HANDS-FREE SPEAKER IDENTIFICATION BASED ON SPECTRAL SUBTRACTION USING A MULTI-CHANNEL LEAST MEAN SQUARE APPROACH

Longbiao Wang[1], Zhaofeng Zhang[2], Atsuhiko Kai[2]

[1]Nagaoka University of Technology, Japan
[2]Shizuoka University, Japan

wang@vos.nagaokaut.ac.jp, zhang@spa.sys.eng.shizuoka.ac.jp, kai@sys.eng.shizuoka.ac.jp

ABSTRACT

A dereverberation method based on generalized spectral subtraction (GSS) using a multi-channel least mean square (MCLMS) approach achieved significantly improved results on speech recognition experiments compared with conventional methods. In this study, we employ this method for hands-free speaker identification. The GSS-based dereverberation method using clean speech models degrades speaker identification performance, although it is very effective for speech recognition. One reason may be that the GSS-based dereverberation method causes distortion such as distortion characteristics between clean speech and dereverberant speech. In this study, we address this problem by training speaker models using dereverberant speech, which is obtained by suppressing reverberation from arbitrary artificial reverberant speech. We also propose a method that combines various compensation parameter sets to improve speaker identification and provide an efficient computational method. The speaker identification experiment was performed on large-scale far-field speech, with reverberant environments different to the training environments. The proposed method achieved a relative error reduction of 87.5%, compared with conventional cepstral mean normalization with beamforming using clean speech models, and 44.8% compared with reverberant speech models.

Index Terms— speaker identification, hands-free, dereverberation, spectral subtraction, multi-channel LMS

1. INTRODUCTION

Because of the existence of reverberation in hands-free environments, hands-free speaker identification performance is drastically degraded. Most dereverberation techniques are employed through signal processing to compensate an input signal. Beamforming is one of the simplest and 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], which has been extensively examined as a simple and effective way of reducing reverberation with normalized cepstral features. Unfortunately, the impulse response of reverberation in a hands-free environment usually has a much longer duration than the analysis window for short-term spectral analysis. Therefore, the performance of dereverberation is not completely effective when using CMN in this environment. A reverberation compensation method for speaker identification 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 hands-free speech recognition method based on generalized spectral subtraction (GSS) employing the multi-channel least mean square (MCLMS) algorithm [4]. They treated late 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 from the impulse response. To estimate the spectra of the impulse responses, they extended the variable step-size unconstrained MCLMS algorithm for transforming the impulse responses in the time domain [6] to the frequency domain. The early reverberation was normalized using CMN.

GSS-based dereverberation has been used in the speech recognition field in a previous study [4]. However, the effect of GSS-based dereverberation on hands-free speaker identification is still unknown. A preliminary experiment on speaker identification using the GSS-based method was performed. The results showed that the GSS-based dereverberation using clean speech models degraded the speaker identification performance, although it was very effective for speech recognition. One reason for this may be that the GSS-based dereverberation method causes distortion like the speaker ' s distortion characteristics between clean and dereverberant speech. We addressed this problem by training speaker models using dereverberant speech obtained by suppressing early and late reverberation from arbitrary artificial reverberant speech. The speaker ' s distortion characteristics in training and test data were similar, so the GSS-based dereverberation method was expected to be effective for speaker identification.

For GSS, it is difficult to determine the optimum compensation parameters (exponent parameter, noise over estimation factor etc.) under various environmental conditions. In this study, we combined multiple speaker identification results obtained using different compensation parameter sets, which meant that special tuning of GSS was not necessary for our proposed method. However, the computational cost increased linearly according to the number of compensation parameter sets. To reduce the computational time, only the speaker models with top N-best likelihood outcomes are rescored and combined to determine the target speaker.

