Dereverberantion based on Generalized Spectral Subtraction for Distant-talking Speaker Recognition

Zhaofeng Zhang*, Longbiao Wang† and Atsuhiko Kai*
* Shizuoka University, Japan
E-mail: zhang@spa.sys.eng.shizuoka.ac.jp,kai@sys.eng.shizuoka.ac.jp
† Nagaoka University of Technology, Japan
E-mail: wang@vos.nagaokaut.ac.jp

***Abstract*—A dereverberation method based on generalized spectral subtraction (GSS) using multi-channel least mean square (MCLMS) was proposed previously. The results on speech recognition experiments showed that this method achieved a significant improvement compare to the conventional methods. In this paper, we employ this method to distant-talking speaker recognition. However, the GSS-based dereverberation method using clean speech models degrades the speaker recognition performance while it is very effective for speech recognition. One of the reason may be that the GSS-based dereverberation method causes some distortions such as speaker characteristics distortion between clean speech and dereverberant speech. In this paper, we address this problem by training speaker models using dereverberant speech which is obtained by suppressing reverberation from arbitrary artificial reverberant speech. The speaker recognition experiment was performed on a large scale far-field speech with different reverberant environments to the training environments. The proposed method achieved a relative error reduction rate of 88.2% compared to conventional CMN with beamforming using clean speech models and 32.8% compared to reverberant speech models, respectively.**

I. Introduction

Because of existing of reverberation in distant-talking environments, distant-talking speaker recognition performance has been drastically degraded. Most dereverberation techniques have been employed through signal processing to compensating an input signal. Beamforming is one of the simplest and the most robust means of spatial filtering to suppress reverberation and background noise, which can discriminate between signals based on the physical locations of the signal sources [1]. The other general approach is cepstral mean normalization (CMN) [2], has been extensively examined as a simple and effective way of reducing reverberation with normalizing the cepstral features. Unfortunately, the impulse response of reverberation in a distant-talking environment usually has a duration which is much longer than analysis window size of short-term spectral analysis. Therefore, the performance of dereverberation is not completely effective by CMN in this environment. 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.

Previously, Wang et al. presented a distant-talking speech recognition method based on generalized spectral subtraction (GSS) employing the multi-channel LMS (MCLMS) algorithm [4]. They treated last reverberation as additive noise, and a noise reduction technique based on GSS [5] was proposed to estimate the spectrum of the clean speech using an estimated spectrum of the impulse response. To estimate the spectra of the impulse responses, they extended the variable step-size unconstrained MCLMS algorithm for identifying the impulse responses in a time domain [6] to a frequency domain. The early reverberation was normalized by CMN. The experiment on speech recognition showed that the GSS-based dereverberation method achieved an average relative word error reduction rate of 32.6% compared to conventional CMN with beamforming.

The GSS-based dereverberation was used in speech recognition filed in previous study [4]. However, the effect of GSS-based dereverberation on distant-talking speaker recognition is still unknown. A preliminary experiment of speaker recognition with GSS-based method was performed. The result showed that the GSS-based dereverberation using clean speech models degraded the speaker recognition performance while it was very effective for speech recognition. One of the reason may be that the GSS-based dereverberation method causes some distortions such as speaker characteristics distortion between clean speech and dereverberant speech. We address this problem by training speaker models using dereverberant speech which is obtained by suppressing early and late reverberation from arbitrary artificial reverberant speech. The speaker characteristics distortion in training data and test data is similar, so the GSS-based dereverberation method is expected to be effective for speaker recognition.

II. Distant-talking speaker recognition system employing dereverberation method

The performance of distant-talking speech/speaker recognition is degraded remarkably by reverberation. Through removing reverberation, we expect to improve the speech/speaker recognition performance. However, very few researches have studied the different on the speech recognition and the speaker recognition in a distant-talking environment. The required characteristics for acoustic features in speech recognition is to maximize the inter-phoneme variation while minimizing the intra-phoneme variation in the feature space, and for speaker variation instead of phoneme variation in speaker recognition. This appearance lead some methods which are effective in speech recognition may be not effective in speaker recognition field especially on hands-free environment. For example, a simple and popular channel normalization method, CMN, removes both the transmission characteristics and speaker characteristics, so the trend of the speaker recognition performance and speech recognition performance is different in some cases. Previous study [7] on distant-talking speaker recognition showed that the results with the conventional CMN was much worse than that without CMN while it was very effective for speech recognition. The reason was that CMN removed the speaker characteristics and the channel distortion (reverberation) was not very large. In speech recognition field, the GSS-based dereverberation using clean speech models obtained a significant improvement [4]. However in speaker recognition field, the experiment we performed in Section IV shows that it degrades the speaker recognition performance. It may be the GSS-based dereverberation method cause some speaker characteristics distortion between clean speech and dereverberant speech.

