ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Speech Recognition by Denoising and Dereverberation Based on Spectral Subtraction in a Real Noisy Reverberant Environment

Kyohei Odani[1], *Longbiao Wang*[2], *Atsuhiko Kai*[2]

[1]Graduate School of Engineering, Shizuoka University, Hamamatsu 432-8561, Japan
[2]Faculty of Engineering, Shizuoka University, Hamamatsu 432-8561, Japan
odani@spa.sys.eng.shizuoka.ac.jp, {wang, kai}@sys.eng.shizuoka.ac.jp

10.21437/Interspeech.2012-384

Abstract

A blind dereverberation method based on spectral subtraction using a multi-channel least mean squares algorithm was previously proposed. The results of a large vocabulary continuous speech recognition task showed that this method achieved significant improvements over the conventional method based on cepstral mean normalization and beamforming in a simulated reverberant environment without additive noise. In this paper, we evaluate the blind dereverberation method in a real noisy reverberant environment. We present a denoising and dereverberation method based on power spectral subtraction or generalized spectral subtraction, and evaluate our proposed method using speech in a real environment. The generalized spectral subtraction based method achieves an average relative word error reduction rate of 39.1% and 11.5% compared to the conventional cepstral mean normalization and power spectral subtraction based methods, respectively.

Index Terms: hands-free speech recognition, blind dereverberation, noise reduction, spectral subtraction, real environment

1. Introduction

In a distant-talking environment, channel distortion drastically degrades speech recognition performance because of mismatches between the training and test environments. The current approaches focusing on robustness issues for automatic speech recognition (ASR) in noisy reverberant environments can be classified as speech enhancement, robust feature extraction, or model adaptation methods.

In this paper, we focus on speech enhancement in a distant-talking environment. Previously, Wang et al. [1] proposed a robust distant-talking speech recognition method based on power spectral subtraction (SS) employing the adaptive multi-channel least mean squares (MCLMS) algorithm. In their study, late reverberation was treated as additive noise, and a noise reduction technique based on power SS was proposed to estimate the power spectrum of clean speech using an estimated power spectrum of the impulse response. To estimate the power spectra of the impulse responses, they extended the MCLMS algorithm for identifying impulse responses in a time domain [2] to a frequency domain. Odani et al. [3] proposed a blind dereverberation method based on generalized SS (GSS), which has been shown to be effective for noise reduction, instead of power SS. The dereverberation method based on GSS with beamforming achieved a relative word error reduction rate of 9.8% and 31.4% compared to the dereverberation method based on power SS with beamforming and the conventional cepstral mean normalization (CMN) with beamforming, respectively. However, both the power SS-based method [1] and GSS-based method [3] were evaluated in a simulated reverberant environment without additive noise.

In this paper, we evaluate the blind dereverberation methods in a real noisy reverberant environment. We present a denoising and dereverberation method based on power SS or GSS. The schematic diagram of our proposed denoising and dereverberation method is shown in Fig. 1. Background noise and the late reverberation are reduced from the spectrum of multi-channel distorted speech by our proposed method using the estimated spectrum of the background noise and impulse response. Thereafter, the early reverberation is normalized by CMN at the feature extraction stage.

2. Outline of Denoising and Dereverberation

2.1. Denoising and dereverberation based on power SS

If speech $s[t]$ is corrupted by convolutional noise $h[t]$ and additive noise $n[t]$, the observed speech $x[t]$ becomes

$$x[t] = h[t] * s[t] + n[t], \quad (1)$$

where * denotes the convolution operation. If the length of the impulse response is much 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 much greater than the analysis window size, the

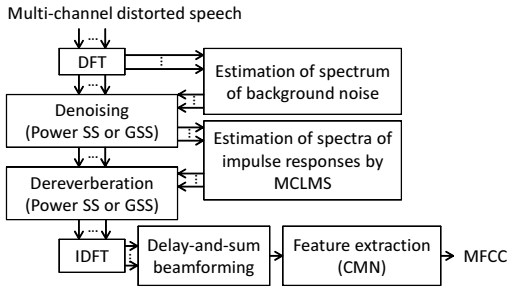

Figure 1: Schematic diagram of denoising and dereverberation method

STFT of the distorted speech is usually approximated by

$$X(f,\omega) \approx S(f,\omega) * H(\omega) + N(f,\omega) = S(f,\omega)H(0,\omega) + \sum_{d=1}^{D-1} S(f-d,\omega)H(d,\omega) + N(f,\omega), \quad (2)$$

