Speech Extraction in a Car Interior using Frequency-Domain ICA with Rapid Filter Adaptations

Daisuke Saitoh¹, Atsunobu Kaminuma¹, Hiroshi Saruwatari², Tsuyoki Nishikawa², Akinobu Lee³

¹Nissan Research Center, Nissan Motor Co., LTD
1 Natsushima-cho, Yokosuka-shi, Kanagawa 237-8523, Japan
{sai-dai, atunob-k}@mail.nissan.co.jp
²Nara Institute of Science and Technology
8916-5 Takayama-cho Ikoma-shi Nara 630-0192, Japan
{sawatari, tsuyo-ri}@is.naist.jp

Abstract

This paper describes two new algorithms for blind source separation (BSS) based on frequency-domain independent component analysis (FDICA). One is FDICA with preprocessing by a speech sub-band passing filter (SPF-FDICA). With this method, the signal-to-noise ratio (SNR) is slowed down, thereby reducing heavy reverberations or diffusive noises present that degrade separation performance seriously. A fast convergence algorithm [4] combines the null-beamforming technique with FDICA to derive the inverse of the mixing matrix with excellent convergence. FDICA with sub-band elimination post-processing [5] judges the separation performance for each sub-band and eliminates the parts with poor performance. However, these ICA-based approaches have the disadvantage that they involve enormous computational complexity. This is particularly true in the case of speech collision in an automobile, such as that involving the conversation of other passengers or a baby’s crying. An extremely large volume of calculations is needed to accomplish sufficient separation filtering. Thus, this approach is hardly practical for use in an automobile interior.

In this paper we propose two approaches to resolve this problem. One way is FDICA with preprocessing by a speech sub-band passing filter (SPF-FDICA). With this method, the learning speed (step size) in the specific sub-band where speech components are difficult to detect because of the low signal-to-noise ratio (SNR) is slowed down, thereby reducing the signal substitutions that are a major problem of FDICA. The other way is FDICA with speech sub-band selection (3S-FDICA). This approach reduces the number of iterations for low SNR sub-bands to provide an effect similar to that of the SPF and thereby accelerates the processing speed markedly. Each technique is focused on the relative difficulty of extracting the speech components in noisy speech signals.

This paper is organized as follows. In section 2, the background of conventional FDICA is explained. In sections 3 and 4, a system overview and the experimental results obtained with SPF-FDICA and 3S-FDICA are described respectively. Finally, our conclusions are presented in section 6.

1. Introduction

The interior of an automobile is a terribly noisy environment where both directional and diffusive noises are present. Independent component analysis (ICA) [1] is one effective method for extracting speech components in such an environment. Frequency-domain ICA (FDICA) is a technique for calculating the complex-valued inverse mixing matrix (separation filter) in the frequency domain [2, 3].

In recent years, some improved FDICA methods have been proposed for treating more realistic environments in which heavy reverberations or diffusive noises are present that degrade separation performance seriously. A fast convergence algorithm [4] combines the null-beamforming technique with FDICA to derive the inverse of the mixing matrix with excellent convergence. FDICA with sub-band elimination post-processing [5] judges the separation performance for each sub-band and eliminates the parts with poor performance.

However, these ICA-based approaches have the disadvantage that they involve enormous computational complexity. This is particularly true in the case of speech collision in an automobile, such as that involving the conversation of other passengers or a baby’s crying. An extremely large volume of calculations is needed to accomplish sufficient separation filtering. Thus, this approach is hardly practical for use in an automobile interior.

In this study, a straight-line microphone array is assumed. The number of microphones is K and the number of multiple sound sources is L. In FDICA, a short-time analysis of the observed signals is first conducted by frame-by-frame discrete Fourier transform (DFT). By plotting the spectral values of each microphone input in a frequency bin, namely sub-bands, frame by frame, we consider them as a time series. We designate the time series here as \(x(f, t) = [x_1(f, t), \ldots, x_k(f, t)]\). Next, we perform signal separation using the complex-valued inverse of the mixing matrix \(w^I(f)\), so that the L time-series output \(y^I(f, t) = [y_1^I(f, t), \ldots, y_L^I(f, t)]^T\) becomes mutually independent. This procedure can be given as

\[ y^I(f, t) = w^I(f) x(f, t) \tag{1} \]

We perform 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 \(y^I(f, t)\), we reconstruct the resultant source signals in the time domain, \(Y^I(i)\).

