Hands-Free Speech Recognition Using Blind Source Separation
Post-Processed by Two-Stage Spectral Subtraction

Masanori Tsujikawa, Ken-ichi Iso

Media and Information Research Laboratories, NEC Corporation, Japan
tujikawa@cb.jp.nec.com, k-iso@bx.jp.nec.com

Abstract
This paper proposes hands-free speech recognition using blind source separation (BSS) post-processed by two-stage spectral subtraction (2S-SS). The BSS using independent component analysis (ICA) estimates a target signal and jammer signals. The 2S-SS removes its residual crosstalk components and suppresses spatially-distributed noise not separated by BSS. In large vocabulary continuous speech recognition (LVCSR) evaluation, utterances by other speakers and computer-room noise were used as a jammer signal and a spatially-distributed noise source, respectively. In all noisy environments, it was confirmed that the proposed method outperformed the BSS with single-channel spectral subtraction (1SS).

1. Introduction

Hands-free speech recognition is an essential technique in man-machine communication. As a method that makes the hands-free speech recognition robust to noise, blind source separation (BSS) has been studied [1]-[8]. However, the BSS-processed signals include residual components of interference signals and spatially-distributed noise, which degrade speech-recognition accuracy.

The problem caused by the residual components of interference signals has been addressed in [1]. Figure 1 shows the structure of the method in [1]. One of the separated signals by BSS, $y_2$, is input into an adaptive digital filter (ADF). The filter coefficients are updated so that the square error $e^2$ between the ADF output $u$ and the other separated signal $y_1$ is minimized. Since the coefficient update is performed while the interference signal is large, its residual components can be eliminated. However, the spatially-distributed noise has not been taken into account.

A method considering both the residual components of interference signals and the spatially-distributed noise has been proposed in [2]. The structure of the method in [2] is depicted in Fig. 2. The mixed signal $x_1$ and the separated signals $y_1, y_2$ are input into desired speaker activity detection (SAD). The SAD calculates a probability measure $\lambda$ for the desired-speaker presence. Then the Wiener filter (WF) estimates the noise with averaging $(1 - \lambda)y_1$. Because $(1 - \lambda) \approx 0$ during the target-speech period, it is difficult to suppress non-stationary noise corrupting the target speech.

This paper proposes hands-free speech recognition using blind source separation post-processed by 2-stage spectral subtraction(2S-SS[9]). The 2S-SS improves separation accuracy of BSS and eliminates spatially-distributed noise. It also performs robust speech detection with speech and noise spectra that are obtained from the separated signals by BSS. The detected-speech spectra are conveyed to a recognition process. In the next section, the proposed method is explained in details. Section 3 presents LVCSR-evaluation results.

2. Proposed system

Figure 3 shows the structure of the proposed method, which mainly consists of the three modules, BSS, Noise estimation, and 2S-SS. The BSS separates target speech and an interference signal. In the Noise estimation module, the separated signals are used for estimating noise components contaminating the target speech. The estimated noise and the separated speech are input into the 2S-SS module. It removes the estimated noise from the target speech with noise and also performs robust target-speech detection. The detected-speech spectra are output to speech recognition. The details of each module are explained in the following.
2.1. BSS

BSS separates target speech and interference signals. There are two types of BSS, namely the time-domain BSS and the frequency-domain BSS. In this paper, the frequency-domain BSS is adopted. It estimates $P$ number of unmixed components from $Q$($Q \geq P$) convolutively-mixed signals, which are observed at $Q$ sets of microphone. The estimation is performed at each frequency bin. We assume that $P = Q = 2$ for simplicity of presentation.

Equation (1) describes the mixing process of source signals in the frequency domain.

$$X(f,t) = H(f)S(f,t) + N(f,t),$$

where $X(f,t) = [X_1(f,t) \ X_2(f,t)]^T$ is the mixed-signal vector. The element $X_q(f,t)$ is the $N$-point discrete Fourier transform (DFT) of the mixed signal from the microphone $q$, $f = 0, \cdots, N/2$ and $t = 0, \cdots$ are the integers representing the frequency bin and time frame, respectively. The operation $^T$ transposes matrices and vectors. $H(f)$ is the $2 \times 2$ mixing matrix. The element $H_{pq}(f)$ of the mixing matrix represents the frequency response between the source $p$ and the microphone $q$. $S(f,t) = [S_1(f,t) \ S_2(f,t)]^T$ is the source-signal vector. $N(f,t) = [N_1(f,t) \ N_2(f,t)]^T$ is the spatially-distributed noise vector. The element $S_p(f,t)$ and the element $N_q(f,t)$ are the frequency spectrum of the source $p$ and that of the distributed noise at the microphone $q$, respectively.

