MMSE STSA ESTIMATOR WITH NONSTATIONARY NOISE ESTIMATION BASED ON ICA FOR HIGH-QUALITY SPEECH ENHANCEMENT

Ryoi Okamoto, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, JAPAN
{ryoio-o, yuu-t, sawatari, shikano}@is.naist.jp

ABSTRACT

In this paper, we propose a new blind speech extraction method consisting of a minimum mean-square error short-time spectral amplitude (MMSE STSA) estimator and noise estimation based on independent component analysis (ICA). First, we perform a computer simulation using the artificial noise whose stationarity could be controlled parametrically, and the obtained results indicate that the proposed method is superior to conventional methods, such as blind spatial subtraction array (BSSA) and the original MMSE STSA estimator under the non-point-source and nonstationary noise condition. Finally, we conduct an experiment in an actual railway-station environment, and objective and subjective evaluations to confirm the advantage of the proposed method in the real world.

Index Terms— Blind source separation, microphone array signal processing, independent component analysis, spectral subtraction, MMSE STSA estimator

1. INTRODUCTION

In recent years, noise reduction techniques based on microphone array signal processing have been studied actively. One of the authors has proposed a blind speech extraction method, which is called blind spatial subtraction array (BSSA), for the realization of a noise-robust hands-free speech recognition system [1]. BSSA consists of a delay-and-sum (DS) [2]-based primary path and a reference path for independent component analysis (ICA) [3]-based noise estimation. Noise reduction in BSSA is achieved by spectral subtraction (SS) [4], that is, the estimated noise power spectrum is subtracted from the target speech power spectrum partly enhanced by DS. This method is used to extract a target speech source under the non-point-source noise condition. However, BSSA is not appropriate for listening use because SS in BSSA often generates a large amount of artificial distortion (so-called musical noise). It can be considered that the generation of its distortion is due to the fact that the short-time spectral amplitude (STSA) of the target speech estimated by SS is not optimal in the minimum mean-square error (MMSE) sense. In this paper, we focus on the MMSE STSA estimator [5] and propose a new blind speech extraction method that comprises ICA for an optimal noise estimator on the basis of MMSE. According to Ephraim and Malah [5], it is considered that the original MMSE STSA estimator is a noise reduction algorithm with a smaller amount of musical noise and a higher sound quality for listening use than SS.

The original MMSE STSA estimator requires the estimation of the spectral amplitude of noise by some means to calculate a priori and a posteriori signal-to-noise ratio (SNR) estimators. It is often conducted by detecting the nonspeech time period of the observed signal. However, this approach cannot estimate exactly the spectral amplitude of noise under the nonstationary noise condition. Therefore, the original MMSE STSA estimator often causes an unavoidable degradation of its output signal. In contrast, the proposed MMSE STSA estimator in this paper can estimate more exactly the noise component under such a condition by ICA; the proposed method can estimate a posteriori SNR dynamically and more exactly. Therefore, it can be considered that the proposed method can generate a small amount of artificial noise and reduce nonstationary noise efficiently. Besides, ICA requires no a priori information for directions-of-arrival (DOAs) of sound sources unlike supervised blocking matrices, such as fixed null beamformer (NBF), and this makes ICA more robust for estimation of noises than using traditional supervised blocking matrices, especially under reverberant conditions.

To show experimentally the effectiveness of the proposed MMSE STSA estimator under the non-point-source and nonstationary noise condition, we perform a simulation experiment using the artificial noise whose stationarity could be controlled parametrically and compare BSSA, the original MMSE STSA estimator and the proposed method. Finally, we conduct an experiment in an actual railway-station environment, and show the advantage of the proposed method in the real world by objective and subjective evaluations.

2. SIGNAL MODEL AND NOISE ESTIMATION BASED ON INDEPENDENT COMPONENT ANALYSIS

We consider an acoustic mixing model where the number of microphones is J and the observed signal contains only one target speech signal, which can be regarded as a point source, and an additive noise signal. This additive noise represents a signal that cannot be regarded as point sources. Hereafter, the observed signal vector in the time-frequency domain, \( x(f, \tau) = [x_1(f, \tau), \ldots, x_J(f, \tau)]^T \), is given by

\[
  x(f, \tau) = h(f)s(f, \tau) + n(f, \tau),
\]

where \( f \) is the frequency bin number, \( \tau \) is the time-frame index number, \( h(f) = [h_1(f), \ldots, h_J(f)]^T \) is a column vector of transfer functions from the target signal component to each microphone, \( s(f, \tau) \) is the target speech signal component, and \( n(f, \tau) = [n_1(f, \tau), \ldots, n_J(f, \tau)]^T \) is the column vector of the additive noise signal.