2. HANDS-FREE SPEAKER IDENTIFICATION SYSTEM EMPLOYING THE DEREVERBERATION METHOD

To mitigate the speaker ' s distortion characteristics caused by dereverberation in the test stage, dereverberant speech obtained by suppressing early and late reverberation from arbitrary artificial reverberant speech was used to train the speaker models. We assumed

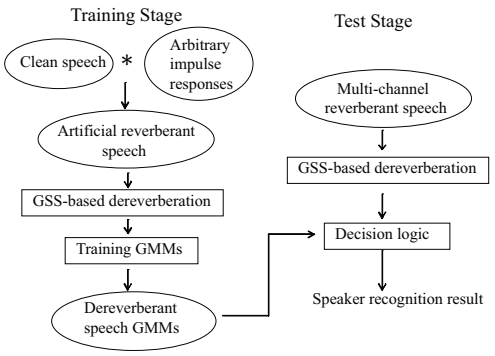

Fig. 1. Schematic diagram of the hands-free speaker identification system

that the speaker ' s distortion characteristics in training and test data were similar. By employing dereverberation in both the training and test stages, the transmission characteristics can be removed and the relative speaker ' s characteristics can remain maximized. Compared with speaker models trained with reverberant speech, our method was expected to have a better speaker identification performance. In previous research, GMMs trained with reverberant speech have been used in hands-free speaker identification. However, the mismatch of hands-free environments between the training and test conditions has still not been addressed. Furthermore, when late reverberations contain large energy, the performance of speaker identification cannot be improved even with GMMs trained with a matched reverberant condition. It means that the GMMs cannot handle severe late reverberations precisely.

In this study, a hands-free speaker identification system employing the dereverberation method was proposed. The schematic diagram of our proposed method is shown in Fig 1. In the training stage, clean speech was convoluted by arbitrary impulse responses to create artificial reverberant speech, which reduced the experimental cost because real reverberant speech was not necessary. A GSS-based dereverberation, which will be introduced in Section 3, was then performed to suppress both the early and late reverberation. Finally, the dereverberant speech was used to train the speaker models. In the test stage, the reverberation of multi-channel distorted speech [1] was removed by the GSS-based dereverberation method, and the dereverberant speech was then used to perform hands-free speaker identification.

3. OUTLINE OF DEREVERBERATION

3.1. 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 transforms (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 the linear spectral domain, but is 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). 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:

$$|\hat{X}(f,\omega)|^{2n} \approx max\left\{|X(f,\omega)|^{2n} - \alpha \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{X}(f-d,\omega)|^{2n}|\hat{H}(d,\omega)|^{2n}\}}{|\hat{H}(0,\omega)|^{2n}}, \beta \cdot |X(f,\omega)|^{2n}\right\}. \tag{3}$$

where α is the noise over the 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 the 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 using the multi-channel LMS algorithm method mentioned in Section 3.2. D and n are the number of reverberation windows and the exponent parameter.

3.2. Blind Estimation of Impulse Responses

In this section, we describe the blind estimation of the spectra of an impulse response $\hat{H}(d,\omega)$ using Eq. (3). In [6], MCLMS in a time domain was proposed to blindly estimate the impulse responses of each channel. In this study, we used a variable step-size unconstrained MCLMS (VSS-UMCLMS) algorithm extended in the time domain to the frequency domain.

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

$$\mathbf{R}_{X_iX_i}(\tau+1)\mathbf{H}_j(\tau) = \mathbf{R}_{X_iX_j}(\tau+1)\mathbf{H}_i(\tau) \quad i,j = 1,2,\cdots,N, i \neq j, \tag{4}$$

$$\mathbf{R}_{X_iX_j}(\tau) = E[\mathbf{X}_i(\tau)\mathbf{X}_j^T(\tau)], \tag{5}$$

$$\mathbf{X}_i(\tau) = [X_i(\tau), X_i(\tau-1), \cdots, X_i(\tau-D+1)]^T, \tag{6}$$

$$\mathbf{H}_i(\tau) = [H_i(\tau,0), \cdots, H_i(\tau,d), \cdots, H_i(\tau,D-1)]^T, \tag{7}$$

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

By summing up the N-1 cross correlations and some further calculation, the spectra of the impulse responses can be estimated blindly. For details of this approach refer to the literature [4] and [6].

[1] artificial reverberant speech or real reverberant speech

4. COMBINATION METHOD AND ITS EFFICIENT COMPUTATION

It is difficult to determine the optimum exponent parameter n and noise over the estimation factor α for GSS. In this study, a combination of the various likelihood speaker models with different compensation parameter sets is used.

