Blind Dereverberation Based on Generalized Spectral Subtraction by Multi-channel LMS Algorithm

Kyohei Odani*, Longbiao Wang* and Atsuhiko Kai*
* Shizuoka University, Japan
E-mail: odani@spa.sys.eng.shizuoka.ac.jp,{wang,kai}@sys.eng.shizuoka.ac.jp

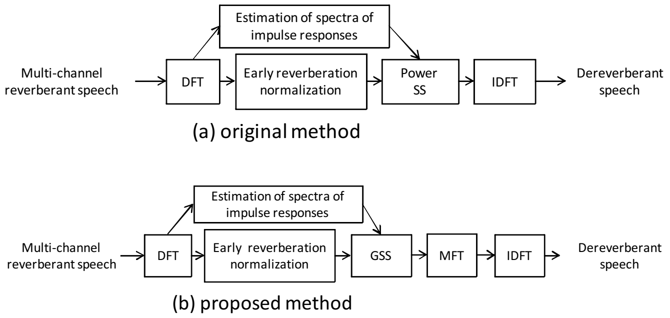
***Abstract*—A blind dereverberation method based on power spectral subtraction (SS) using a multi-channel least mean squares algorithm was previously proposed. The results of isolated word speech recognition experiments showed that this method achieved significant improvement over conventional cepstral mean normalization (CMN). In this paper, we propose a blind dereverberation method based on generalized spectral subtraction (GSS), which has been shown to be effective for noise reduction, instead of power SS. Furthermore, we extend the missing feature theory (MFT), which was initially proposed to enhance the robustness of additive noise, to dereverberation. The reliability of each spectral component is calculated through the signal-to-reverberation ratio obtained from the spectrum of dereverberant speech based on GSS. The proposed dereverberation method based on GSS with MFT is evaluated on a large vocabulary continuous speech recognition task. The dereverberation method based on GSS with MFT and beamforming achieves a relative word error reduction rate of 11.4% and 32.6% compared to the dereverberation method based on power SS with beamforming and the conventional CMN with beamforming, respectively.**

I. Introduction

In a distant-talking environment, channel distortion drastically degrades speech recognition performance due to a mismatch between the training and testing environments. Compensating an input feature is the main method for reducing the mismatch. Cepstral mean normalization (CMN) has been especially employed as a simple and effective way of normalizing the cepstral feature to reduce channel distortion. However, the impulse response of reverberation in a distant-talking environment usually has a much longer tail than the window length of the short-term spectral analysis. Therefore, conventional CMN is not totally effective under these conditions. Several studies have focused on mitigating this problem. Raut et al. [2] used preceding hidden Markov model (HMM) states as units of preceding speech segments, and they adapted models accordingly by estimating their contributions to the current state using a maximum likelihood function. However, model adaptation using *a priori* training data makes the models less practical to use because the true impulse response or matched reverberant utterance is not always as expected in various environments. A reverberation compensation method for speaker recognition using spectral subtraction, in which late reverberation is treated as additive noise, was proposed in [3]. However, the drawback of this approach is that the optimum parameters for spectral subtraction are empirically estimated from a development dataset and the late reverberation cannot be subtracted correctly as it is not modeled precisely.

In previous work [1], we proposed a robust distant-talking speech recognition method based on power spectral subtraction (SS) employing the adaptive multi-channel least mean squares (MCLMS) algorithm (see Fig. 1(a)). We treated the late reverberation as additive noise, and a noise reduction technique based on power SS was proposed to estimate the power spectrum of the clean speech using an estimated power spectrum of the impulse response. To estimate the power spectra of the impulse responses, we extended the variable step-size unconstrained MCLMS (VSS-UMCLMS) algorithm for identifying the impulse responses in a time domain [7] to a frequency domain. The early reverberation was normalized by CMN. By combining the proposed method with beamforming, a relative error reduction rate of 24.5% compared to the conventional CMN with beamforming was achieved on an isolated word recognition task.