The separating process of the mixed signals in the frequency domain is as follows.

$$Y(f,t) = W(f)X(f,t),$$

where $Y(f,t) = [Y_1(f,t) \ Y_2(f,t)]^T$ is the separated-signal vector. $W(f)$ is the $2 \times 2$ separating matrix for the mixing matrix $H(f)$. The matrix $W(f)$ is estimated using the method [3], which is based on independent component analysis (ICA). The ambiguity of amplitude in ICA is removed by the method in [4]. The permutation problem in the frequency-domain BSS based on ICA is solved by the directivity-based initializing method in [8], which improves on [5] in the low-frequency band.

The BSS outputs are spectral amplitudes of the separated signals. They are approximated as follows.

$$|Y_1(f,t)| \approx |S_1(f,t)| + \epsilon_1(f)|S_2(f,t)| + R(f,t)|N'(f,t)|,$$

$$|Y_2(f,t)| \approx |S_2(f,t)| + \epsilon_2(f)|S_1(f,t)| + |N'(f,t)|,$$

where $\epsilon_q(f)(<1.0)$ is the coefficient representing a residual measure of each source signal. $N'(f,t)$ denotes the distributed-noise component. $R(f,t)$ is the ratio of the distributed noise included in $|Y_1(f,t)|$ to that in $|Y_2(f,t)|$. The residual and distributed-noise components may degrade speech-recognition accuracy and should be removed.

2.2. Noise estimation

The noise component in the target speech with noise is estimated for the 2S-SS procedure. Assuming that target speech is $S_1(f,t)$, the noise included in $|Y_1(f,t)|$ is the sum of the residual component $\epsilon_1(f)|S_2(f,t)|$ of interference signal and the distributed-noise $R(f,t)|N'(f,t)|$. To facilitate the estimation, we assume the following:

- $S_1(f,t)$ and $S_2(f,t)$ do not give large values simultaneously because speech signals are non-stationary.

Then the noise component $|Z(f,t)|$ in $|Y_1(f,t)|$ is estimated by

$$|Z(f,t)| = \begin{cases} |Y_2(f,t)|, & (|Y_1(f,t)| \geq |Y_2(f,t)|) \\ |Y_1(f,t)|/R(f,t-1). & (\text{otherwise}) \end{cases}$$
\( \hat{R}(f, t) \) is the estimate of the ratio \( R(f, t) \) in (3), computed in the 2S-SS.

In case where \(|Y_1(f, t)| \geq |Y_2(f, t)|\),
\[
|Z(f, t)| \approx \epsilon_2(f)|S_1(f, t)| + |N'(f, t)|, \tag{6}
\]
for the rest,
\[
|Z(f, t)| \approx \epsilon_1(f)|S_2(f, t)|/\hat{R}(f, t) - 1 + |N'(f, t)|. \tag{7}
\]
Consequently, one can summarize as follows:

- when the target speech \(|S_1(f, t)|\) is large, \(\hat{R}(f, t)|Z(f, t)|\) mainly represents the spatially-distributed noise in \(|Y_1(f, t)|\).
- when the target speech \(|S_1(f, t)|\) is small, \(\hat{R}(f, t)|Z(f, t)|\) represents the distributed noise and the residual component of interference signal in \(|Y_1(f, t)|\).

2.3 Two-stage spectral subtraction (2S-SS)\[9]\]

The 2S-SS has been proposed in [9]. It involves two separately-positioned microphones, one collecting noisy speech and the other measuring the environmental noise. In the proposed method, the BSS output \(Y_1(f, t)\) and the Noise estimation output \(|Z(f, t)|\) are input into the 2S-SS as the noisy speech and the noise, respectively. The 2S-SS eliminates the noise \(|Z(f, t)|\) from the noisy speech \(|Y_1(f, t)|\), \(|Y_1(f, t)|\) and \(|Z(f, t)|\) are also used for target-speech detection. The detected target-speech spectra are conveyed to a recognition process. The 2S-SS shown in Fig. 3 is composed of the following three components:

- Speech detection
- Single-channel spectral subtraction (1SS)
- Two-input spectral subtraction (2SS)

The target-speech detection is based on the comparison of thresholds and the ratio \(D(t)\) given by
\[
D(t) = \frac{\sum_{f=f_l}^{f_u} \sum_{\tau=0}^{\tau_w-1} |Y_1(f, t - \tau)|^2}{\sum_{f=f_l}^{f_u} \sum_{\tau=0}^{\tau_w-1} \hat{R}(f, t)^2 |Z(f, t - \tau)|^2}, \tag{8}
\]
\(f_l\) and \(f_u\) are the lower and upper limit of processed-frequency band, respectively. \(\tau_w\) denotes the average-window width. \(\hat{R}(f, t)\) represents the amplitude-correction coefficient between \(|Y_1(f, t)|\) and \(|Z(f, t)|\). It is obtained by
\[
\hat{R}(f, t) = \frac{\hat{P}_1(f, t)}{\hat{P}_2(f, t)}, \tag{9}
\]
\[
\hat{P}_1(f, t) = \frac{1}{\tau_b} \sum_{\tau=0}^{\tau_b-1} |Y_1(f, t - \tau)|^2, \tag{10}
\]
\[
\hat{P}_2(f, t) = \frac{1}{\tau_b} \sum_{\tau=0}^{\tau_b-1} |Z(f, t - \tau)|^2. \tag{11}
\]
\(\tau_b\) is the average-window width. \(\hat{R}(f, t)\) and \(\hat{P}_1(f, t)\) are calculated during non-speech period, where target speech is absent. Therefore, \(D(t) \approx 1.0\) during non-speech period, and \(D(t) > 1.0\) while target speech is present.