To mitigate the speaker characteristics distortion caused by dereverberation in test stage, dereverberant speech which is obtained by suppressing early and late reverberation from arbitrary artificial reverberant speech is used to train speaker models. We assume that the speaker characteristics distortion in training data and test data is similar. By employing dereverberation in both training stage and test stage, the transmission characteristics can be removed and the relative speaker characteristics can be keep maximally. Comparing with speaker models trained with reverberant speech, our method is expected to have a better speaker recognition performance. In the previous research, GMMs trained with reverberant speech have been used in distant-talking speaker recognition. However, the mismatch of distant-talking environments between the training condition and the test condition has still not been addressed. Furthermore, when the late reverberations have a large energy, the performance of speaker recognition cannot be improved even with GMMs trained with a matched reverberant condition [8]. It means that the GMMs cannot handle severe late reverberations precisely.

In this paper, a distant-talking speaker recognition system employing dereverberation method was proposed. The schematic diagram of our proposed method is shown in Fig 1. In training stage, clean speech is convoluted by arbitrary impulse responses to create artificial reverberant speech, which can reduce the experimental cost because a real reverberant speech is not necessary. Then a GSS-based dereverberation which will be introduced in Section III is performed to suppress both the early and late reverberation. Finally, the dereverberant speech is used to train speaker models. In test stage, the reverberation of multi-channel distorted speech [1] is removed by the GSS-based dereverberation method, and then the dereverberant speech is used to perform distant-talking speaker recognition.

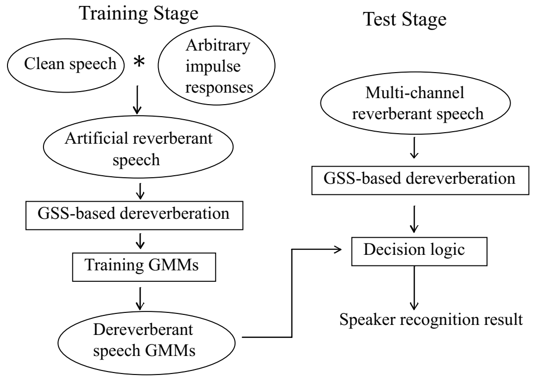
Training Stage
Test Stage
Clean speech
*
Arbitrary impulse responses
Multi-channel reverberant speech
Artificial reverberant speech
GSS-based dereverberation
GSS-based dereverberation
Training GMMs
Decision logic
Dereverberant speech GMMs
Speaker recognition result

Fig. 1. Schematic diagram of distant-talking speaker recognition system

III. Outline of Dereverberation Based on GSS

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

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

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.

In [4], Wang et al. proposed a dereverberation method based on generalized spectral subtraction to estimate the STFT of the clean speech $\hat{S}(f,\omega)$ based on Eq. (2). Assuming that phases of different frames are noncorrelated for simplification, the power spectrum of Eq. (2) can be approximated as Eq. (3)

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

The spectral subtraction is used to suppress the late reverberation, and the early reverberation is compensated by subtracting the cepstral mean of the utterance at the stage of feature extraction. The spectrum $|\hat{X}(f,\omega)|^{2n}$ obtained by reducing the late reverberation can be estimated as

$$|\hat{X}(f,\omega)|^{2n} \approx max\left\{|X(f,\omega)|^{2n} - \alpha \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{X}(f-d,\omega)|^{2n}|\hat{H}(d,\omega)|^{2n}\}}{|\hat{H}(0,\omega)|^{2n}}, \beta\cdot|X(f,\omega)|^{2n}\right\}. \tag{4}$$

[1] artificial reverberant speech or real reverberant speech

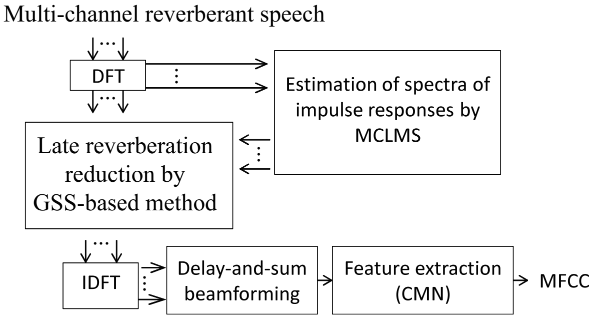
Multi-channel reverberant speech
DFT
Estimation of spectra of impulse responses by MCLMS
Late reverberation reduction by GSS-based method
IDFT
Delay-and-sum beamforming
Feature extraction (CMN)
MFCC