where f is the frame index, $H(\omega)$ is the STFT of the impulse response, $S(f,\omega)$ 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 [1], Wang et al. proposed a dereverberation method based on power SS to estimate the STFT of the clean speech $\hat{S}(f,\omega)$ based on Eq. (2). To estimate the spectrum of the impulse response for the SS, they extended the MCLMS algorithm for identifying the impulse responses in a time domain to a frequency domain. Assuming that the phases of different frames are noncorrelated for the sake of simplicity, the power spectrum of Eq. (2) can be approximated as

$$|X(f,\omega)|^2 \approx |S(f,\omega)|^2|H(0,\omega)|^2 + \sum_{d=1}^{D-1} |S(f-d,\omega)|^2|H(d,\omega)|^2 + |N(f,\omega)|^2. \quad (3)$$

Furthermore, the late reverberation is reduced based on the power SS, while the early reverberation is normalized by CMN at the feature extraction stage. SS is used to prevent the estimated power spectrum obtained by reducing the late reverberation from being a negative value; the estimated power spectrum $|\hat{X}(f,\omega)|^2$ obtained by reducing the late reverberation then becomes

$$|\hat{X}(f,\omega)|^2 \approx max\Big\{|\hat{X}_N(f,\omega)|^2 - \alpha_1 \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{X}(f-d,\omega)|^2|\hat{H}(d,\omega)|^2\}}{|\hat{H}(0,\omega)|^2}, \beta_1 \cdot |\hat{X}_N(f,\omega)|^2\Big\}, \quad (4)$$

$$|\hat{X}_N(f,\omega)|^2 \approx max\{|X(f,\omega)|^2 - \alpha_2 \cdot |\hat{N}(\omega)|^2, \beta_2 \cdot |X(f,\omega)|^2\}, \quad (5)$$

where $|\hat{X}(f,\omega)|^2 = |\hat{S}(f,\omega)|^2|\hat{H}(0,\omega)|^2$, $|\hat{S}(f,\omega)|^2$ is the spectrum of estimated clean speech, $\hat{H}(f,\omega)$ is the STFT of the impulse response obtained by frequency domain MCLMS algorithm [1], $\hat{N}(\omega)$ is the mean of noise spectrum $N(f,\omega)$, and $\hat{X}_N(f,\omega)$ is the spectrum obtained by subtracting the estimated mean spectrum of noise $\hat{N}(\omega)$ [1] from the spectrum of the observed speech, respectively. In this paper, we set parameter β_1 equal to β_2.

Table 1: Details of recording conditions for impulse response measurement. "RT60 (s)": reverberation time in room. "S": small, "L": large.

array #	database	room	RT60
1	RWCP	echo room (cylinder)	0.38
2	RWCP	tatami-floored room (S)	0.47
3	RWCP	tatami-floored room (L)	0.60
4	CENSREC-4	lounge	0.50
5	CENSREC-4	Japanese style bath	0.60
6	CENSREC-4	living room	0.65
7	CENSREC-4	elevator hall	0.75

2.2. Denoising and dereverberation based on GSS

In [3], we proposed a blind dereverberation method based on GSS, which has been shown to be effective for noise reduction, instead of power SS. Instead of the power SS-based denoising and dereverberation given in Eq. (4), the modified GSS-based denoising and dereverberation is defined as

$$|\hat{X}(f,\omega)|^{2n} \approx max\Big\{|\hat{X}_N(f,\omega)|^{2n} - \alpha_1 \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{X}(f-d,\omega)|^{2n}|\hat{H}(d,\omega)|^{2n}\}}{|\hat{H}(0,\omega)|^{2n}}, \beta_1 \cdot |\hat{X}_N(f,\omega)|^{2n}\Big\}, \quad (6)$$

$$|\hat{X}_N(f,\omega)|^{2n} \approx max\{|X(f,\omega)|^{2n} - \alpha_2 \cdot |\hat{N}(\omega)|^{2n}, \beta_2 \cdot |X(f,\omega)|^{2n}\}, \quad (7)$$

where n is the exponent parameter. For power SS, the exponent parameter n is equal to 1. In this paper, the optimal value of exponent parameter n was empirically determined using simulated noisy reverberant speech.