When a combination of multiple methods is used to identify the speaker, the likelihood of speaker models with different compensation parameter sets is linearly coupled to produce a new score L^k_{comb} given by:

$$L^k_{comb} = \frac{1}{I}\sum_{i=1}^{I} L^k_i, \qquad k = 1, 2, \cdots, K, \tag{8}$$

where L^k_i is the likelihood produced by the *k-th* speaker model with the *i-th* compensation parameter set. K is the number of speakers registered and I denotes the number of compensation parameter sets. A speaker with the maximum likelihood is decided as the target speaker. By doing this, special tuning is not necessary for GSS.

However, the computational time is linearly increased according to the number of compensation parameter sets. In this study, an efficient computational method is proposed. Initially, the power SS (that is, compensation parameter $n = 1$) is used to suppress the reverberation and the likelihoods of all speaker models are calculated. Second, only the speaker models with the top N-best likelihoods are used to calculate a new likelihood according to different compensation parameter sets. Finally, the likelihood calculated by a different compensation parameter set is combined to decide the target speaker.

The total computational time T_A for speaker identification is about $T_F + T_L$, where T_F and T_L are the computational times for feature extraction and likelihood calculation of K speaker models. The computational time for the combination (that is, conventional combination method) of various results with I parameter set is $T^{comb}_A = I(T_F + T_L) = IT_A$. The computational time for our proposed efficient combination method using the N-best likelihood is:

$$\begin{aligned} T^{comb}_E &= T_F + T_L + (I-1)T_F + \frac{(I-1)N}{K}T_L \\ &= T_A + \frac{1}{\gamma+1}(\frac{(I-1)N}{K}\gamma + I - 1)T_A, \end{aligned} \tag{9}$$

where T_L equals γT_F. The computational cost decreased compared with the conventional combination method.

4.1. Experimental Setup

The proposed method for hands-free speaker identification was evaluated in artificial reverberant speech for the sake of convenience [2]. Eight types of multi-channel impulse responses were selected from the Real World Computing Partnership (RWCP) sound scene database [8] and the CENSREC-4 database [9], which were convoluted with clean speech to create artificial reverberant speech. A large-scale database, the Japanese Newspaper Article Sentence (JNAS) [10] corpus, was used as clean speech. The utterances of training data are composed of 130 male and female speakers with 10 utterances taken from each. Each speaker made 20 utterances for the test.

[2]For real reverberant speech, the processing step is the same as for artificial reverberant speech.

Table 1. 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 2. Conditions for speaker identification.

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 3. 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 1 lists the impulse responses for the training and test sets. For the RWCP database, a four-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 microphones of the linear microphone array were located at 2.83 cm intervals. Impulse responses were measured at several positions 2 m from the microphone array. For the CENSREC-4 database, four-channel microphones were taken from a linear microphone array (seven 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 2 gives the conditions for speaker identification. 25-dimension MFCCs and GMMs with 128 mixtures were used. Table 3 gives the conditions for GSS-based dereverberation. The parameters shown in Table 3 were determined empirically.

Four methods were compared in this study. The description of these methods are presented in Table 4. For all these methods, CMN with delay-and-sum beamforming was performed. Clean speech models, which were directly trained by clean speech, 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 the test stage for

Table 4. The description of each speaker identification 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

Table 5. Hands-free speaker identification 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	94.0	90.6	91.0	90.5	92.3	91.7

method 2, which is the same as the condition for hands-free speech recognition [4]. Reverberant speech models, which were trained using artificial reverberant speech with three types of CENSREC-4 impulse responses (see Table 1(a)), were used as speaker models for *method 3*. *Method 4* is our proposed method. For the proposed method, both the reverberation of the training and test data were suppressed by GSS-based dereverberation, and the dereverberant speech was used to train dereverberant speech GMMs.