Fig. 2. Schematic diagram of dereverberation method

where α is the noise over estimation factor, β is the spectral floor parameter to avoid negative or under flow values, $|\hat{X}(f,\omega)|^{2n} = |\hat{S}(f,\omega)|^{2n}|\hat{H}(0,\omega)|^{2n}$, $|\hat{S}(f,\omega)|^{2n}$ is the spectrum of estimated clean speech and $\hat{H}(d,\omega), d = 0, 1...D-1$ is the STFT of the impulse response, which can be blindly estimated by the multi-channel LMS algorithm method mentioned in [4]. D is the number of reverberation windows. n is the exponent parameter. In this paper, the exponent parameter n is set to 0.5 empirically.

The schematic diagram of our proposed GSS-based dereverberation method is shown in Fig. 2. GSS-based method uses the spectra of impulse responses which are estimated by MCLMS algorithm to reduce the late reverberation from reverberant speech. Then the spectrum of dereverberant speech is inverted into time domain and delay-and-sum bemaforming is performed on mutil-channel speech. At last, the early reverberation is normalized by CMN at the feature extraction stage.

IV. Experiments

A. Experimental Setup

The proposed method for distant-talking speaker identification was evaluated in artificial reverberant speech for the sake of convenience[2]. Eight kinds of multi-channel impulse responses were selected from the Real World Computing Partnership (RWCP) sound scene database [9] and the CENSREC-4 database [10], which were convoluted with clean speech to create artificial reverberant speech. A large scale database, Japanese Newspaper Article Sentence (JNAS) [11] corpus, was used as clean speech. The utterances of training data is composed of both 130 male and female speakers with 10 utterances taken from each one. Each speaker has 20 utterances for test.

Table I lists the impluse responses for training set and test set. For the RWCP database, a 4 channel circular or linear microphone array was taken from a circular + linear microphone array (30 channels). The circle type microphone array had a diameter of 30 cm. The microphone of the linear microphone array was located at 2.83 cm intervals. Impulse responses were measured at several positions 2 m from the microphone array. For the CENSREC-4 database, 4 channel microphones were taken from a linear microphone array (7 channels) with the microphones located at 2.125 cm intervals. Impulse responses were measured at several positions 0.5 m from the microphone array.

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

array no	room	mic type	RT60(s)
(a) CENSREC-4 database for training			
1	Japanese style room	linear	0.40
2	Japanese style bath	linear	0.60
3	elevator hall	linear	0.75
(b) RWCP database for test			
4	echo room (cylinder)	circle	0.38
5	tatami-floored room (S)	circle	0.47
6	tatami-floored room (L)	circle	0.60
7	conference room	circle	0.78
8	echo room (panel)	linear	1.30

TABLE II
Conditions for speaker recognition.

sampling frequency	16 kHz
frame length	25 ms
frame shift	10 ms
feature space	25 dimensions with CMN (12 MFCCs + Δ + Δpower)
acoustic model	GMMs with 128 diagonal covariance matrices

TABLE III
Conditions for GSS-based dereverberation.

analysis window	Hamming
window length	32 ms
window shift	16 ms
number of reverberant windows D	6 (192 ms)
spectral floor parameter β	0.15
noise over estimation factor a	0.5
exponent parameter n	0.5

Table II gives the conditions for speaker recognition. 25-dimension MFCCs and GMMs with 128 mixtures were used. Table III gives the conditions for GSS-based dereverberation. The parameters shown in Table III were determined empirically.

TABLE IV
The description of each speaker recognition method.

Method #	Speaker models	Processing at test stage
1 (Baseline)	Clean speech models	CMN with beamforming
2 (Method in [4])	Clean speech models	GSS-based dereverberation
3	Reverberant speech models	CMN with beamforming
4 (Proposed method)	Dereverberant speech models	GSS-based dereverberation

[2] For real reverberant speech, the processing step is same as the artificial reverberant speech, and it is considered that a similar trend should be obtained comparing with the artificial reverberant speech.