3. Evaluation data

3.1. Simulated noisy reverberant speech

For the simulated noisy reverberant speech, we used multi-channel distorted speech signals simulated by convolving multi-channel impulse responses with clean speech and then adding background noise. Seven kinds of multi-channel impulse responses measured in various acoustic reverberant environments were selected from the Real World Computing Partnership (RWCP) sound scene database [4] and the CENSREC-4 database [5]. Table 1 lists the conditions for the seven recordings using a two-channel microphone array. For the RWCP database, two-channel microphones located at 5.85 cm intervals were taken from a circular microphone array (16 channels). For the CENSREC-4 database, two-channel microphones located at 2.125 cm intervals were taken from a linear microphone array (7 channels). Computer room noise

[1] In this study, stationary noise is assumed.

Table 2: 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 speakers	5 male speakers
number of utterances	100 utterances (about 20 utterances per speaker)
background noise	electric fan
sampling frequency	16 kHz
quantization bit rate	16 bits

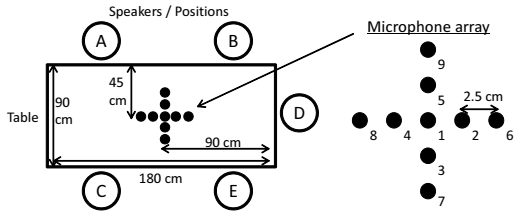

Figure 2: Illustration of recording settings and microphone array

used as background noise was added at an SNR (signal to noise ratio) of 20 dB. One hundred utterances from the Japanese Newspaper Article Sentences (JNAS) corpus were used as clean speech.

3.2. Real noisy reverberant speech

To evaluate our proposed denoising and dereverberation method in a real environment, we recorded multi-channel speech degraded simultaneously by background noise and reverberation. Table 2 gives the conditions and content of the recordings. One hundred utterances from the JNAS corpus, uttered by five male speakers seated on the chairs labeled 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. An electric fan with high air volume located behind the speaker in position A was used as background noise. An average SNR of the speech was about 18 dB. We used a microphone array with 9 channels (Fig. 2) and a pin microphone to record speech in the distant-talking environment and close-talking environment, respectively.

4. Experiments

4.1. Experimental setup

Table 3 gives the conditions for speech recognition. 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 20K sentences (clean speech) uttered by 132 male speakers were used. Table 4 gives the conditions for SS-based

Table 3: 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 4: Conditions for SS-based denoising and dereverberation. "DN": denoising. "DR": dereverberation.

method	Power SS		GSS	
	DN	DR	DN	DR
analysis window	Hamming			
window length	32 ms			
window shift	16 ms			
noise overestimation factor α	α_2= 3.0	α_1= 1.0	α_1=α_2= 0.1	
spectral floor parameter β	β_1=β_2=0.15			

denoising and dereverberation. The parameters shown in Table 4 were determined using the simulated noisy reverberant speech. The number of reverberant windows D was set to 6 (192 ms). For the proposed dereverberation method based on SS, the previous clean power spectra estimated with a skip window were used to estimate the current clean power spectrum since in this study we used a frame shift of half the frame length [1]. The spectrum of the impulse response $\hat{H}(d, \omega)$ was estimated for each utterance to be recognized. An open-source LVCSR decoder, Julius [6], based on word trigram and triphone context-dependent hidden Markov models (HMMs), was used.

4.2. Experimental results for the simulated noisy reverberant speech

In both our power SS-based and GSS-based denoising and dereverberation methods, speech signals from multiple microphones were used to estimate blindly the compensation parameters for the power SS and GSS. Thereafter, noise and reverberation were suppressed by our method and the spectrum of denoising and dereverberant speech was inverted into a time domain. Finally, delay-and-sum beamforming was performed on the multi-channel dereverberant speech.

Table 5 shows the speech recognition results for the power SS and GSS-based denoising and dereverberation methods for the simulated noisy reverberant speech (development set). The word accuracy rate for LVCSR with clean speech was 92.6%. "Distorted speech #", "DN" and "DNR" in Table 5 denote the "array #" in Table 1, "denoising", and "denoising and dereverberation", respectively. The speech recognition performance of conventional CMN was drastically degraded owing to the noisy reverberant conditions and the fact that CMN did

Table 5: Word accuracy for LVCSR with the simulated noisy reverberant speech (%). Delay-and-sum beamforming was performed for all methods.