In ICA, we perform signal separation using an unmixing matrix \( W(f) \), so that the output signals \( y(f, \tau) = [y_1(f, \tau), y_J(f, \tau)]^T \) become...
mutually independent; this is calculated as
\[ y(f, \tau) = [y_1(f, \tau), y_2(f, \tau)]^T = W(f)x(f, \tau), \]
where \( W \) is the mixing matrix, \( x(f, \tau) \) is the input signal, and \( y(f, \tau) \) is the output signal. The noise component is calculated as
\[ y_N(f, \tau) = y(f, \tau) - \hat{y}(f, \tau). \]

The estimated noise component by ICA is subtracted from the primary path signal, resulting in the enhanced target speech signal:
\[ y_{\text{enhanced}}(f, \tau) = y_I(f, \tau) - y_N(f, \tau). \]

Finally, the noise reduction is carried out by SS, subtracting the estimated noise power spectrum of \( y_N(f, \tau) \) from the noise-only vector \( y^{\text{noise}}(f, \tau) \) as
\[ y_{\text{final}}(f, \tau) = y(f, \tau) - y^{\text{noise}}(f, \tau). \]

The proposed method can estimate temporal a priori vector \( y^{\text{noise}}(f, \tau) \) from the ICA's output signal (2) as
\[ y^{\text{noise}}(f, \tau) = [0, y_0(f, \tau)]^T. \]

Following this, we apply the projection back (PB) operation to remove the ambiguity of magnitude, and construct the estimated noise signal \( z(f, \tau) \) by applying DS as
\[ z(f, \tau) = W^{-1}(f)y^{\text{noise}}(f, \tau). \]

Finally, noise reduction is carried out by SS, subtracting the estimated noise power spectrum of (10) from the partly enhanced target speech power spectrum of (6). This procedure is given as
\[ y_{\text{final}}(f, \tau) = y(f, \tau) - y^{\text{noise}}(f, \tau). \]

where the superscripts \( (D) \) and \( (B) \) denote the real and imaginary parts, respectively.

Under the non-point-source noise condition, it is pointed out that ICA is not proficient in target speech estimation, whereas it is proficient in noise estimation [6]. That is, ICA is strongly recommended to be used for noise estimation under the non-point-source noise condition. ICA is also more proficient in noise estimation under the reverberant condition than the block-matrix with fixed coefficient, e.g., NBF.

On the basis of the above-mentioned facts, one of the authors has proposed a noise reduction method, i.e., BSSA [1], which utilizes ICA as a noise estimator.

3. PROPOSED METHOD

3.1. Our previous work: BSSA [1]

Figure 1 shows a block diagram of BSSA consisting of a DS-based primary path and a reference path for ICA-based noise estimation. The estimated noise component by ICA is subtracted from the primary path in the power spectral domain neglecting phase information; this makes BSSA realize error-robust noise reduction.

First, in the primary path, the observed signal is partly enhanced by DS. This procedure can be given as
\[ y_{DS}(f, \tau) = w_{DS}^T(f)x(f, \tau), \]
where \( y_{DS}(f, \tau) \) is a slightly enhanced target speech signal by DS, \( w_{DS}(f) \) is the filter coefficient vector of DS, \( N \) is the DFT size, \( f_s \) is the sampling frequency, \( d_j \) (\( j = 1, \ldots, J \)) denotes a microphone position, \( c \) is the sound velocity, and \( \theta_i \) represents the estimated DOA of the target speech signal given by the ICA part in the reference path. We estimate \( \theta_i \) in (8) from the unmixing matrix \( W(f) \) without a priori information of target speech’s location [7].

Next, in the reference path, the estimated target speech signal is not required and thus discarded because we want to estimate only

\[ y_{\text{final}}(f, \tau) = y(f, \tau) - y^{\text{noise}}(f, \tau). \]

where the superscripts \( (D) \) and \( (B) \) denote the real and imaginary parts, respectively.

3.2. Problem of BSSA and motivation

Indeed, BSSA can reduce non-point-source noise efficiently, but it often causes artificial noise, the so-called musical noise. Such musical noise significantly degrades the enhanced speech quality for listening use. It can be considered that its generation is due to the fact that SS is not an optimal spectral amplitude estimator on the basis of MMSE. On the other hand, the original MMSE STSA estimator [5] is an efficient noise reduction algorithm even with a small amount of musical noise. This method, however, assumes the stationarity of noise, and thus, noise reduction performance often degrades for nonstationary noise. Therefore, in this section, we propose a new MMSE STSA estimator that utilizes ICA as an optimal and frame-by-frame noise estimator in the MMSE sense. This new method has the following possible advantages: it can generate a small amount of artificial noise and suppress nonstationary noise efficiently.

3.3. Algorithm of proposed method

Figure 2 shows a block diagram of the proposed method, which can estimate noise dynamically by ICA and determine an a posteriori SNR unlike the original MMSE STSA estimator. The detailed signal processing will be shown below.