TABLE V
DISTANT-TALKING SPEAKER RECOGNITION RATES (%)

Method #	# of impulse response condition for test					Ave.
	4	5	6	7	8	
1	66.7	53.3	43.2	43.7	38.3	49.0
2	53.1	32.9	25.6	25.3	29.1	33.2
3	91.6	88.4	86.5	87.6	88.0	88.4
4	96.2	91.1	91.1	90.1	92.3	92.2

Four kinds of methods were compared in this study. The description of these methods are shown in Table IV. For all these methods, CMN with delay-and-sum beamforming was performed. Clean speech models which were training by clean speech directly were used as speaker models for *method 1* and *method 2*. For *method 1*, only CMN with beamforming was used to reduce the reverberation. The GSS-based dereverberation was performed at test stage for *method 2*, which is the same as the condition for distant-talking speech recognition [4]. Reverberant speech models which were training by artificial reverberant speech using 3 kinds of CENSREC-4 impulse responses (see Table I(a)) were used as speaker models for *method 3*. *Method 4* is our proposed method. For the proposed method, both of the reverberation of the training data and the test data were suppressed by GSS-based dereverberation, and the dereverberant speech was used to train dereverberant speech GMMs.

B. Experimental Results

The distant-talking speaker recognition results of 4 kinds of methods are compared in Table V. "# of impulse response condition for test" in Table V denotes the "array no" in Table I(b). In the previous research, the speech recognition results in reverberant environments with clean speech models were improved by using GSS-based dereverberation method [4]. However, *method 2* proposed in [4] degraded the speaker recognition performance in the speaker recognition filed. The reason for this phenomenon may be that the GSS-based dereverberation causes some speaker characteristics distortion as we discussed in Section II. The result of *method 3* which was based on reverberant speech models improved the speaker recognition performance significantly because multiple reverberant environments were trained. However, the reverberation was not suppressed, so a furthermore improvement could be expectation by employing blind dereverberation. The proposed method which suppressed the reverberation in both training data and test data outperformed than all of the other methods under all reverberant environments. The proposed method achieved a relative error reduction rate of 88.2% compared to the baseline (*method 1*) and 32.8% compared to reverberant speech models (*method 3*), respectively.

V. CONCLUSIONS AND FUTURE WORK

Previously, Wang et al. proposed a blind dereverberation method based on GSS employing the mutil-channel LMS algorithm for distant-talking speech recognition [4]. In this paper, we applied this method to distant-talking speaker recognition. However, in speaker recognition field, the method proposed in [4] worked worse than the baseline method, which was an opposite trend as the speech recognition has been shown. The speaker identification performance was degraded about 16% of recognition rate in this condition. One of the reason may be that the GSS-based dereverberation method causes some distortions between clean speech and dereverberant speech. We addressed this problem by training speaker models using dereverberant speech which was obtained by suppressing reverberation from arbitrary artificial reverberant speech, and the reverberant speech for test data was also compensated by GSS-based dereverberation. The proposed method based on dereverberant speech models achieved a relative error reduction rate of 88.2% compared to the conventional CMN with beamforming using clean speech models and 32.8% compared to reverberant speech models, respectively.

In this paper, the optimal parameters for GSS-based dereverberation were determined empirically. In the future, we will try to find the optimal parameters automatically which are selected by maximum likelihood among various parameters for GSS-based dereverberation. The combination of log likelihood of GMMs with different parameters for dereverberation is another simple way to address the above problem. Moreover, we will evaluate the speaker identification experiments with proposed method in a real environment.

REFERENCES

[1] T. B. Hughes, H. S. Kim, J. H. DiBiase and H. F. Silverman, "Performane of an an HMM Speech Recognizer Using a Real-time Tracking Microphone Array as Input," IEEE Trans. Speech, and Audio Processing, vol. 7, no. 3, pp. 346-349, May 1999.

[2] S. Furui, "Cepstral Analysis Technique for automatic speaker verification," IEEE Trans. Acoust. Speech Singnal Process., vol.29, no.2, pp.254-272, 1981.

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

[4] L. Wang, K. Odani and A. Kai, "Dereverberation and denoising based on generalized spectral subtraction by nutil-channel LMS algorithm using a small-scale microphone array," Eurasip Journal on Advances in Signal Processing 2012:12, Jan. 2012.

[5] 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.

[6] Y. Huang, J. Benesty, J. Chen, "Acoustic MIMO Signal Processing," Springer-Verlag, Berlin, 2006.

[7] L. Wang, N. Kitaoka and S. Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," Speech Communication, Vol. 49, No.6, pp. 501–513, June 2007.

[8] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with rasta-plp," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 1997, vol. 2, pp. 1259-1262.

[9] 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.

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

[11] K. Itou, M. Yamamoto, K. Takeda, T. Kakezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, "JNAS: Janpanese speech corpus for large vocabulary continuous speech recognition research, " J. Acoust Soc Jpn (E). 20(3), 199-206, 1999.