Distorted Speech #	CMN only	Power SS		GSS	
		DN	DNR	DN	DNR
1	28.2	37.4	48.8	30.3	48.3
2	16.0	25.9	33.5	18.8	36.3
3	9.5	21.3	31.3	13.9	32.8
4	55.8	72.2	69.9	60.4	68.2
5	17.2	24.4	32.0	20.9	37.7
6	26.1	32.8	45.3	30.0	51.7
7	54.4	64.6	66.5	57.7	68.8
Average	29.6	39.8	46.7	33.1	49.1

not suppress the late reverberation. The power SS-based DN using Eq. (5) improved speech recognition performance significantly compared to the CMN for all reverberant conditions. The GSS-based DN using Eq. (7), however, did not improve the speech recognition performance compared to the power SS-based DN. On the other hand, the power SS-based DNR using Eq. (4) achieved a marked improvement in the speech recognition performance compared with that of CMN. The GSS-based DNR using Eq. (6) improved speech recognition performance significantly compared to both the CMN method and the power SS-based DNR for almost all reverberant conditions.

4.3. Experimental results for the real noisy reverberant speech

Table 6 shows the speech recognition results for the real noisy reverberant speech under the same conditions as the simulated noisy reverberant speech. The word accuracy rate for close-talking speech recorded in a real environment was 88.3%. We investigated the best channel combination in the real environment and the best speech recognition performance was obtained when channels 6, 7, 8, and 9 described in Fig. 2 were used. Therefore, this channel combination was used in this study. Power SS-based DN and GSS-based DN achieved a smaller improvement in recognition performance compared with the simulated noisy reverberant speech because the type of background noise in the real environment was different from that in the simulated noisy reverberant speech. On the other hand, the power SS-based DNR markedly improved the speech recognition performance compared to CMN. The GSS-based DNR improved speech recognition performance significantly compared to both the CMN method and the power SS-based DNR for almost all speakers. The GSS-based DNR achieved an average relative word error reduction rate of 39.1% and 11.5% compared to conventional CMN and power SS-based DNR, respectively. These results show that our proposed method is also effective in a real environment under the same denoising and dereverberation conditions as the simulated noisy reverberant speech.

Table 6: Word accuracy for LVCSR with the real noisy reverberant speech (%). Delay-and-sum beamforming was performed for all methods.

Speakers / Position	CMN only	Power SS		GSS	
		DN	DNR	DN	DNR
A	60.2	67.7	78.9	64.7	79.5
B	75.6	72.2	78.5	72.5	83.2
C	67.4	63.2	69.4	66.7	77.5
D	59.1	53.9	74.9	60.8	78.7
E	42.9	51.0	62.8	50.0	61.7
Average	60.9	61.6	73.1	62.9	76.2

5. Conclusion and future work

Previously, we proposed a blind dereverberation method based on power SS and GSS. The results of an LVCSR task showed that these methods achieved significant improvements over conventional CMN in a simulated reverberant environment without additive noise. In this paper, we presented a denoising and dereverberation method based on power SS or GSS and evaluated it in a real noisy reverberant environment. The GSS-based method achieved an average relative word error reduction rate of 39.1% and 11.5% compared to conventional CMN and the power SS-based method, respectively. These results show that our proposed method is also effective in a real noisy reverberant environment.

In the future, we intend to extend our proposed method to deal with real-world speech data including overlapping speech that involves multiple persons speaking simultaneously.

6. References

[1] L. Wang, N. Kitaoka and S. Nakagawa, "Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm, " IEICE Trans. Information and Systems, Vol.E94-D, No.3, pp. 659-667, Mar. 2011.

[2] Y. Huang, J. Benesty and J. Chen, "Acoustic MIMO Signal Processing," Springer, 2006.

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

[4] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada, "Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition, " Proc. of LREC2000, pp. 965-968, May, 2000.

[5] M. Nakayama et al., "CENSREC-4: Development of Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments, " Proc. of INTERSPEECH-2008, pp. 968-971, Sep. 2008.

[6] 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.