In conventional FDICA, the optimal \(w^I(f)\) is obtained by the following iterative equation,

\[ w^I_{n+1}(f) = \frac{1}{\sum_{i=1}^{I} w^I_i(f)} \left( \Phi(i) \left( \Phi(i) w^I_i(f) - \langle \Phi(i) \rangle \langle \Phi(i) w^I_i(f) \rangle \right) \right) \tag{2} \]

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, \(\eta\) is the step size parameter, and \(\Phi\) is a nonlinear vector function such as the sigmoid function. Thus, the separation filter can calculate equations (1) and (2) repeatedly.

---

September 4-8, Lisbon, Portugal
3. Proposed method 1: SPF-FDICA

Letting $S$ denote the speech components and $N$ the noise components observed in a car interior, the noisy speech signals can be divided between high-SNR and low-SNR sub-bands. In the high-SNR sub-bands, FDICA can provide a separation filter with high convergence and stability because the speech components can be detected easily. When an observed signal contains many low-SNR sub-bands, FDICA learns the separation filter, taking into account noise as erroneous speech. As a result, substitution of sound sources occurs frequently and the performance of the filter deteriorates drastically. This is referred to as the permutation problem. Though the sub-band elimination method [5] can improve this problem, the non-linear processing done by this method causes speech distortion. FDICA combined with pre-filtering by a speech sub-band passing filter (SPF-FDICA) is a new approach for resolving this problem by controlling the step size of each sub-band so as to slow down the learning speed in the low-SNR sub-bands.

3.1. Algorithm

A diagram of this approach is shown in figure 1-(a). In this approach, some specific sub-bands possessing no speech elements or rather small speech energy are considered as low-SNR sub-bands. Therefore, the SPF is designed based on this presupposition. In actuality, the SPF is designed as a band-pass filter that passes the high-SNR sub-bands and suppresses the low-SNR sub-bands. The filtering process done by the SPF is defined as

$$X'(f,t) = H(f)X(f,t)$$

(3)

where $H(f)$ is the SPF and $X'(f,t)$ is the filtered output of the input signal $X(f,t)$. Then, by substituting equation (3) into equation (1), we obtain the following expression,

$$Y(f,t) = W(f)X'(f,t)$$

(4)

The process of SPF-FDICA corresponds to iterative computations performed with equations (2) and (4). Focusing on $Y(f,t)$ in equations (2) and (4), it is observed that its power among the low-SNR sub-bands is smaller in the latter equation than in the former one because of the SPF. When this $Y(f,t)$ is applied to equation (2), the amount of change produced by the filter in the low-SNR sub-bands decreases. As a result, the learning step size can be controlled by the SPF, with the result that a small step size is applied to low-SNR sub-bands.

3.2. Evaluation and Discussion

Speech recognition experiments were conducted to verify the performance of this method, compared with conventional FDICA.

3.2.1. Speech data

The speech condition was defined as speech collisions in a car interior when the vehicle was traveling at 60 km/h. Car interior noises were recorded by a microphone array consisting of two microphones located over the rearview mirror and spaced at a 40-mm interval. Loudspeakers were arranged at the driver's and passenger's mouth position, and transfer functions between the microphones and the loudspeakers were measured. The microphones were at an angle of -63.2 deg. to the driver's mouth and at an angle of 56.7 deg to the passenger's mouth. The configuration of microphones and loudspeakers is shown in figure 2. In addition, 1,587 clean utterances by 17 men and 5 women (69 utterances per person) were recorded in a semi-anechoic room. One test set was made as follows: first, 300 utterances each were randomly chosen for the driver and for the passenger.

The transfer functions were then convoluted for each person. Utterances for the driver and for the passenger were added to make speech collisions. Finally, interior noise was added. The speech gain was adjusted to obtain SNR = 10 dB, corresponding to the interior noise condition under engine idling. All signals were digitized to 16 bits at a sampling frequency of 48 kHz and downsampled to 11.025 kHz.

