Applied Acoustics 72 (2011) 124–126

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Technical Note

The effect of reverberation on the performance of cepstral mean subtraction in speaker verification

Noam R. Shabtai [a,*], Boaz Rafaely [a], Yaniv Zigel [b]

[a] Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel
[b] Department of Biomedical Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

ARTICLE INFO

Article history:
Received 1 April 2010
Received in revised form 20 September 2010
Accepted 21 September 2010
Available online 28 October 2010

Keywords:
Speaker recognition
Cepstral mean subtraction
Reverberation

ABSTRACT

Speaker verification (SVR) performance is degraded under reverberation conditions. Cepstral mean subtraction (CMS) is often applied to the feature vectors in order to compensate for convolutive effects of transmission channels, which are considered to have a short-duration impulse response. The effect of reverberation on the performance of CMS applied to the feature vectors in SVR is investigated. Although CMS was found effective in reducing the effect of reverberation for short reverberation time (RT), in cases of long RT, it is shown that CMS may degrade SVR performance rather than improve it. Hence, CMS should not to be used in these cases. In addition, the effect of the room volume was tested and found less critical than the effect of long RT.

1. Introduction

In speaker recognition, features are extracted from speech signals to form feature vectors, and statistical pattern recognition methods are applied in order to model the distribution of the feature vectors in the feature space. Speakers are recognized by pattern matching of the statistical distribution of their feature vectors with target models. Speaker verification (SVR) is the task of deciding, upon receiving tested feature vectors, whether to accept or reject a speaker hypothesis, according to the speaker's model. A popular feature extraction method for speech signal processing is the *mel-frequency cepstral coefficients* (MFCC) [1], and *Gaussian mixture models* (GMM) has become a dominant approach for statistical modeling of speech feature vectors for text-independent SVR [2].

Speaker verification is widely used in telecommunication or conference room applications, where reverberation is often present due to the surrounding room environment. The presence of reverberation adds distortion to the feature vectors, which results in performance degradation of SVR systems due to mismatched conditions between trained models and test segments.

Feature normalization techniques such as *cepstral mean subtraction* (CMS) were originally developed to compensate for the effect of a telephone channel [3], or for the effect of slowly varying convolutive noises in general [2]. For that reason, CMS can be used to reduce the effect of reverberation, if it is characterized by a short-duration *room impulse response* (RIR). However, it is difficult to find research studies in the literature on the effect of CMS on SVR performance under reverberation conditions of long duration RIR, which is often the case in room acoustics.

This work investigates the effect of reverberation on the performance of CMS applied to the MFCC feature vectors in SVR with reverberant speech, including the effect of long *reverberation time* (RT). In that sense, it is an extension of an early study by the authors [4], where only simulated RIRs were used to form reverberant speech. Here both simulated and measured RIRs are employed.

2. Room parameters

Room parameters can either have a direct relation to the physical characteristics of the room, or some relation to the RIR. Associated with the physical characteristics of the room we have the geometrical characteristics, which are the volume V and the surface area S, and the reflection coefficient of the room boundaries, R. The absorption coefficient of the room boundaries a is defined as [5]

$$a = 1 - |R|^2, \tag{1}$$

and thus the absorption area is

$$A = \bar{a}S, \tag{2}$$

* Corresponding author.
E-mail addresses: shabtay@ee.bgu.ac.il (N.R. Shabtai), br@ee.bgu.ac.il (B. Rafaely), yaniv@bgu.ac.il (Y. Zigel).

0003-682X/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apacoust.2010.09.009

where $\bar{a}$ is the average absorption coefficient along the room boundaries.

An important room parameter that can be measured from the RIR is RT, which is the time that takes the energy in a room to decay by 60 dB once the source is turned off. By assuming that until the source was turned off it had been producing a stationary white noise, RT can be calculated from the RIR by using Schroeder's energy decay curve [6]

$$e(t) = 10\log_{10}\int_t^\infty h^2(\tau)\mathrm{d}\tau - 10\log_{10}\int_0^\infty h^2(\tau)\mathrm{d}\tau, \tag{3}$$

where $h(t)$ is the RIR, and numerically solving

$$e(\mathrm{RT}) = -60\ \mathrm{dB}. \tag{4}$$

In the ISO 3382 standard [7], RT is calculated from a least squares based linear fitting of Schroeder's energy decay curve in order to compensate for the non-linearity and for the noise-floor effect.

