Speech Enhancement Based on Blind Source Separation in Car Environments

Hiroshi Saruwatari, Katsuyuki Sawai, Tsuyoki Nishikawa, Akinobu Lee, Kiyohiro Shikano
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan
sawatari@is.naist.jp

Atsunobu Kaminuma, Masao Sakata, Daisuke Saitoh
NISSAN MOTOR CO., LTD.
Natsushima-cho, Yokosuka-shi, Kanagawa 237-8523, Japan

Abstract

This paper describes a new speech enhancement system used in car environments, where independent component analysis (ICA)-based blind source separation (BSS) and subband elimination (SBE) processing are combined to achieve a high separation performance. The proposed method consists of two parts, fast-convergence frequency domain ICA and SBE processing based on the independence among the separated subband signals. The 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 diffuse 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. In recent works of BSS based on independent component analysis (ICA) [1], various methods have been proposed for acoustic-sound separation [2, 3, 4, 5]. This technique is applicable to the realization of noise-robust speech recognition and high-quality hands-free telecommunication systems. The one promising application of the BSS is a navigation system in a car environment, where many kinds of noises, e.g., engine noise and air-conditioner noise must be reduced. Some of these noises are classified into directional noise which is easily eliminated [6], but the others are diffuse noise which are difficult to be dealt with in the traditional BSS framework with small microphone arrays. To our knowledge, however, the feasibility of the BSS utilized in car environments has been reported in the limited number of papers [7], and the approach to address the diffuse noise reduction has not been presented.

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 $\theta_l$ ($l = 1, \cdots, L$) (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:

$$X(f) = A(f)S(f),$$

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

$$X(f) = [X_1(f), \cdots, X_K(f)]^T,$$

$$S(f) = [S_1(f), \cdots, S_L(f)]^T,$$

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

\[ X(f, t) = [X_1(f, t), \ldots, X_K(f, t)]^T. \] (5)

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

\[ Y(f, t) = W(f) X(f, t), \] (6)

where

\[ Y(f, t) = [Y_1(f, t), \ldots, Y_L(f, t)]^T, \]

\[ W(f) = \begin{bmatrix} W_{11}(f) & \cdots & W_{1K}(f) \\ \vdots & \ddots & \vdots \\ W_{L1}(f) & \cdots & W_{LK}(f) \end{bmatrix}. \] (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 \( Y(f, t) \), we reconstruct the resultant source signals in the time domain.

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

\[ W_{i+1}(f) = \eta \left[ \text{diag} \left( \langle \Phi(Y(f, t)) Y^H(f, t) \rangle_t \right) + \langle \Phi(Y(f, t))^H(f, t) \rangle_t \right] W_i(f) + W_i(f), \] (9)

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 \( \eta \) is the step-size parameter. Also, we define the nonlinear vector function \( \Phi(\cdot) \) as

\[ \Phi(Y(f, t)) = [\Phi(Y_1(f, t)), \ldots, \Phi(Y_L(f, t))]^T, \] (10)

\[ \Phi(Y(f, t)) = [1 + \exp(-Y_{i(R)}(f, t))]^{-1} + j \cdot [1 + \exp(-Y_{i(I)}(f, t))]^{-1}, \] (11)

where \( Y_{i(R)}(f, t) \) and \( Y_{i(I)}(f, t) \) are the real and imaginary parts of \( Y_i(f, t) \), respectively.

3. Proposed algorithm

3.1. Fast-convergence algorithm [8]

The conventional ICA method inherently has a significant disadvantage which is due to low convergence through nonlinear optimization in ICA [5]. In order to resolve the problem, we propose an algorithm [8] based on the temporal alternation of learning between ICA and beamforming; the inverse of the mixing matrix, \( W(f) \), obtained through ICA is temporarily 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 \( W_i(f) \), i.e., \( W_0(f) \), to an arbitrary value, where the subscripts \( i \) is set to be 0.

