Speech Recognition Using Blind Source Separation and Dereverberation Method for Mixed Sound of Speech and Music

Longbiao Wang*, Kyohei Odani†, Atsuhiko Kai† and Weifeng Li‡
*Nagaoka University of Technology, Nagaoka 9402188, Japan
E-mail: wang@vos.nagaokaut.ac.jp
†Graduate School of Engineering, Shizuoka University, Hamamatsu 4328561, Japan
E-mail: odani@spa.sys.eng.shizuoka.ac.jp, kai@sys.eng.shizuoka.ac.jp
‡Tsinghua University, Shenzhen 100084, China
E-mail:li.weifeng@sz.tsinghua.edu.cn

***Abstract*—In this paper, we propose a method for performing a non-stationary noise reduction and dereverberation method. We use a blind dereverberation method based on spectral subtraction using a multi-channel least mean square algorithm has been proposed in our previous study. To suppress the non-stationary noise, we used a blind source separation based on an efficient fast independent component analysis algorithm. This method is evaluated using a mixed sound of speech and music, and achieves an average relative word error reduction rate of 41.9% and 7.9% compared with a baseline method and the state-of-the-art multi-step linear prediction-based dereverberation, respectively, in a real environment.**

Index Terms: hands-free speech recognition, blind dereverberation, blind source separation, multi-channel least mean square, generalized spectral subtraction

I. Introduction

In a distant-talking environment, background noise and reverberation drastically degrade speech recognition performance because of a mismatch between the training and test environments. Current approaches to robustness issues for automatic speech recognition (ASR) in noisy reverberant environments can be classified as speech enhancement, robust feature extraction, or model adaptation methods. Several previous studies have focused on speech enhancement, especially multi-channel speech is often used. As conventional method to suppress the background noise, blind source separation based on independent component analysis (ICA) was proposed [1], [2], [3]. In this technique, a set of original signals are retrived from their mixtures based on the assumption of their mutual statistical independence. For suppressing the reverberation, [4] proposed the method was based on constructing the null subspace of the data matrix in the presence of colored noise and employing a generalized singular-value decomposition or generalized eigenvalue decomposition of the respective correlation matrices. In [10], an adaptive multi-channel least mean square (MCLMS) algorithm was proposed to blindly identify the channel impulse response in time domain. A novel dereverberation method was proposed that utilized long-term multiple-step linear prediction [5]. This enabled the linear prediction coefficients to be estimated in a time domain and the amplitude of late reflections to be suppressed through spectral subtraction in a frequency domain.

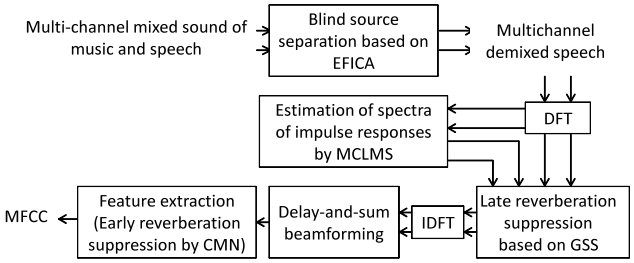

Fig. 1. Schematic diagram of our method

In our previous study, we proposed a blind dereverberation method based on generalized spectral subtraction (GSS) employing the adaptive MCLMS algorithm in a frequency domain, and this method was effective in several reveberant environments [6]. However, dereverberation method is not effective in environment that there are background noise and reverberation simultaneously. Assuming stationary noise, we proposed a blind denoising and dereverberation method that combines noise reduction and dereverberation based on GSS [7]. However, whereas GSS-based noise reduction is effective for stationary noise, it is not effective against non-stationary noise such as music.

In this paper, we present a non-stationary noise reduction and dereverberation method, and evaluate this method using a mixed sound of speech and music. To suppress the non-stationary noise, we use blind source separation based on ICA by applying Efficient FastICA (EFICA) [3], an improved version of the popular FastICA algorithm [2]. A schematic diagram of our proposed method is shown in Fig. 1. At first, blind source separation based on EFICA is used to demix the speech and music. Next, GSS-based blind dereverberation reduces the late reverberation using the impulse responses

estimated from the demixed speech. Thereafter, the early reverberation is normalized by cepstral mean normalization (CMN) at the feature extraction stage.

II. Outline of blind source separation

In this section, we briefly explain ICA-based blind source separation. Let $\mathbf{W}$ be the demixing matrix trained from the observed signal $\mathbf{X}$ such that the demixed signal $\mathbf{Y}$ can be written as $\mathbf{Y} = \mathbf{WX}$. Here, $\mathbf{X}$ is an $M \times N$ matrix and $\mathbf{W}$ is an $M \times M$ matrix, where M is the number of mixed signals and N is the number of samples in each signal.