Power spectral subtraction is the most commonly used spectral subtraction method. A previous study has shown that generalized SS (GSS) with a lower exponent parameter is more effective than power spectral subtraction for noise reduction [4]. In this paper, instead of using power SS, GSS is employed to suppress late reverberation. We also investigate the use of missing feature theory (MFT) [5] to enhance the robustness to noise, in combination with GSS, since the reverberation cannot be suppressed completely owing to the estimation error of the impulse response. Soft-mask estimation based MFT calculates the reliability of each spectral component from the signal-to-noise ratio (SNR). This idea is applied to reverberant speech. However, the reliability estimation is complicated in a distant-talking environment. In [6], reliability is estimated from the time lag between the power spectrum of the clean speech and that of the distorted speech. In this paper, reliability is estimated by the signal-to-reverberation ratio (SRR) since the power spectra of clean speech and the reverberation signal can be estimated by power SS or GSS using MCLMS. A diagram of the modified proposed method combining GSS with MFT is shown in Fig.1(b).



Estimation of spectra of impulse responses
Multi-channel reverberant speech
DFT
Early reverberation normalization
Power SS
IDFT
Dereverberant speech
Estimation of spectra of impulse responses
Multi-channel reverberant speech
DFT
Early reverberation normalization
GSS
MFT
IDFT
Dereverberant speech

(a) original method

(b) proposed method

Fig. 1. Schematic diagram of blind dereverberation methods.

II. Outline of Blind Dereverberation

A. Dereverberation based on power spectral subtraction

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

$$x[t] = h[t] * s[t] + n[t]. \tag{1}$$

In this paper, additive noise is ignored for simplification, so (1) becomes $x[t] = h[t] * s[t]$.

If the length of the impulse response is much smaller than the size T of the analysis window used for 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 STFT of the distorted speech is usually approximated by

$$\begin{aligned} X(f,\omega) &\approx S(f,\omega) * H(\omega) \\ &= S(f,\omega)H(0,\omega) + \sum_{d=1}^{D-1} S(f-d,\omega)H(d,\omega), \end{aligned} \tag{2}$$

where f is the frame index, $H(\omega)$ is the STFT of the impulse response, $S(f,\omega)$ is the STFT of clean speech s, and $H(d,\omega)$ denotes the part of $H(\omega)$ corresponding to the frame delay d. That is, with a long impulse response, the channel distortion is no longer of a multiplicative nature in a linear spectral domain, but is rather convolutional [2].

In [1], we proposed a dereverberation method based on power spectral subtraction to estimate the STFT of the clean speech $\hat{S}(f,\omega)$ based on (2). The spectrum of the impulse response for the spectral subtraction is blindly estimated using the method described in Section II-C. Furthermore, we compensate the early reverberation by subtracting the cepstral mean of the utterance. If we assume that phases of different frames are noncorrelated for simplification, the power spectrum of (2) can be approximated as [1]

$$|\tilde{X}(f,\omega)|^2 = \frac{|X(f,\omega)|^2}{|\bar{X}(f,\omega)|^2} \approx |\tilde{S}(f,\omega)|^2 + \frac{\sum_{d=1}^{D-1}\{|\tilde{S}(f-d,\omega)|^2|H(d,\omega)|^2\}}{|H(0,\omega)|^2}, \tag{3}$$

where $|\tilde{S}(f,\omega)|^2 = \frac{|S(f,\omega)|^2}{|\bar{S}(f,\omega)|^2}$, $|\tilde{X}(f,\omega)|^2$ is the power spectrum of reverberant speech after early reverberation normalization and $\bar{S}(f,\omega)$ is the mean vector of $S(f,\omega)$. The power spectrum of clean speech $|\hat{S}(f,\omega)|^2$ can be estimated as

$$\begin{aligned} &|\hat{S}(f,\omega)|^2 \approx max\{|\tilde{X}(f,\omega)|^2 - \\ &\alpha \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{S}(f-d,\omega)|^2|H(d,\omega)|^2\}}{|H(0,\omega)|^2}, \beta \cdot |\tilde{X}(f,\omega)|^2\}, \end{aligned} \tag{4}$$

where α is the noise over estimation factor, β is the spectral floor parameter to avoid negative or under flow values, and $H(d,\omega), d = 0, 1...D-1$ is the STFT of the impulse response, which can either be calculated from a known impulse response or be blindly estimated. D is the number of reverberation windows.