The proposed method can estimate temporal a priori and a posteriori SNRs, and a spectral gain using the estimated noise signal obtained by ICA. First, the a posteriori SNR estimate \( \hat{\gamma}(f, \tau) \) is given as
\[ \hat{\gamma}(f, \tau) = \frac{|y_{DS}(f, \tau)|^2}{\hat{\lambda}(f, \tau)}, \]
where \( \hat{\lambda}(f, \tau) \) is the power spectrum of the estimated noise (10) and
is given as
\[
\lambda(f, \tau) = \begin{cases} 
E \left[ |z(f, \tau)|^2 \right]_{\tau=\tau_0}, & (\tau \leq \tau_0), \\
E \left[ |z(f, \tau)|^2 \right]_{(\tau-\tau_0)} & (\tau > \tau_0).
\end{cases}
\] (13)
Here, \( \tau_0 \) is a smoothing parameter which denotes a certain time frame window and \( E[\cdot] \) denotes an expectation operator through \( A \) to \( B \). Note that we can estimate momentarily the a posteriori SNR estimate (12) utilizing the estimated noise signal by ICA (10), unlike the original MMSE STSA estimator. Therefore, it can be considered that our proposed method can suppress nonstationary noise more efficiently than the conventional MMSE STSA estimator.

Next, using (12), the a priori SNR estimate \( \hat{\xi}(f, \tau) \) is given as
\[
\hat{\xi}(f, \tau) = \frac{\alpha \hat{\gamma}(f, \tau - 1)}{1 + \alpha \hat{\gamma}(f, \tau - 1)} 
+ (1 - \alpha)P \left[ \hat{\gamma}(f, \tau) - 1 \right] \quad (0 \leq \alpha < 1),
\] (14)
where \( \alpha \) is the weighting factor of decision-directed estimation. \( G(f, \tau) \) is a spectral gain function and the operator \( P[\cdot] \) is defined by
\[
P[l] = \begin{cases} 
1 & (l \geq 0), \\
0 & \text{(otherwise)}. 
\end{cases}
\] (15)
Also, the spectral gain function is defined by
\[
G(f, \tau) = \Gamma(1.5)\frac{v(f, \tau)}{\hat{\gamma}(f, \tau)} \exp \left( -\frac{v(f, \tau)}{2} \right) \cdot \left[ (1 + v(f, \tau))I_0 \left( \frac{v(f, \tau)}{2} \right) + v(f, \tau)I_1 \left( \frac{v(f, \tau)}{2} \right) \right].
\] (16)
\( \Gamma(\cdot) \) denotes the gamma function and \( I_0(\cdot) \) and \( I_1(\cdot) \) denote the modified Bessel functions of zero and first-order, respectively. Moreover, \( v(f, \tau) \) is defined by
\[
v(f, \tau) = \frac{\hat{\xi}(f, \tau)}{1 + \hat{\xi}(f, \tau)} \hat{\gamma}(f, \tau).
\] (17)
Finally, noise reduction is carried out as follows:
\[
y_{\text{PROP}}(f, \tau) = G(f, \tau)y_{\text{DS}}(f, \tau),
\] (18)
where \( y_{\text{PROP}}(f, \tau) \) is the final output of this method.

In this paper, we refer to [5, eq. 7] as the spectral gain function. But even if using an alternative gain function [5, eq. 30] with speech uncertainty, ICA-based dynamic noise estimation is still available and effective.

4. SIMULATION EXPERIMENT

4.1. Experimental setup
To confirm the effectiveness of the proposed method under the non-point-source and nonstationary noise condition, we conducted a computer-simulation-based experiment. In the simulation, we compared BSSA, the proposed method and the original MMSE STSA estimator cascaded with the ICA (ICA+MMSE STSA). Figure 3(a) shows a layout of the reverberant room used in this experiment where the reverberation time is about 200 ms. We used 16 kHz-sampled signals as test data convoluted with impulse responses recorded in this reverberant room. We used 14 speakers (7 males and 7 females) as sources of the original target speech signal.

Also, to mimic nonstationary noise, the following amplitude-modulated noise signal \( n_{\text{mod}}(t) \) was introduced as a source of the original noise signal:
\[
n_{\text{mod}}(t) = P[\pi r + (1 - m)] \cdot n_{\text{white}}(t),
\] (19)
where \( m (0 \leq m < \infty) \) is a modulation parameter, \( r \) is a time index, \( r (0 < r < 1) \) is a uniform random variable ingenerated at regular time intervals and \( n_{\text{white}}(t) \) is a white Gaussian noise signal. In this amplitude modulation, we can easily control only the nonstationarity of the white Gaussian noise by changing the modulation parameter.