FastICA [2] is one of the most popular algorithms for estimating the demixing matrix $\mathbf{W}$. In this paper, we use the improved EFICA algorithm [3]. EFICA combines symmetric FastICA with an adaptive choice of nonlinearities g, which are fixed in FastICA. The algorithm consists of: (1) Running the original symmetric FastICA. (2) Adaptively choosing the different nonlinearities g_k. (3) Refining or fine-tuning each of the source components found by one-unit FastICA. In this paper, we use the source separation tool T-ABCD [9] which implements the EFICA algorithm.

III. Outline of Blind Dereverberation

A. Dereverberation based on GSS

If speech $s[t]$ is corrupted by convolutional noise $h[t]$, the observed speech $x[t]$ becomes $x[t] = h[t] * s[t]$, where $*$ denotes the convolution operation. If the length of the impulse response is smaller than the analysis window length T used in the short-time Fourier transform (STFT), the STFT of the distorted speech equals that of the clean speech multiplied by the STFT of the impulse response $h[t]$. However, if the length of the impulse response is greater than the analysis window size, the STFT of the distorted speech is usually approximated by

$$X(\tau,\omega) \approx S(\tau,\omega)H(0,\omega) + \sum_{d=1}^{D-1} S(\tau-d,\omega)H(d,\omega) \quad (1)$$

where τ is the frame index, $H(\omega)$ is the STFT of the impulse response, $S(\tau,\omega)$ is the STFT of the clean speech s, D is the number of reverberation windows, and $H(d,\omega)$ denotes the part of $H(\omega)$ corresponding to the frame delay d.

In [6], we proposed a dereverberation method based on GSS to estimate the STFT of the clean speech $\hat{S}(\tau,\omega)$ based on Eq. (1). To estimate the spectrum of the impulse response for the GSS, the MCLMS algorithm was extended to identify the impulse responses in the frequency domain. The estimated spectrum of clean speech may not be very accurate due to estimation errors in the impulse response, especially the earlier parts. In addition, an unreliable estimated spectra in previous frame causes further estimation error in the current frame. In this paper, we reduce late reverberation using GSS and normalize early reverberation by CMN at the feature extraction stage.

Assuming, for simplicity, that the phases of different frames are non-correlated, the estimated spectrum $\hat{X}(\tau,\omega)$ obtained by reducing the late reverberation becomes

$$|\hat{X}(\tau,\omega)|^{2n} \approx max\left\{|X(\tau,\omega)|^{2n} - \alpha \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{X}(\tau-d,\omega)|^{2n}|\hat{H}(d,\omega)|^{2n}\}}{|\hat{H}(0,\omega)|^{2n}}, \beta \cdot |X(\tau,\omega)|^{2n}\right\} \quad (2)$$

where $|\hat{X}(\tau,\omega)|^{2n} = |\hat{S}(\tau,\omega)|^{2n}|\hat{H}(0,\omega)|^{2n}$, $|\hat{S}(\tau,\omega)|^2$ is the estimated spectrum of clean speech, $\hat{H}(\tau,\omega)$ is the STFT of the impulse response obtained by the frequency-domain MCLMS algorithm (discussed in Sec. 3.2), α is the noise overestimation factor, β is a spectral floor parameter to avoid negative or underflow values, and n is an exponent parameter.

B. Blind Estimation of Impulse Responses

In this section, we explain the blind estimation of impulse response spectra $\hat{H}(d,\omega)$ using Eq. (2). In [10], time-domain MCLMS was proposed as a technique for blindly estimating the impulse responses of each channel. In this paper, we use a variable step-size unconstrained MCLMS (VSS-UMCLMS) algorithm to extend from the time domain to the frequency domain.

In the absence of additive noise, we have the following relation between the correlation matrix and the impulse response.

$$\mathbf{R}_{X_iX_i}(\tau+1)\mathbf{H}_j(\tau) = \mathbf{R}_{X_iX_j}(\tau+1)\mathbf{H}_i(\tau) \quad i,j = 1,2,\cdots,N, i \neq j \quad (3)$$

$$\mathbf{R}_{X_iX_j}(\tau) = E[\mathbf{X}_i(\tau)\mathbf{X}_j^T(\tau)] \quad (4)$$

$$\mathbf{X}_i(\tau) = [X_i(\tau), X_i(\tau-1), \cdots, X_i(\tau-D+1)]^T \quad (5)$$

$$\mathbf{H}_i(\tau) = [H_i(\tau,0), \cdots, H_i(\tau,d), \cdots, H_i(\tau,D-1)]^T \quad (6)$$

where i is the channel number, $\mathbf{X}_i(\tau)$ is the spectrum of the observed signal in frame τ, $\mathbf{H}_i(\tau)$ is the spectrum of the impulse response in frame τ, and $H_i(\tau,d)$ is the spectrum of the impulse response in frame τ corresponding to the frame delay d.