B. Dereverberation based on generalized spectral subtraction

Previous studies have showed that GSS with an arbitrary exponent parameter is more effective than power SS for noise reduction. In this paper, we extend GSS to suppress late reverberation. Instead of the power SS based dereverberation given in (4), GSS based dereverberation is modified as

$$\begin{aligned} &|\hat{S}(f,\omega)|^{2n} \approx max\{|\tilde{X}(f,\omega)|^{2n} - \\ &\alpha \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{S}(f-d,\omega)|^{2n}|H(d,\omega)|^{2n}\}}{|H(0,\omega)|^{2n}}, \beta \cdot |\tilde{X}(f,\omega)|^{2n}\}, \end{aligned} \tag{5}$$

where n is the exponent parameter. For power SS, the exponent parameter n is equal to 1. In this paper, the exponent parameter n is set to 0.1 as this value yielded the best results in [4].

The methods given in (4) and (5) are referred to as *SS-based (original)* and *GSS-based (proposed) dereverberation methods*, respectively.

C. Compensation parameter estimation for spectral subtraction by multi-channel LMS algorithm

In [7], an adaptive multi-channel LMS algorithm for blind Single-Input Multiple-Output (SIMO) system identification was proposed.

In the absence of additive noise, we can take advantage of the fact that

$$x_i * h_j = s * h_i * h_j = x_j * h_i, \;\; i, j = 1, 2, \cdots, N, i \neq j, \tag{6}$$

and have the following relations at time t:

$$\mathbf{x}_n(t) = [x_n(t) \; x_n(t-1) \; ... \; x_n(t-L+1)]^T, \tag{7}$$

$$\mathbf{h}_n(t) = [h_n(t,0) \; h_n(t,1) \; ... \; h_n(t,L-1)]^T, \tag{8}$$

$$n = 1, 2, ..., N,$$

where n is the channel index, $\mathbf{x}_n(t)$ is the speech signal received at time t, $\mathbf{h}_n(t)$ is the impulse response at time t, $h_n(t,l)$ is the *l-th* tap of the impulse response at time t, and L is the number of taps of the impulse response.

An estimated error vector at time t is expressed as

$$e_{ij}(t+1) = \mathbf{x}_i^T(t+1)\mathbf{h}_j(t) - \mathbf{x}_j^T(t+1)\mathbf{h}_i(t), \tag{9}$$

$$i, j = 1, 2, ..., N, i \neq j.$$

This error can be used to define a cost function at time t

$$\mathbf{J}(t+1) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_{ij}^2(t+1). \tag{10}$$

By minimizing the cost function J of (10), the impulse response can be blindly derived. We extended this VSS-UMCLMS algorithm [7], which identifies the multi-channel impulse responses, for processing in a frequency domain with SS applied in combination [1].

III. Missing Feature Theory for Dereverberation

Missing feature theory (MFT) [5] enhances the robustness of speech recognition to noise by rejecting unreliable acoustic features using a missing feature mask (MFM). The MFM is the reliability corresponding to each spectral component, with 0 and 1 being unreliable and reliable, respectively. The MFM is typically a hard and a soft mask. The hard mask applies binary reliability values of 0 or 1 to each spectral component and is generated using the signal-to-noise ratio (SNR). The reliability is 0 when the SNR is greater than a manually-defined threshold, otherwise it is 1. The soft mask is considered a better approach than the hard mask and applies a continuous value between 0 and 1 using a sigmoid function.

In a distant-talking environment, it is difficult to estimate the reliability of each spectral component since it is difficult to estimate the spectral components of clean speech and reverberant speech. Therefore, in [6], the reliability was estimated from $a\ priori$ information by measuring the difference between the spectral components of clean speech and reverberant speech at given times. In this paper, a soft mask is calculated using the SRR. From (5) the SRR is calculated as

$$SRR(f,\omega) = 10\log_{10}\left(\frac{|\hat{S}(f,\omega)|^{2n}}{\sum_{d=1}^{D-1}\{|\tilde{S}(f-d,\omega)|^{2n}|H(d,\omega)|^{2n}\}}\right). \quad (11)$$

The reliability $r(f,\omega)$ for the soft mask is generated as

$$r(f,\omega) = \frac{1}{1+exp(-a(SRR(f,\omega)-b))}, \quad (12)$$

where a and b are the gradient and center of the sigmoid function, respectively, and are empirically determined. Finally, the estimated spectrum of clean speech from (5) is multiplied by the reliability $r(f,\omega)$.