Room response from a source to a receiver can be given in the frequency domain by the *room transfer function* (RTF). In rectangular rooms, the RTF is known to be a combination of *natural* or eigenmodes. At frequencies where the density of the eigenmode is more than three eigenmodes for a 3 dB bandwidth of a given eigenmode, the sound field is usually considered to sufficiently satisfy the assumptions of diffuse field theory. In diffuse fields, RT is related to the volume by Sabine formula [8]

$$\mathrm{RT} = 0.161\frac{V}{A}. \tag{5}$$

3. Feature extraction and normalization

A commonly used procedure of MFCC feature extraction is shown in Fig. 1 [9]. The pre-emphasis filter is applied to enhance the high frequencies of the spectrum, which are generally reduced by the speech production process. The STFT block splits the signal in the time domain into overlapping frames where the signal is considered to be stationary, and calculates the *fast fourier transform* (FFT) of each frame. Then, filter banking is applied by integrating the magnitude FFT of the signal frames with triangular windows in the mel-frequency domain. Afterwards, the dB level is calculated. This results in a series of energy scalars for every frame. *Discrete cosine transform* (DCT) is calculated, from which coefficients are selected to form MFCC feature vectors. Applying a discrete-time derivative results in ΔMFCC feature vectors, such that

$$\mathbf{c}_t = \left[c_1^t \dots c_N^t, \Delta c_1^t \dots \Delta c_N^t\right]^{\mathsf{T}}, \tag{6}$$

is the feature vector of the t'th frame (t here is a discrete time index), where N is the number of MFCC coefficients.

Transmission channels may add a convolutive effect to the speech signal prior to the process of feature extraction. This may result in feature vectors distortion. For that reason feature normalization may be used. In this chapter we discuss the CMS technique, which is the operation of subtracting the sample mean [9]

$$\tilde{\mathbf{c}}_t = \mathbf{c}_t - \boldsymbol{\mu} \qquad t = 0 \dots T-1, \tag{7}$$

where $\boldsymbol{\mu}$ is the sample mean of the series $\mathbf{c}_0 \dots \mathbf{c}_{T-1}$. The operation of CMS may include variance normalization [3] by dividing the components by the sample *standard deviation* (STD), i.e.,

$$\tilde{c}_n^t = \frac{\tilde{c}_n^t}{\sigma_n} \qquad \begin{matrix} t = 0 \dots T-1 \\ n = 1 \dots N \end{matrix}, \tag{8}$$

where for every $n = 1,\dots,N$, σ_n is the sample STD of the series $c_n^0 \dots c_n^{T-1}$.

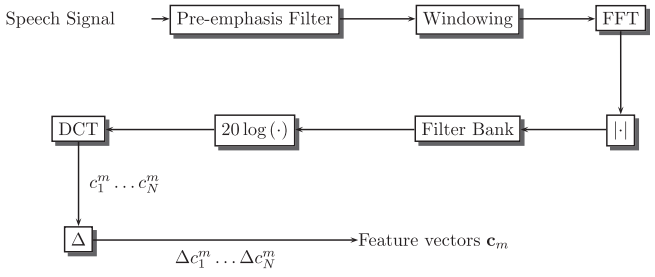

Fig. 1. Extraction of MFCC and ΔMFCC feature vectors from speech signal [9].

4. Experimental study of the performance of SVR with CMS and reverberant speech

An early study by the authors [4] investigated the effect of room parameters on the efficiency of CMS in improving the performance of SVR. The performance of an SVR system was measured by calculating the *equal error rate* (EER) in rooms with different RTs and volumes. Test speech segments were made reverberant with RIRs that were simulated using the image method of Allen and Barkley [10]. It was shown that for long RTs, the efficiency of CMS decreases. It was also shown that for certain room volumes, using CMS may increase the EER rather than decrease it in cases of long RT.

In this work we extend the research to reverberant speech generated by convolution with measured RIRs. The environments in which the RIRs were measured are tabulated in Table 1. Measured RIRs 1–10 were measured with a Brüel & Kjær 4295 Omni-Source loudspeaker and a Brüel & Kjær 4942 $\frac{1}{2}$-in. diffuse-field microphone, at selected rooms in *Ben-Gurion University of the Negev, Israel* (BGU). Measured RIRs 11–14 were taken from the *Concert Hall Research Group* (CHRG) project [11]. In order to compare the results with simulated RIRs, the image method was used to simulate RIRs of rooms with similar dimensions and RTs to the rooms in Table 1.