Transposing the right-hand side of Eq. (3) and combining terms over all channels, we obtain Eq. (7).

$$\mathbf{R}_{X+}(\tau+1)\mathbf{H}(\tau) = \mathbf{0} \quad (7)$$

$$\mathbf{H}(\tau) = [\mathbf{H}_1(\tau)^T, \mathbf{H}_2(\tau)^T, \cdots, \mathbf{H}_N(\tau)^T]^T \quad (8)$$

$$\mathbf{R}_{X+}(\tau) = \begin{bmatrix} \sum_{i\neq 1}\mathbf{R}_{X_iX_i}(\tau) & -\mathbf{R}_{X_2X_1}(\tau) & \cdots & -\mathbf{R}_{X_NX_1}(\tau) \\ -\mathbf{R}_{X_1X_2}(\tau) & \sum_{i\neq 2}\mathbf{R}_{X_iX_i}(\tau) & \cdots & -\mathbf{R}_{X_NX_2}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{X_1X_N}(\tau) & -\mathbf{R}_{X_2X_N}(\tau) & \cdots & \sum_{i\neq N}\mathbf{R}_{X_iX_i}(\tau) \end{bmatrix} \quad (9)$$

Eq. (7) uses the true impulse response in the absence of additive noise. When additive noise is present or the estimated impulse response is used, however, the observed estimation error is given by Eq. (10).

$$\tilde{\mathbf{R}}_{X+}(\tau+1)\hat{\mathbf{H}}(\tau) = \mathbf{E}(\tau+1) \quad (10)$$

where $\tilde{\mathbf{R}}_{X+}$ is the matrix of Eq. (9) calculated using a noisy observed signal and $\mathbf{E}$ is the estimation error. $\hat{\mathbf{H}}$ is adaptively trained by minimizing the cost function obtained from the estimation error. The learning equation in unconstrained MCLMS is as follows:

$$\hat{\mathbf{H}}(\tau+1) = \hat{\mathbf{H}}(\tau) - 2\mu\tilde{\mathbf{R}}_{X+}(\tau+1)\hat{\mathbf{H}}(\tau) \tag{11}$$

where μ is the step-size. Multi-channel impulse responses can be estimated by updating Eq. (11).

The VSS-UMCLMS algorithm automatically determines the step-size μ in Eq. (11). This is updated according to Eq. (12).

$$\mu_{opt}(\tau+1) = \frac{\hat{\mathbf{H}}^T(\tau)\Delta\mathbf{J}(\tau+1)}{||\Delta\mathbf{J}(\tau+1)||^2} \tag{12}$$

where

$$\Delta\mathbf{J}(\tau+1) \approx \frac{2\tilde{\mathbf{R}}_{X+}(\tau+1)\hat{\mathbf{H}}(\tau)}{||\hat{\mathbf{H}}(\tau)||^2} \tag{13}$$

The impulse response spectra can be blindly estimated using VSS-UMCLMS.

IV. Experiments

A. Evaluation data

A. Simulated mixed sound of speech and music

We simulated multi-channel distorted speech signals by combining speech and music that had been recorded separately in a real environment. Table I gives the recording conditions and content. One hundred utterances from the Japanese Newspaper Article Sentences (JNAS) corpus, uttered by five male speakers seated on chairs A to E in Fig. 2, were recorded by a multi-channel recording device. The heights of the microphone array and the utterance position of each speaker were about 0.8 m and 1.0 m, respectively. We used a nine-channel microphone array (Fig. 2) and a pin microphone to record speech in distant-talking and close-talking environments, respectively. The average signal-to-noise ratio (SNR) of the speech was about 21 dB. Monaural music was played by a Logicool LS11 2.0 Stereo Loudspeaker on the table and recorded by a multi-channel recording device. The recorded music, which was categorized as hard rock and did not include a singing voice, was added to the speech at an SNR of 10 dB and 20 dB for each channel.