4.2. Experimental Results

The hands-free speaker identification results for the four methods are compared in Table 5. "# of impulse response condition for test" in Table 5 denotes the "array no" in Table 1(b). In previous research, the speech recognition results in reverberant environments with clean speech models were improved when using the GSS-based dereverberation method [4]. However, *method 2* proposed in [4] degraded the speaker identification performance in the speaker identification field. The result of *method 3*, which was based on reverberant speech models, improved speaker recognition significantly because multiple reverberant environments were trained. However, the reverberation was not suppressed, so a further improvement could be the outcome of employing blind dereverberation. The proposed method without parameter tuning (that is, $\alpha = n = 1$), which suppressed the reverberation in both training and test data, outperformed all the other methods under all reverberant environments. The proposed method achieved a relative error reduction of 83.7% compared with the baseline (*method 1*) and 28.4% compared with reverberant speech models (*method 3*).

The performance of the proposed GSS-based dereverberation method may vary with different compensation parameters. We confirmed this and compared the performance with different parameters in Table 6. For GSS, most exponent parameters n are set from 0.1 to 1 in any particular study. Thus, in this study, the exponent parameter n was set as 0.1, 0.3, 0.5, 0,7, 1.0, and the noise over estimation factor α was set as $\alpha = n$ or $\alpha = 2n$. The results show that the optimum parameter depends on the reverberant environment and is very difficult to determine. By combining the results with various

Table 6. Comparison of results with different compensation parameter sets and combination methods for speaker identification (%)

parameters (n,α)	# of impulse response condition for test					Ave.
	4	5	6	7	8	
(0.1, 0.1)	95.2	90.2	87.2	90.5	92.6	91.1
(0.3, 0.3)	**96.2**	89.9	89.6	87.8	89.9	90.7
(0.5, 0.5)	**96.2**	**91.1**	91.4	90.1	92.3	**92.2**
(0.7, 0.7)	95.0	90.3	91.4	**90.8**	92.5	92.0
(1.0, 1.0)	94.0	90.6	91.0	90.5	92.3	91.7
(0.1, 0.2)	94.6	88.9	88.0	86.0	90.4	89.6
(0.3, 0.6)	95.8	89.9	89.8	88.8	91.3	91.1
(0.5, 1.0)	95.3	91.0	91.0	90.5	**93.1**	**92.2**
(0.7, 1.4)	94.5	90.9	**91.7**	90.6	92.9	92.1
((1.0, 2.0)	93.8	89.8	90.9	90.3	92.0	91.3
Conventional combination	96.2	92.5	92.6	92.5	94.1	93.6
Efficient combination	96.2	92.5	92.6	92.5	94.1	93.6

compensation parameter sets, the combined result achieved a relative error reduction of 17.9% compared with the individual results with the optimum parameter. The determination of the parameters for GSS was solved while the computational cost increased. For the conventional combination method, the computational time T_A^{comb} is 10 (the number of the parameter sets I is 10) times the computational time for the individual method T_A. The computational time T_E^{comb} for our proposed efficient combination method is $1.27T_A$, [3] and about $1/8T_A^{comb}$ when the performance is the same as the conventional combination method, which uses all likelihoods by all the speaker models. As a result, the proposed efficient combination method achieved a relative error reduction of 87.5% compared with the baseline, and 44.8% compared with reverberant speech models with almost the same computational cost.

5. CONCLUSIONS AND FUTURE WORK

Previously, Wang et al. proposed a blind dereverberation method based on GSS employing the multi-channel LMS algorithm for hands-free speech recognition [4]. In this study, we applied this method to hands-free speaker identification. However, in the speaker identification field, the method proposed in [4] performed worse than the baseline method, which was opposite to the trend for speech recognition. We addressed this problem by training speaker models using dereverberant speech, which was obtained by suppressing reverberation from arbitrary artificial reverberant speech. The reverberant speech for test data was also compensated using GSS-based dereverberation. By combining the various compensation parameter sets for GSS and calculating the likelihood efficiently, a more robust result was obtained without parameter tuning. Without increasing computational time, the proposed method based on dereverberant speech models achieved a relative error reduction of 87.5% compared with the conventional CMN with beamforming using clean speech models, and 44.8% compared with reverberant speech models.

In the future, we will evaluate the speaker identification experiments using the proposed method in a real environment.

[3] In this study, the values of I, N, K, γ in Eq. 9 are 10, 5, 260 and 92.

6. 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 identification," 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 identification 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] 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 Hands-free Speech Recognition under Reverberant Environments, " Proc. of INTERSPEECH-2008, pp. 968-971, Sep. 2008.

[10] 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.