IV. Experiments

A. Experimental setup

Multi-channel distorted speech signals simulated by convolving multi-channel impulse responses with clean speech were used to evaluate our proposed algorithm. Seven kinds of multi-channel impulse responses measured in various acoustical reverberant environments were selected from the Real World Computing Partnership (RWCP) sound scene database [8] and the CENSREC-4 database [9]. Table I lists the conditions for the seven recordings using a two-channel microphone array. For the RWCP database, two-channel microphones were taken from a circular microphone array (16 channels), with the two microphones located at 5.85 cm intervals. Impulse responses were measured at several positions 2 m from the microphone array. For the CENSREC-4 database, two-channel microphones were taken from a linear microphone array (7 channels) with the two microphones located at 2.125 cm intervals. Impulse responses were measured at several positions 0.5 m from the microphone array. The Japanese Newspaper Article Sentences (JNAS) corpus was used as clean speech. 100 utterances from the JNAS database convolved with the multi-channel impulse responses shown in Table I were used as test data. The average time for all utterances was about 5.8 s.

TABLE I
Details of recording conditions for impulse response measurement. "RT60 (second)": reverberation time in room. "S": small, "L": large.

array no	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

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 spectral subtraction based dereverberation.

analysis window	Hamming
window length	32 ms
window shift	16 ms
number of reverberant windows D	6 (192 ms)
noise overestimation factor α	1.0 (Power SS) 0.1 (GSS)
spectral floor parameter β	0.15 (both)
soft mask gradient parameter a	0.05 (Power SS) 0.01 (GSS)
soft mask center parameter b	0.0 (both)

Table II gives the conditions for speech recognition. The acoustic models were trained with the ASJ speech databases of phonetically balanced sentences (ASJ-PB) and the JNAS. In total, around 20K sentences (clean speech) uttered by 132 speakers were used for each gender. Table III gives the conditions for spectral subtraction based dereverberation. The parameters shown in Table III were determined empirically. For the proposed dereverberation method based on spectral subtraction, the previous clean power spectra estimated with a skip window were used to estimate the current clean power spectrum since the frame shift was half the frame length in this study. The spectrum of the impulse response $H(d,\omega)$ was estimated for each utterance to be recognized. The word accuracy rate for large-vocabulary continuous speech recognition (LVCSR) with clean speech was 92.59%.

B. Experimental results and discussion

In both our SS-based and GSS-based dereverberation methods, speech signals from two microphones were used to estimate blindly the compensation parameters for the power SS and GSS (that is, the spectra of the channel impulse

TABLE IV
WORD ACCURACY RATES FOR LVCSR (%). DELAY-AND-SUM BEAMFORMING WAS PERFORMED FOR ALL METHODS.

Distorted	CMN	Power SS		GSS (proposed)	
Speech #	only	w/o MFT	MFT	w/o MFT	MFT
1	44.35	63.34	65.15	65.95	66.47
2	27.59	40.79	44.03	49.16	47.56
3	25.61	42.55	45.75	49.29	48.31
4	73.90	79.26	78.17	80.77	80.96
5	27.06	42.28	44.91	45.38	47.83
6	29.62	50.78	54.60	56.13	58.87
7	65.24	71.67	68.31	74.35	75.93
Ave.	41.91	55.81	57.27	60.15	60.85

TABLE V
BREAKDOWN OF SPEECH RECOGNITION ERRORS (%).

	CMN	Power SS		GSS (proposed)	
	only	w/o MFT	MFT	w/o MFT	MFT
Sub	40.61	30.48	29.37	27.39	27.42
Del	13.82	9.27	9.26	8.99	8.06
Ins	3.67	4.44	4.10	3.47	3.67

responses), and then reverberation was suppressed by SS and the spectrum of dereverberant speech was inverted into a time domain. Finally, delay-and-sum beamforming was performed on the two-channel dereverberant speech. The schematic of dereverberation is shown in Fig. 1.