[Step 2: 1-time ICA iteration] Optimize \( W_i(f) \) using the following 1-time ICA iteration:

\[ W_{i+1}(f) = \eta \left[ \text{diag} \left( \langle \Phi(Y(f, t)) Y^H(f, t) \rangle_t \right) + \langle \Phi(Y(f, t))^H(f, t) \rangle_t \right] W_i(f) + W_i(f), \] (12)

where the superscript "(ICA)" is used to express that the inverse of the mixing matrix is obtained by ICA.
Figure 3. Proposed algorithm combining frequency-domain ICA and beamforming with subband elimination.

[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

\[
F_l(f, \theta) = \sum_{k=1}^{K} W^{(ICA)}_{lk}(f) \exp[j2\pi f d_k \sin \theta/c],
\]

where \( W^{(ICA)}_{lk}(f) \) is the element of \( W^{(ICA)}(f) \). 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

\[
\hat{\theta}_l = 2 \sum_{m=1}^{N/2} \theta_l(m_f)/N,
\]

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

\[
\begin{align*}
\theta_1(f_m) & = \min_{\theta} \left[ F_1(f_m, \theta), F_2(f_m, \theta) \right], \\
\theta_2(f_m) & = \max_{\theta} \left[ F_1(f_m, \theta), F_2(f_m, \theta) \right],
\end{align*}
\]

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, \( W^{(BF)}(f) \), 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

\[
W^{(BF)}_{11}(f_m) = \frac{1}{W^{(BF)}_{21}(f_m)} = -\exp\left[-j2\pi f_m d_1 \sin \hat{\theta}_2/c\right] \\
\times \left\{ -\exp\left[j2\pi f_m d_1 (\sin \hat{\theta}_1 - \sin \hat{\theta}_2)/c\right] + \exp\left[j2\pi f_m d_2 (\sin \hat{\theta}_1 - \sin \hat{\theta}_2)/c\right] \right\}^{-1},
\]

\[
W^{(BF)}_{12}(f_m) = \frac{1}{W^{(BF)}_{22}(f_m)} = \exp\left[-j2\pi f_m d_2 \sin \hat{\theta}_2/c\right] \\
\times \left\{ -\exp\left[j2\pi f_m d_1 (\sin \hat{\theta}_1 - \sin \hat{\theta}_2)/c\right] + \exp\left[j2\pi f_m d_1 (\sin \hat{\theta}_1 - \sin \hat{\theta}_2)/c\right] \right\}^{-1}.
\]

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

\[
W^{(BF)}_{21}(f_m) = \exp\left[-j2\pi f_m d_1 \sin \hat{\theta}_1/c\right] \\
\times \left\{ -\exp\left[j2\pi f_m d_1 (\sin \hat{\theta}_2 - \sin \hat{\theta}_1)/c\right] + \exp\left[j2\pi f_m d_2 (\sin \hat{\theta}_2 - \sin \hat{\theta}_1)/c\right] \right\}^{-1},
\]

\[
W^{(BF)}_{22}(f_m) = \exp\left[-j2\pi f_m d_2 \sin \hat{\theta}_1/c\right] \\
\times \left\{ -\exp\left[j2\pi f_m d_1 (\sin \hat{\theta}_2 - \sin \hat{\theta}_1)/c\right] + \exp\left[j2\pi f_m d_2 (\sin \hat{\theta}_2 - \sin \hat{\theta}_1)/c\right] \right\}^{-1}.
\]

[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

\[
\begin{align*}
J^{(ICA)}(f) & = \frac{\left| \langle Y_1^{(ICA)}(f, t) Y_2^{(ICA)}(f, t)^* \rangle_t \right|^2}{\left( \langle Y_1^{(ICA)}(f, t)^2 \rangle_t \right)^{1/2} \left( \langle Y_2^{(ICA)}(f, t)^2 \rangle_t \right)^{1/2}}, \\
J^{(BF)}(f) & = \frac{\left| \langle Y_1^{(BF)}(f, t) Y_2^{(BF)}(f, t)^* \rangle_t \right|^2}{\left( \langle Y_1^{(BF)}(f, t)^2 \rangle_t \right)^{1/2} \left( \langle Y_2^{(BF)}(f, t)^2 \rangle_t \right)^{1/2}}.
\end{align*}
\]
where \( Y_l^{(ICA)}(f, t) \) is the separated signal by ICA, and \( Y_l^{(BF)}(f, t) \) is the separated signal by beamforming. If the separation performance of beamforming is superior to that of ICA, we obtain the condition, \( J^{(ICA)}(f) > J^{(BF)}(f) \); otherwise \( J^{(ICA)}(f) \leq J^{(BF)}(f) \). Thus, an observation of the conditions yields the following algorithm:

\[
W(f) = \begin{cases} 
W_{i+1}^{(ICA)}(f), & \text{if } J^{(ICA)}(f) \leq J^{(BF)}(f) \\
W_{i+1}^{(BF)}(f), & \text{if } J^{(ICA)}(f) > J^{(BF)}(f) 
\end{cases} 
\]  

(23)

If the \((i+1)\)th iteration was the final iteration, go to step 6; otherwise go back to step 2 and repeat the ICA iteration inserting the \( W(f) \) given by Eq. (23) into \( W_i(f) \) 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 [5].

Figure 4 shows a good example of the behavior on the SNR improvement regarding the proposed method’s convergence under a reverberant room condition [8]. This figure contains the following three curves.

**Proposed Method** : Our proposed BSS method described in Section 3.1.

**Conventional ICA** : The conventional ICA-based BSS method described in Section 2. This also corresponds to the special case that ICA is always chosen in step 5 of the proposed algorithm.

**Null Beamformer** : The iteratively optimized null beamformer which corresponds to the special case that the null beamformer is always chosen in step 5 of the proposed algorithm.

From Fig. 4, it is evident that the proposed algorithm can show a rapid convergence, and the separation performances of the proposed algorithm are superior to those of the conventional ICA-based BSS method at every iteration point.

Figure 5 shows the example of alternation results between ICA and null beamforming through iterative optimization by the proposed algorithm. The symbol, black box, indicates that the null beamforming is used at the iteration point and frequency bin.

- The unmixing matrix obtained by ICA is substituted by the matrix based on null beamforming through all iteration points at particular frequency bins where the independence between the sources is low.

From these results, although null beamforming is not suitable for signal separation under the condition that direct sounds and their reflections exist, we can confirm that the temporal utilization of null beamforming for algorithm diversity through ICA iterations is effective for improving the separation performance and convergence.
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 diffuse 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. 6).

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

\[
J^{(s)}(f) = \frac{1}{f_m} \int_{f-f_m}^{f+f_m} \min \left[ J^{(ICA)}_{\text{final}}(f'), J^{(BF)}_{\text{final}}(f') \right] df',
\]

(24)

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

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

\[
R(f) = \begin{cases} 
1, & \text{if } J^{(s)}(f) \leq J_T \\
\varepsilon, & \text{if } J^{(s)}(f) > J_T 
\end{cases}
\]

(25)

where \( J_T \) is the threshold for the decision of the elimination and \( \varepsilon \) is the small value less than 1. By using \( R(f) \), we can finally obtain the separated signals as follows:

\[
\hat{Y}(f, t) = R(f) W(f) X(f, t).
\]

4. Noise reduction 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. 7. The target signal is a driver’s speech which arrives from the left-hand side of the array. As for the typical noise in car environment, we use six kinds of noises as follows: (1) the speaker in the assistant seat which arrives from the right-hand side of the array (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 \( \eta \) is set to be \( 1.0 \times 10^{-5} \). In the SBE, \( J_T \) is set to be 0.7, and \( \varepsilon \) is 0.

4.2. Objective evaluation of separated signals

In order to evaluate the performance of the proposed algorithm, the noise reduction rate (NRR) [5], defined as the output signal-to-noise ratio (SNR) in dB minus input SNR in dB, is shown hereinafter.

Figure 8 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 also the directional noise, and BSS can separate such kind of noise easily [6].