In two 1SS modules, the stationary noise \(\hat{P}_1(f, t)\) included in \(|Y_1(f, t)|\) and \(\hat{P}_2(f, t)\) in \(|Z(f, t)|\) are respectively subtracted as
\[
\hat{Y}_1(f, t) = \text{max} \left[ |Y_1(f, t)| - \hat{P}_1(f, t), 0 \right] + \alpha \hat{P}_1(t), \tag{12}
\]
\[
\hat{Z}(f, t) = \text{max} \left[ |Z(f, t)| - \hat{P}_2(f, t), 0 \right] + \alpha \hat{P}_2(t), \tag{13}
\]
\[
\hat{P}_i(t) = \frac{1}{f_u - f_l + 1} \sum_{f=f_l}^{f_u} \hat{P}_i(f, t)^2 \text{ for } i = 1, 2. \tag{14}
\]
Here, \(\alpha\) is a floor parameter.

Finally, the 2SS eliminates the non-stationary noise remaining in \(|\hat{Y}_1(f, t)|\) with
\[
|\hat{S}_1(f, t)| = \text{max} \left[ |\hat{Y}_1(f, t)| - \hat{R}(f, t)|\hat{Z}(f, t)|, 0 \right] + \alpha \hat{R}(f, t)\hat{P}_2(t). \tag{15}
\]
This amplitude \(|\hat{S}_1(f, t)|\) is input into speech recognition. As derived from (7), the 2SS in (15) can effectively eliminate the residual component of interference signal in the frequency bins while the spectrum \(|S_1(f, t)|\) is relatively small. This results in significant reduction of speech recognition errors as presented in the next section.

3. LVCSR Evaluation

3.1. Experimental setup

To evaluate the proposed method with LVCSR, utterances were recorded with two microphones and two loudspeakers, which were positioned as shown in Fig. 4. One of the loudspeakers was for the target speech and the other was for interference speech. The target speech consisted of 3,000 travel-conversation sentences (20 males × 150 sentences). Different sentences by other speakers were used as interference speech input. The target speech to the interference speech power ratio was 0dB. The outputs from
the two microphones were sampled at the rate of 11025Hz. For the spatially-distributed noise, the computer-room noise was used. It was recorded in a computer room with two microphones spaced 3cm apart. The recorded computer-room noise was added to the original mixed speech. The target speech to the computer-room noise power ratio (SNR) was 15dB and 10dB.

The LVCSR system [10] for travel conversation was used. The processed-frequency band was from 150 to 5000Hz. The 23-dimension (MFCC1-10, ΔMFCC1-10, ΔΔpower, pitch, Δpitch) feature vectors were extracted from each utterance. The acoustic model was the triphone HMM with 8,000 gaussians (1,000 states × 8 gaussians). The n-gram language model was prepared for travel conversation.

The proposed method [BSS+2S-SS] was compared to [1SS] and [BSS+1SS]. In the [1SS], the mixed speech from a microphone was processed by the 1SS, and the word accuracy in recognition result was measured. In the [BSS+1SS], the output \( \hat{Y}_1(f, t) \) in (12) was recognized. In [1SS] and [BSS+1SS], the power-based single-channel speech detection was performed.

3.2. Results

Figure 5 illustrates the word accuracies. The left graph is without and the right one with the interference speech. The bottom figures represents the SNR of spatially-distributed noise (computer-room noise).

In all noisy environments, the speech signals processed with the proposed method [BSS+2S-SS] were recognized with the highest success rate. This was because the 2S-SS effectively worked for both the residual component of interference speech and the spatially-distributed noise. In case where SNR = \( \infty \)dB (without interference speech), all of results were much alike due to the environment nearly silent. In case where SNR = 15dB (with interference speech), the residual component fell below the distributed-noise level, resulting in similar levels of performance for both the proposed method and [BSS+1SS]. In cases with low SNR (10dB), the proposed method significantly outperformed the method with [BSS+1SS].

4. Conclusions

Hands-free speech recognition using BSS post-processed by 2S-SS has been proposed. The ICA-based BSS has estimated a target signal and jammer signals. The 2S-SS has removed its residual crosstalk component and suppressed spatially-distributed noise. In LVCSR evaluation, utterances by other speakers and computer-room noise were used as a jammer signal and a spatially-distributed noise source, respectively. In all noisy environments, it was confirmed that the proposed method outperformed the BSS with 1SS. In future studies, we plan to make a detailed experimental comparison between the proposed and other speech enhancement methods [1],[2] in LVCSR.

5. References