Table IV shows the speech recognition results for the original and proposed methods. "Distorted speech #" in Table IV corresponds to "array no" in Table I. The word accuracy rate by CMN without beamforming was 40.46%. The speech recognition performance was drastically degraded under reverberant conditions because the conventional CMN did not suppress the late reverberation. Delay-and-sum beamforming with CMN (41.91%) could not improve the speech recognition performance markedly because of the small number of microphones and the small distance between the microphone pair. On the other hand, the power SS based dereverberation using (4) markedly improved the speech recognition performance. The GSS-based dereverberation using (5) improved speech recognition performance significantly compared with the original proposed (power SS based dereverberation) method and CMN for all reverberant conditions. The GSS-based method without MFT achieved an average relative word error reduction rate of 31.4% compared to the conventional CMN and 9.8% compared to the power SS-based method without MFT. When MFT was combined with both our methods, a further improvement was achieved. Finally, the GSS-based method with MFT achieved an average relative word error reduction rate of 32.6% compared to conventional CMN and 11.4% compared to the original proposed method [1].

Table V gives a breakdown of the word error rates obtained by the power SS- and GSS-based methods. The power SS-based method improved the substitution and deletion error rates, but degraded the insertion error rate compared with CMN. The GSS-based method improved all error rates compared with the power SS-based method, and achieved almost the same word insertion error as CMN.

V. CONCLUSIONS AND FUTURE WORK

Previously [1], we proposed a blind dereverberation method based on power SS employing the multi-channel LMS algorithm for distant-talking speech recognition. Previous studies showed that GSS with an arbitrary exponent parameter is more effective than power SS for noise reduction. In this paper, instead of power SS, GSS is applied to suppress late reverberation. However, reverberation cannot be completely suppressed owing to the estimation error of the impulse response. MFT is used to enhance the robustness of noise. Soft-mask estimation based MFT calculates the reliability of each spectral component from SNR. In this paper, reliability was estimated through the signal-to-reverberation ratio. Furthermore, delay-and-sum beamforming was also applied to the multi-channel speech compensated by the reverberation compensation method. Our SS and GSS-based dereverberation methods were evaluated using distorted speech signals simulated by convolving multi-channel impulse responses with clean speech. The GSS-based method without MFT achieved an average relative word error reduction rate of 31.4% compared to conventional CMN and 9.8% compared to the power SS-based method without MFT. When MFT was combined with both our methods, further improvement was obtained. The GSS-based method with MFT achieved average relative word error reduction rates of 32.6% and 11.4% compared to conventional CMN and the original proposed method, respectively.

So far, additive noise has been ignored in our study for the sake of simplicity, but background noise cannot be ignored in a real environment. In the future, we will attempt to extend our proposed methods to real-world speech data simultaneously degraded by additive noise and convolutional noise.

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] C. Raut, T. Nishimoto, S. Sagayama, "Adaptation for long convolutional distortion by maximum likelihood based state filtering approach, " Proc. of ICASSP-2006, Vol. 1, pp. 1133-1136, 2006.

[3] Q. Jin, T. Schultz and A. Waibel, "Far-field speaker recognition," IEEE Trans. ASLP, Vol. 15, No. 7, pp. 2023-2032, 2007.

[4] B. L. Sim, Y. C. Tong, J. S. Chang and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method, " IEEE Trans. on Speech and Audio Processing, vol.6, no.4, pp. 328-337, 1998.

[5] Bhiksha Raj and Richard M. Stern, "Missing-Feature Approaches in Speech Recognition," IEEE Signal Processing Magazine, pp. 101-116, Sep. 2005.

[6] Kalle J. Palomaki, Guy J. Brown and Jon Barker, "Missing Data Speech Recognition in Reverberant Conditions, " Proc. of ICASSP-2002, pp. 65-68, 2002.

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

[8] 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.

[9] T. Nishiura et al., "Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments, " Proc. of INTERSPEECH-2008, pp. 968-971, Sep. 2008.