B. Real mixed sound of speech and music

To evaluate our proposed method in a real environment, we recorded multi-channel speech that was simultaneously degraded by music and reverberation. One hundred utterances from the JNAS corpus, uttered by one male speaker seated on chair A in Fig. 2, were recorded by a multi-channel recording device. We played music continuously until all utterances were finished. The other recording conditions were the same as for the simulated mixed sound recordings. The average SNR of the speech was about 3.4 dB.

TABLE I
CONDITIONS FOR RECORDING.

microphone	SONY ECM-C10
A/D board	Tokyo Electron device TD-BD-16ADUSB
recording room size [m]	7.1(D) × 3.3(W) × 2.5(H)
number of utterances	100 utterances
sampling frequency	16 kHz
quantization bit rate	16 bits

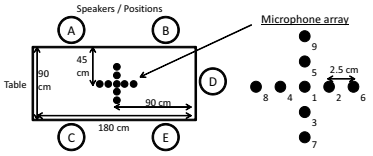

Fig. 2. Illustration of recording settings and microphone array

B. Experimental setup

Table II lists the speech recognition conditions. The acoustic models were trained with the Acoustical Society of Japan's (ASJ) speech database of phonetically balanced sentences (ASJ-PB) and the JNAS. In total, around 20,000 sentences (clean speech) uttered by 132 male speakers were used. Table III gives the conditions for SS-based denoising and dereverberation. The parameters shown in Table III were determined using the simulated noisy reverberant speech. The number of reverberant windows was set to $D = 6$ (192 ms). For the proposed SS-based dereverberation method, the clean power spectrum for each frame was incrementally estimated by the clean power spectra of preceding non-overlapping frames, as this study used a frame shift of half the frame length. The spectrum of the impulse response $\hat{H}(d,\omega)$ was estimated to allow each utterance to be recognized. We used the open-source Julius large vocabulary continuous speech recognition (LVCSR) decoder [11], which is based on word trigrams and triphone context-dependent hidden Markov models (HMMs).

TABLE II
CONDITIONS FOR SPEECH RECOGNITION.

sampling frequency	16 kHz
frame length	25 ms
frame shift	10 ms
acoustic model	5 states, 3 output probability left-to-right triphone HMMs
feature space	25 dimensions with CMN (12MFCCs + Δ + Δpower)

TABLE III
CONDITIONS FOR GSS-BASED DEREVERBERATION.

analysis window	Hamming
window length	32 ms
window shift	16 ms
noise overestimation factor α	0.1
spectral floor parameter β	0.15
exponent parameter n	0.1

C. Experimental results

The speech recognition results from the proposed method are shown in Figs. 3 and 4. There were obtained using two channels (Mic. 1 and 2 in Fig. 2) and four channels (Mic. 6, 7, 8, and 9 in Fig. 2) for the blind estimation of impulse response and delay-and-sum beamforming, respectively. In Figs. 3 and 4, "CMN only", "EFICA", "EFICA+MSLP", and "EFICA+MCLMS" denote results from conventional CMN, blind source separation based on EFICA, the combination of EFICA with blind dereverberation based on multiple-step linear prediction (MSLP) [5], and the combination EFICA with blind dereverberation based on MCLMS (our proposed method). Dereverberation utilizing MSLP shows good performance under various reverberant environments. In this paper, delay-and-sum beamforming was performed for all methods.

In Figs. 3 and 4, "Real (w/o music)", "Simulated (20 dB)", and "Simulated (10 dB)" indicate speech without music and the simulated mixed sound of speech and music added at an SNR of 20 dB and 10 dB, respectively. The speech recognition performance of "CMN only" was drastically degraded owing to the noisy reverberant conditions and the fact that CMN does not suppress the music or late reverberation. "EFICA" improved the speech recognition performance significantly compared with "CMN only" under all conditions, and "EFICA+MSLP" further improved the performance compared with "EFICA". Our proposed method ("EFICA+MCLMS") outperformed all of the other methods, including "EFICA+MSLP", especially when using two channels. With two channels in the simulated environment at an SNR of 10 dB, our proposed method achieved an average relative word error reduction rate of 48.2% and 12.4% compared with "CMN only" and "EFICA+MSLP", respectively.