3.2.2. SPF-FDICA conditions

SPF: The SPF was constructed with a normalized least mean square (NLMS) adaptive algorithm. The reference signal was human speech like noise (HSLN) [6], and the input signal was operating noise at 100 km/h combined with HSLN. Each signal was digitized to 16 bits at a sampling frequency of 48 kHz and downsampled to 11.025 kHz. The data length for learning was 10 sec and the filter length was 128 taps. Figure 3 shows the frequency response for the SPF. High response was detected near 1000 Hz, whereas low response was observed under 200 Hz and above 3000 Hz. These results indicate that the former sub-bands have a high SNR and the latter a low SNR.

FDICA: Iterative computations to obtain $W(f)$ were executed under 11.025-kHz sampling and 16-bit resolution. The filter length was 1024 taps. The data length for learning was 3 seconds, and 30, 50, 100 and 150 iterative computations were performed. The initial value of $W(f)$ was defined as a null-beam forming filter, and the nulls faced -60 and 0 deg. These iteration processes were executed only for the first data of a test set. This means that this FDICA system operates only when noise variation is detected.

Speech recognition: An isolated word speech recognition system was used. This system was capable of recognizing 69 keywords (all the same as the clean speech keywords). The recognition decoder used was VORERO ver. 4.3 [7].
3.2.3. Experimental results

Figure 4 shows the speech recognition performance for FDICA and SPF-FDICA. The horizontal axis shows the number of iterations, and the vertical axis shows the word recognition accuracy (WA). In 30 and 50 iterations, the WA of SPF-FDICA exceeded that of FDICA by 3 percentage points. However, in the case of 100 and 150 iterations, its WA was inferior to that of FDICA.

These results indicate that the SPF-FDICA method is more effective for fewer iterations such as 50. Under this situation, the permutation problem would be reduced because of the step size control effects of the SPF. It is assumed that this technique is effective in the case of small computational resources that do not allow sufficient learning iterations.

4. Proposed method 2: 3S-FDICA

We propose another improvement for FDICA by applying a method for reducing its computational complexity. To accomplish this, the speech sub-band selection (3S) technique is used instead of the SPF. A diagram of this approach is shown in figure 1-(b).

4.1. Algorithm

In this method, the 3S technique refers to the SPF’s frequency response shown in figure 3. Figure 5 shows the concept of this technique. In this figure, one grid means one learning calculation in one sub-band. The conventional FDICA (figure 5-(a)) performs the same number of iterations for all sub-bands. On the other hand, 3S-FDICA (figure 5-(b)) applies more iterations for sub-bands where the SPF has a high frequency response and reduces the number of iterations for those having a low frequency response. Assuming that the number of iterations in each sub-band $\omega$ is denoted as $\text{Iter}(\omega)$, the procedure of 3S is given by

$$
\text{Iter}(\omega) = R \left( \frac{\text{Iter}_{\max} - \text{Iter}_{\min}}{\text{step}} \right) \times \text{step} + \text{Iter}_{\min}
$$

where $R(\cdot)$ is the rounding function, $\max(\cdot)$ chooses the maximum value from all the elements in parentheses, $\text{res}(\omega)$ is the frequency response of the SPF for each sub-band, $\text{step}$ is the step size parameter of the iterations, $\text{res}_{\min}$ and $\text{res}_{\max}$, $\text{Iter}_{\min}$ and $\text{Iter}_{\max}$, $\omega_{\min}$ and $\omega_{\max}$ are maximum and minimum values of the frequency response, number of iterations, and selected sub-bands, respectively, and $\text{res}_{\max}$ specifies the available range of the SPF’s frequency response ($0 < \text{res}_{\max} \leq 1$). In this equation, $\text{Iter}_{\min}$, $\text{Iter}_{\max}$, $\omega_{\min}$, $\omega_{\max}$, and $\text{res}_{\max}$ are given by the user.

An example of sub-band selection is shown below.

$$
\text{Iter}(\omega) = \begin{cases} 
40 & \text{if } 100 \leq \omega \leq 5512 \\
70 & \text{if } 100 \leq \omega \leq 3000 \\
100 & \text{if } 110 \leq \omega \leq 2000 
\end{cases}
$$

These results were obtained under these conditions: $\text{Iter}_{\min} = 40$, $\text{Iter}_{\max} = 100$, $\omega_{\min} = 100$ Hz, $\omega_{\max} = 5512$ Hz, $\text{res}_{\max} = 0.4$.