The SVR system used 20 ms speech frames in which MFCC and ΔMFCC were calculated to form 24-dimensional feature vectors, for which CMS was either applied or not. Target models were trained using the *adaptive GMM* (AGMM) approach [2]. A *background GMM* (BGM) of 1024 Gaussians was created from one-minute long non-reverberant speech segments of 50 speakers taken from the NIST-1998 SRE database. This BGM was used to train target AGMMs for 198 male speakers, with 1-min long non-reverberant speech segments, taken from the NIST-1999 SRE [12] database. Test speech segments were taken from NIST-1999 SRE for 686 male speakers with a half-min long speech segment each.

Table 1
Rooms in which RIRs were measured.

RIR	Environment	RT (s)	V_i(m^3)
1	Building 33 office 126	0.8	37
2	Building 33 office 427	0.6	42
3	Building 34 classroom 103	0.6	120
4	Building 33 lecture room 102	0.5	147
5	Building 34 classroom 202	1	301
6	Building 33 teaching lab 204	0.6	339
7	Building 26 auditorium 4	1.5	793
8	Building 26 auditorium 5	1.2	1142
9	Building 26 auditorium 6	1.3	1142
10	Sonnenfeld lecture room	1	2529
11	Mechanics hall (Worchester, MA)	2.4	8367
12	Troy music hall (Troy, NY)	2.6	11320
13	Boston symphony hall (Boston, MA)	2.6	16611
14	Kleinhans music hall (Buffalo, NY)	1.9	18241

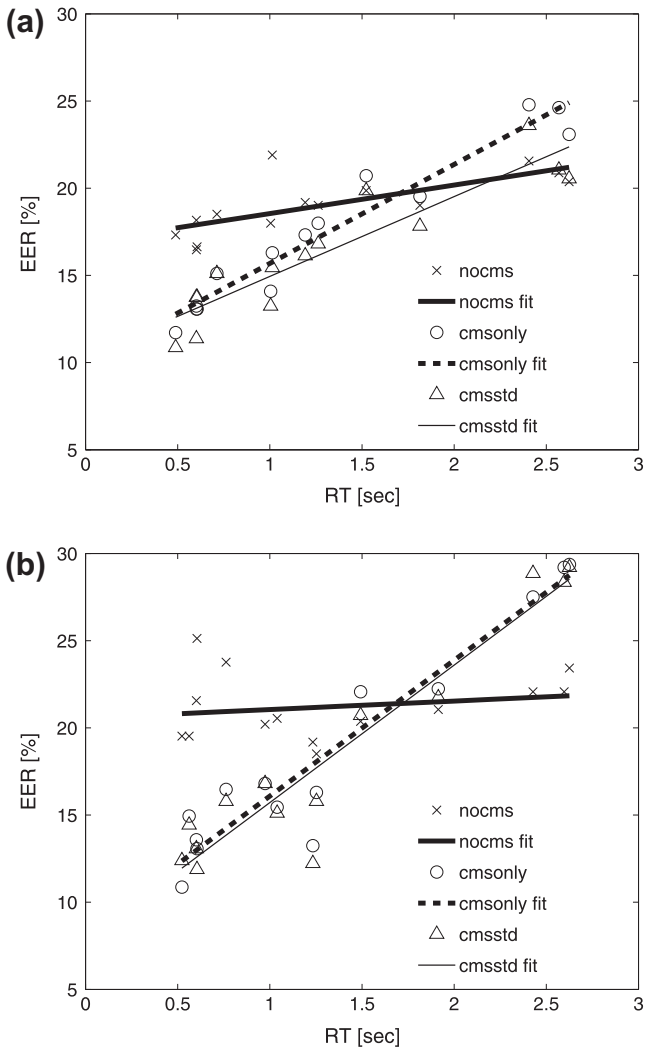

Fig. 2. EER values of SVR with reverberant speech as a function of RT. Crosses ("×") denote using no feature normalization (linear fitting with thick solid line), circles ("○") denote using CMS (linear fitting with thick dashed line), and triangles ("△") denote using CMS with variance normalization (linear fitting with thin solid line). Test speech segments were made reverberant by convolution with (a) simulated and (b) measured RIRs.