"Real (with music)" indicates the real mixed sound of speech and music. The speech recognition performance of "CMN only" with this real mixed sound was again considerably degraded for the reasons mentioned previously. When using two channels, "EFICA" did not offer any improvement over "CMN only" because of the smaller SNR than in the simulated speech, the smaller number of microphones, and the shorter distance between the microphone pair. Consequently, "EFICA+MSLP" and "EFICA+MCLMS" achieved only a small increase in word recognition accuracy. On the other hand, when using four channels, "EFICA" improved the speech recognition performance significantly compared with "CMN only". "EFICA+MCLMS" gave a marked improvement over "EFICA", and also outperformed "EFICA+MSLP". When using four channels, our proposed method achieved an average relative word error reduction rate of 41.9% and 7.9% compared with "CMN only" and "EFICA+MSLP".

V. Conclusion and Future Work

In this paper, we presented a method that combined blind dereverberation based on MCLMS with blind source separation based on EFICA under a reverberant environment with non-stationary noise. To evaluate our proposed method, we prepared simulated and real mixtures of speech and music. The results showed that our proposed method was more effective than combining EFICA with the state-of-the-art MSLP. In future work, we intend to extend our proposed method to deal with real-world speech data, including overlapping speech that involves multiple persons speaking simultaneously.

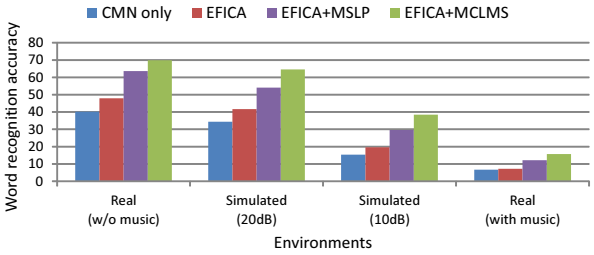

Fig. 3. Word accuracy using two channels for LVCSR

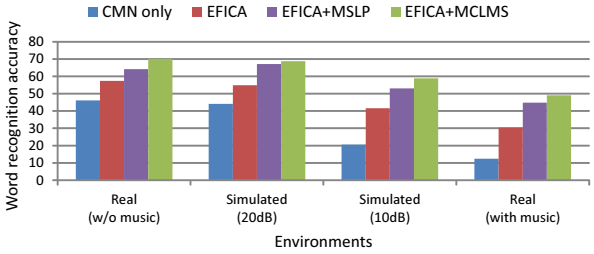

Fig. 4. Word accuracy using four channels for LVCSR

References

[1] A. J. Bell and T. J. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," Signal Processing, Vol. 7, No. 6, pp. 1129-1159, 1995.

[2] A. Hyvaerinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," IEEE Transactions on Neural Networks, Vol. 10, No. 3, pp. 626-634, 1999.

[3] Z. Koldovský, P. Tichavský and E. Oja, "Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound," IEEE Trans. on Neural Networks, Vol. 17, No. 5, pp. 1265-1277, September 2006.

[4] S. Gannot and M. Moonen, "Subspace Methods for Multimicrophone Speech Dereverberation," EURASIP Journal on Advances in Signal Processing, Vol. 2003, Issue 1, pp. 1074-1090, 2003.

[5] K. Kinoshita, M. Delcroix, T. Nakatani and M. Miyoshi, "Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction," IEEE Trans. on ASLP, Vol. 17, N0. 4, pp. 534-545, May 2009.

[6] K. Odani, L. Wang and A. Kai, "Blind Dereverberation Based on Generalized Spectral Subtraction by Multi-channel LMS Algorithm, " Proc. of APSIPA ASC 2011, Oct. 2011.

[7] K. Odani, L. Wang and A. Kai, "Speech Recognition by Denoising and Dereverberation Based on Spectral Subtraction in a Real Noisy Reverberant Environment, " Proc. of INTERSPEECH 2012, Sep. 2012.

[8] T. Yoshioka, T. Nakatani, M. Miyoshi and H. G. Okuno, "Blind Separation and Dereverberation of Speech Mixtures by Joint Optimization, " IEEE Trans. on ASLP, Vol. 19, No. 1, Jan. 2011.

[9] Z. Koldovský and P. Tichavský, "Time-Domain Blind Separation of Audio Sources on the basis of a Complete ICA Decomposition of an Observation Space," IEEE Trans. on ASLP, Vol. 19, No. 2, pp. 406-416, Feburary 2011.

[10] Y. Huang, J. Benesty and J. Chen, "Optimal Step Size of the Adaptive Multi-channel LMS Algorithm for Blind SIMO Identification," IEEE Signal Processing Letters, Vol. 12, No. 3, pp. 173-175, Mar. 2005.

[11] A. Lee, T. Kawahara and K. Shikano, "Julius — an Open Source Real-Time Large Vocabulary Recognition Engine," Proc. of European Conference on Speech Communication and Technology, pp. 1691-1694, Sep. 2001.