For example, \( n_{\text{mod}}(t) \) is equal to the (stationary) white Gaussian noise when \( m \) is zero and becomes more nonstationary as \( m \) increases. In this simulation, the uniform random variable \( r \) was ingenerated every 200 ms and the modulation parameter \( m \) was changed from 0 to 6 in 0.5 steps. The input SNR of test data was set to 10 dB. We used a two-element microphone array with an interelement spacing of 2.15 cm. The over-subtraction parameter \( \beta \) is 2, the flooring parameter \( g \) is 0 in BSSA, the smoothing parameter \( \tau_0 \) is 3 time frame windows denoting 96 ms. and the weighting factors \( \alpha \) of decision-directed estimation are 0.98 for ICA+MMSE STSA and 0.97 for the proposed method (these values are optimally fixed via experiments).

4.2. Experimental results
The simulation results are shown in Figs. 3(b) and (c) which show results for the average noise reduction rate (NRR) [7] and cepstral distortion (CD) [9] of all the target speakers for each modulation parameter. NRR is defined as the output SNR in dB minus the input SNR in dB, and CD is a measure of the degree of spectral envelope distortion via the cepstrum domain.

From Fig. 3(b), we can observe that the NRR of ICA+MMSE STSA decreases significantly as \( m \) increases. In contrast, the NRR of the proposed method negligibly decreases even if the modulation parameter \( m \) increases. This finding indicates that the proposed method is robust against nonstationary noise unlike the simple combination of the ICA and MMSE STSA estimator. Also, the NRR of the proposed method is superior to that of BSSA. From Fig. 3(c), we can confirm that the CD of BSSA is almost equal to the proposed method. However, the CD of ICA+MMSE STSA fluctuates along with the modulation parameter. It can be considered that over-reduction or under-reduction arises owing to an inaccurate noise estimation for nonstationary noise. As a result, we can reveal that the proposed method is more robust against nonstationary noise than the original MMSE STSA estimator, and is superior to BSSA.

5. EXPERIMENT IN REAL WORLD

5.1. Experimental setup
We conducted experiments in an actual railway-station environment to confirm the effectiveness of the proposed method in the real world. In this experiment, we compare the following four methods: BSSA, the proposed method, ICA+MMSE STSA and the MMSE STSA estimator with noise estimation based on NBF (MMSE STSA-NBF). In this MMSE STSA-NBF, NBF for the noise estimation is fixed to steer its spatial null to the assumed target speech direction of 0° (normal to the microphone array). Figure 4 shows a layout of the railway-station environment where the reverberation time is about 1000 ms. We used 16 kHz-sampled signals as test data. We used 4 speakers (2 males and 2 females) as the target speech signal, and 5 noise signals recorded in distinct time periods in this environment as noise signal. The noise in this environment is nonstationary and is almost diffused. The input SNR of test data is set to 0 dB. We used a two-element microphone array with an interelement spacing of 2.15 cm. The direction of the target speech is not exactly 0° but fluctuates around 0° according to speakers. The parameters for each method are the same as those shown in Sect. 4.1 and a weight factor \( \alpha \) of decision-directed estimation is 0.97 for MMSE STSA-NBF.

5.2. Experimental results
The simulation results are shown in Fig. 5. Figures 5(a) and (b) show results for the average NRR and CD of all the target speakers and the noise signals. Figure 5(c) shows results for a subjective evaluation by the human ear. Six examinees participated in the subjective evaluation, where a pair of processed speech signals were presented
From Figs. 5(a) and (b), we can observe that the NRR of the proposed method is significantly superior to those of the conventional methods, and the CD of the proposed method is slightly inferior to that of ICA+MMSE STSA but is significantly superior to that of BSSA and MMSE STSA-NBF. From these results, we can confirm that the proposed method has a higher performance of noise reduction than the conventional methods in the real world, although the proposed method generates a small amount of distortion compared with the original MMSE STSA estimator. Finally, we show the result of subjective evaluation in Fig. 5(c). From Fig. 5(b), we can observe that the CD of MMSE STSA-NBF is significantly inferior to those of the others, and the authors evaluated that the sound quality of MMSE STSA-NBF is obviously lower. Therefore, we conducted a subjective evaluation except MMSE STSA-NBF. In Fig. 5(c), the performance of the proposed method indicates that the proposed method shows a significant preference compared with the conventional methods, and we can also confirm that this method is more appropriate for listening use than the conventional methods in the real world.

6. CONCLUSION

In this paper, we proposed a new blind speech extract method based on the MMSE STSA estimator and ICA-based noise estimation. In a simulation experiment, we confirmed that the proposed method is more effective than BSSA and the original MMSE STSA estimator under the non-point-source and nonstationary noise condition. We had an experiment in the real world, and confirmed the advantage of the proposed method by objective and subjective evaluations.

7. REFERENCES