The test speech segments were made reverberant by convolving them with simulated and measured RIRs. The EER results were calculated by introducing the reverberant test speech segments to the target AGMMs and BGM of non-reverberant speech.

Fig. 2 shows a scatter plot of EER values as a function of RT. The cross, circle, and triangle marks on Fig. 2 represent EER values. When feature normalization was not used, or CMS was applied, or CMS was applied along with variance normalization, respectively. Linear fitting to the EER values is shown in Fig. 2. Thick solid curves denote using no feature normalization, dashed curves denote using CMS, and thin solid curves denote using CMS along with variance normalization. Fig. 2a and b refer to simulated and measured RIRs, respectively. In the case of simulated RIRs, as well as in the case of measured RIRs, it can be seen that CMS improves the performance of SVR to a lesser extent than with the increase of RT. Moreover, it can be seen that for some high values of RT, CMS may increase the EER rather than decrease it. These results support previous results [4] in which it was shown that CMS improves the performance of SVR to a lesser extent with the increase of RT, and validate them with measured RIRs.

In order to compare the effect of the room volume to the effect of RT on the performance of SVR with CMS, a 3-D analysis was performed with EER as a function of RT and the volume V, along with a planar fitting. Since it may be difficult for the reader to trace 3-D mesh plots on a 2-D hard-copy, the plots are not included here. Six planar fittings have been performed for the permutations of the possible configurations, using simulated or measured RIRs, and using no feature normalization, CMS, or CMS with variance normalization. The axes of RT and V were normalized with the maximum values in Table 1. The STD of the distance from the planar fitting ranged from 0.8% to 1.2% for simulated RIRs, and from 1.8% to 2.0% for measured RIRs. Since the values of the EER roughly range between 10% and 30%, this implies that the planar fitting is a relatively fair estimator of the EER values as a function of RT and V. In the cases where CMS was used, the slope of the EER as a function of the normalized RT was found to range from 16.1% to 20.6% per normalized time unit, while as a function of V it ranged only from −4.1% to 0.2% per normalized room-volume unit. Hence, it seems that RT has greater effect on the performance of CMS than the room volume may have.

5. Conclusions

The effect of room volume and RT on the performance of CMS applied to MFCC feature vectors in SVR was investigated. It was shown that the performance of CMS may degrade with the increase of RT. In some cases of long RT, CMS may increase the EER of SVR rather than decrease it. Hence, in these cases, CMS should not automatically be used. The effect of the room volume was found less critical than the effect of long RT. As a future work, we purpose combining a CMS decision block in SVR.

References

[1] Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 1980;ASSP-28:357–66.
[2] Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. Digit Signal Process 2000;10:19–41.
[3] Mammone RJ, Zhang X, Ramachandran RP. Robust speaker recognition: a feature-based approach. IEEE Signal Process Mag 1996;13:58–71.
[4] Shabtai NR, Rafaely B, Zigel Y. The effect of room parameters on speaker verification using reverberant speech. In: Proceedings of the IEEEI; 2008. p. 231–5.
[5] Kuttruff H. Room acoustics. New York: Spon Press; 2000.
[6] Schroeder MR. New method of measuring the reverberation time. J Acoust Soc Am 1965;37:409–12.
[7] ISO 3382:1997. Acoustics – measurement of the reverberation time of rooms with reference to other acoustical parameters; 1997.
[8] Kinsler LE, Frey AR, Coppens AB, Sanders JV. Fundamentals of acoustics. New York: John Wiley; 2000.
[9] Bimbot F, Bonastre JF, Fredouille C, Gravier G, Meignier S, Merlin T, et al. A tutorial on text-independent speaker verification. EURASIP J Appl Signal Process 2004;2004:430–51.
[10] Allen JB, Berkley DA. Image method for efficiently simulating small room acoustics. J Acoust Soc Am 1979;65:943–50.
[11] Concert Hall Research Group CD v.3, Concert Hall Research Group, 327F Boston Post Road. Sudbury, MA 01776; 2004 [Attention: Timothy J. Foulkes, email: chrg@cavtocci.com, phone: 978 443 7871, fax: 978 443 7873].
[12] Martin A, Przybocki M. The NIST 1999 speaker recognition evaluation - an overview. Digit Signal Process 2000;10:1–18.