However, regarding the diffuse noise like the engine noise or the road noise, the separation performance of the BSS without SBE remarkably degrade. Figures 9–11 show the typical examples of cosine distance in the final iteration of ICA, \( \min[J^{(ICA)}_{\text{final}}(f), J^{(BF)}_{\text{final}}(f)] \), 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 \( (f < 200 \text{ Hz}) \), mid-frequency region \( (1500 < f < 2000 \text{ Hz}) \), and high-frequency region \( (3500 \text{ Hz} < f) \), 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 diffuse 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, but the combination with SBE is beneficial to improve the separated speech quality.
5. Speech recognition in car environment

5.1. Conditions for experiments

The aim of this section is to provide an experimental evaluation of applicabilities of BSS in speech recognition. Table 1 shows the experimental conditions for speech recognition. 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. 7. The target signal is set to a driver's speech, and the interference noise to be reduced is (a) assistant speech or (b) air-conditioner noise. As for the background noise, we consider the following situations where several diffuse noises are added: (1) engine noise at idle (idling), (2) engine noise and road noise from the car tires at a speed of 60 km/h with the car window close (60km/h(C)), (3) engine noise and road noise from the car tires at a speed of 60 km/h with the car window open (60km/h(O)), and (4) engine noise and road noise from the car tires at a speed of 100 km/h with the car window close (100km/h(C)).

We introduce an extended ICA method, Multistage ICA...
<table>
<thead>
<tr>
<th>Task</th>
<th>69 isolated words recognition with network grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic model</td>
<td>diphone HMM by single Gaussian mixture (speaker-independent)</td>
</tr>
<tr>
<td>Number of testing speakers</td>
<td>23 speakers (69 sentences / 1 speaker)</td>
</tr>
<tr>
<td></td>
<td>17 males and 5 females</td>
</tr>
<tr>
<td>Decoder</td>
<td>VORERO Ver.4.3 [9]</td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>11 kHz</td>
</tr>
<tr>
<td>Processing for noise-robust</td>
<td>(1) continuous spectral subtraction [10]</td>
</tr>
<tr>
<td></td>
<td>(2) normalized least mean squares error</td>
</tr>
<tr>
<td></td>
<td>(3) exact cepstrum mean normalization [10]</td>
</tr>
</tbody>
</table>

**Table 1. Experimental conditions for speech recognition.**

(MSICA) [12], proposed by one of the authors as well as FDICA. MSICA consists of the FDICA part described in Sect. 3 and time-domain ICA (TDICA), and is conducted with the following steps (see Fig. 12). In the first stage, we perform FDICA to separate the source signals to some extent with the high-stability advantages. In the second stage, we regard the separated signals from FDICA as the input signals for TDICA, and we can remove the residual crosstalk components of FDICA by using TDICA. Finally, we regard the output signals from TDICA as the resultant separated signals.

### 5.2. Experimental results

Figures 13(a)–(c) show the results in terms of word accuracy under different noise conditions. In these figures, the black bars represent the speech recognition results for the observed signals at the single microphone, the gray bars represent the results by the proposed FDICA given in Sect. 3, and the white bars represent the results by MSICA, respectively.

The remarkable improvements of word accuracy can be found in Fig. 13 in both FDICA and MSICA compared with the results using a single microphone. Regarding the reduction of the assistant speech, we can confirm an MSICA's slight outperformance from FDICA in all situations for the background noise. As for the reduction of the air-conditioner noise, there are no obvious improvements in MSICA compared with FDICA, but also no deterioration; this means that MSICA has no serious side-effects. Figure 13(c) shows the results of the assistant speech reduction in the case that a defroster noise is further added into the background noises. We can see the same tendency as in Fig. 13(a).

In summary, these results indicate that the proposed FDICA and MSICA are applicable to the speech recognition system, particularly when the small number of microphones are used. For further improvement, we have already developed another combination technique that utilizes a speech sub-band passing filter (see [13] for more details).
markably effective to improve the qualities of the separated speech and word recognition rates for both directional and diffuse noises.

References


6. Conclusion

In this paper, we described an application of BSS for speech enhancement in car environments. First, fast-convergence BSS algorithm which combines null beamforming and ICA was provided for temporal algorithm diversity through ICA iterations. Also we 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 diffuse noises.