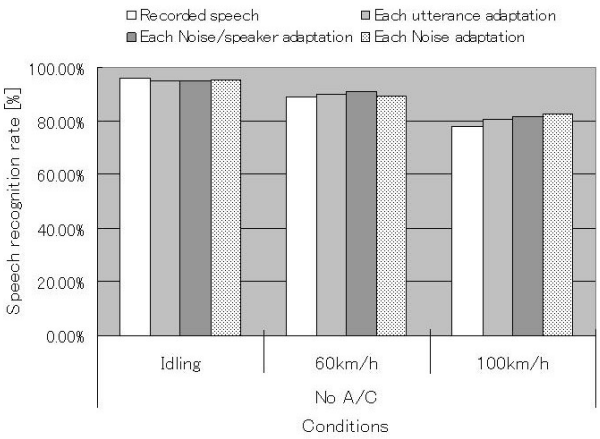

Fig. 8: **Speech recognition rate**

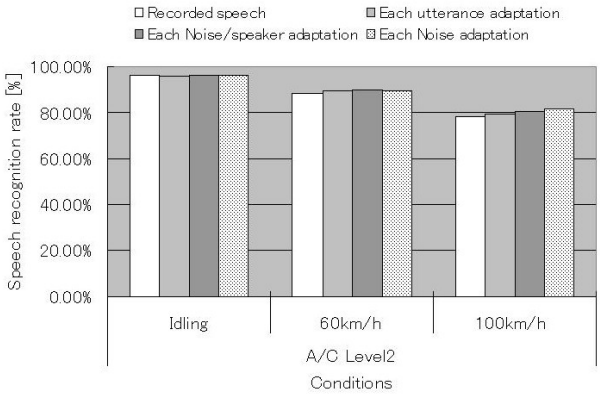

Fig. 9: **Speech recognition rate**

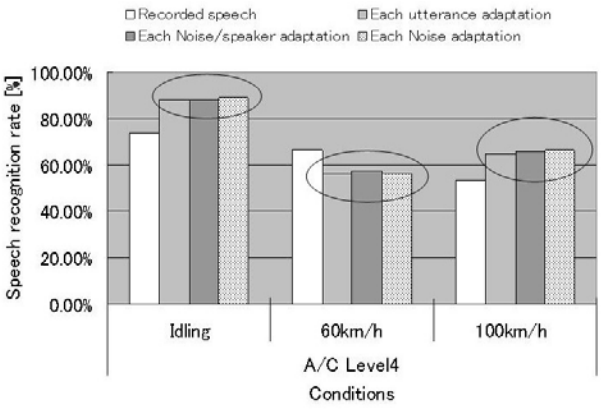

Fig. 10: **Speech recognition rate**

In b), the technique for processing filter adaptation only when either the speaker or noise changes, the system adapted to the filter 200 times only when the speaker or car compartment noise changed, and it separated the sound source by using the generated filter. In c), the technique for processing filter adaptation only when the car compartment noise changes, the system adapted to the

filter 200 times only when car compartment noise changed, and it separated the sound source of all speech test sets by using the generated filter.

5.1.3. Calculation of speech recognition rate

The recognition decoder used VORERO Ver. 4.3 [5] that had spectral subtraction. The language dictionary waited for the 69 isolated words at the same time. We calculated the speech recognition rate for three speech test sets in each noise condition.

5.2. Results

Figure 8 presents the evaluation results of the speech recognition rate at each velocity when the air conditioner was not used. The vertical axis shows the speech recognition rate, and the 3-bar chart shows the average speech recognition rate of three speech test sets in each velocity condition. The bar chart for each noise condition shows all three techniques, including our method.

In all cases, the speech recognition rate of the speech processed by FDICA showed hardly any improvement. We think that the engine noise and road noise do not necessarily come from a specific direction. The FDICA suppresses only the sound from a specific direction. Therefore, the effect of sound source separation is not sufficiently reflected in diffusive noise, such as that of the engine or road. In addition, there was little difference between the adaptation techniques.

Figure 9 presents the evaluation results of the speech recognition rate at each speed with air conditioner level two. The condition concerning the display is the same as in Fig. 8.

In the 60 and 100 km/h constant running speeds, the speech recognition rate after FDICA processing improved to about three points. We think that the noise coming from the air conditioner duct was suppressed by FDICA. In addition, we once again found little difference between the techniques.

Figure 10 presents the evaluation results of the speech recognition rate at each speed with air conditioner level four. The condition concerning the display is the same as in Fig. 8. In the idling and 100 km/h constant running speed conditions, the speech recognition rate after FDICA processing improved to about 12 pts from 15. We think that the noise coming from the air conditioner duct was suppressed by FDICA, as in Fig. 9. In the case in which the speech recognition rate was greater than in Fig. 9, we found the reason to be that the sound pressure level of the air conditioner noise was high, and the noise reduction rate by FDICA also increased. However, in the 60 km/h constant running speed condition, we found about a 10-pt decrease in the speech recognition rate after FDICA processing. A further investigation revealed this to be a filter divergence caused by over-learning.

In this condition, there was little difference between the techniques

6. DISCUSSION & CONCLUSION

Section 5.1 showed two findings. One is that the speech recognition rate can be improved by using FDICA when the direction of the speech is clear and noise of a high level is included. At this time, FDICA functioned extremely well under an environment where the engine noise and road noise were included as diffusive noise.

The other finding is that the performance equal to a conventional technique was obtained even when the filter was adapted when the noise changed in various car noise conditions. These results show that a performance equal to that of the conventional technique was obtained even when the adapted filters were switched according to changes in the noise condition. That is, when a driver gives a speech command, the speech only needs to be filtered by using fixed filters. Operation in real time is possible merely by filtering the speech signals. Consequently, we were able to clarify the real-time operation of the noise reduction system, one that uses frequency-domain ICA.

8. ACKNOWLEDGEMENTS

We thank Masaru Yamazaki and Hiroyuki Tateno at the Nissan Research Center for helping us with our experiments and computer simulations.

9. REFERENCES

[1] T. W. Lee, “Independent Component Analysis-Theory and Applications” Kluwer Academic Publishers, 1998.

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

[3] H. Saruwatari, K.Sawai, A. Lee, K. Shikano, A. Kaminuma, M. Sakata, “Speech enhancement and recognition in car environment using blind source separation and sub-band elimination processing,” Proc. ICA2003, pp. 367-372.

[4] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction” IEEE Trans. ASSP, Vol.27, No.2, pp. 113-120(1979).

[5] M. Shozakai, “Development for automatic speech recognition middleware VORERO for embedded applications” Proc. ASJ Spring 1-8-13, pp. 31-32, March 2004(in Japanese).
http://www.asahi-kasei.co.jp/vorero/