4.2. Evaluation and Discussion

The speech recognition experiment is held to verify the performance of this method, comparing with conventional FDICA.

4.2.1. Conditions

In this experiment, the 0 km/h running situations were supposed. And 5 test sets were made for each conditions by the same way mentioned in section 3.
Two parameter settings for 3S are applied as follows
3S-(a): \( \text{iter}_{\text{max}} = 30, 50, 100, 150, \text{iter} = \text{iter}_{\text{max}} \times 0.4 \),
\( \omega_{\text{ref}} = 5512 \text{ Hz}, \omega_{\text{ref}} = 100 \text{ Hz}, \)
\( \text{res}_{\text{ref}} = 0.3 \text{ and step} = \text{iter}_{\text{max}} - \text{iter}_{\text{min}}. \)
3S-(b): \( \text{iter}_{\text{max}} = 30, 50, 100, 150, \text{iter} = \text{iter}_{\text{max}} \times 0.5 \),
\( \omega_{\text{ref}} = 5512 \text{ Hz}, \omega_{\text{ref}} = 100 \text{ Hz}, \)
\( \text{res}_{\text{ref}} = 0.3 \text{ and step} = \text{iter}_{\text{max}} - \text{iter}_{\text{min}}. \)

4.2.2 Experimental result
Figure 6 shows a WA comparison for FDICA, FDICA with 3S-(a) (3S-FDICA-(a)) and FDICA with 3S-(b) (3S-FDICA-(b)). The number of iterations is shown along the vertical axis, and WA is shown along the horizontal axis. The average WA of the five test sets in each iteration is plotted and maximum and minimum values were plotted overlaying the average WA as the vertical lines. Without FDICA, the WA under the 0 km/h condition was 25.4%. The WA obtained with 3S-FDICA-(b) exceeded that of FDICA by 2~7 percentage points in iterations below 100. Although not shown here, the same tendency was also observed at 60 km/h. These results suggest that a more stable separation filter can be obtained with this method. By reducing the learning of specific sub-bands, the 3S technique achieves the same effect as the SPF.

On the other hand, the 3S-FDICA-(a) method showed an improvement of 4 percentage points only for 30 iterations, and WA deteriorated for 50, 100 and 150 iterations. It is presumed that permutation errors occurred in low-SNR sub-bands because 3S-FDICA-(a) performed more iterations for these sub-bands than 3S-FDICA-(b) did.

The WA differences between FDICA and 3S-FDICA-(b) decreased as the number of iterations increased and reversed for 150 iterations. It is evident that this technique cannot reach the final performance of FDICA (after an enormous number of iterations) because of its smaller number of learning iterations. However, it is effective in improving performance under a condition of fewer iterations.

The results in figure 6 are presented again in figure 7 in terms of the effect of filter adaptation on increasing the computational speed. The figure shows the relationship between WA and the time consumed for each learning iteration. For easy comparison, the FOICA results are plotted for every 10 iterations from 10 to 150 iterations, while those between WA and the time consumed for each learning iteration. Looking at the 54% WA level, it is seen that FDICA required 18.3 sec to accomplish that WA level, whereas the 3S-FDICA-(b) method needed only 7.4 sec. The second method not only improved WA, it also speeded up the computational process.

5. Conclusion
This paper has proposed SPF-FDICA and 3S-FDICA methods both of which are focused on the difficulty of recognizing speech components at each frequency. The results of speech recognition experiments conducted under a speech collision condition showed that both methods improved word recognition accuracy for a small number of iterations. Furthermore, the second method not only improved WA, it also speeded up the computational process.

We aim to apply this technique to automobiles because it is well suited to their processing systems which have much less computational power than PCs.

6. Acknowledgements
The authors would like to thank Masaru Yamazaki and Hiroyuki Tateno at the Nissan Research Center for running our signal processing and speech recognition experiments.

7. References
http://www.asahi-kasei.co.jp/voreo