6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A NEW CEPSTRUM-BASED CHANNEL COMPENSATION METHOD FOR SPEAKER VERIFICATION

T.F. Lo, M.W. Mak and K.K. Yiu

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University

enmwmak@polyu.edu.hk

ABSTRACT

A new cepstrum-based channel compensation method is proposed for speaker verification over the telephone network. The method consists of intra-frame and inter-frame cepstral processing. For the former, a pole-removed cepstrum is derived, where the LP poles with frequency higher than a certain threshold are removed. For the latter, we introduce a novel way of cepstral mean subtraction called differential-partial cepstral mean subtraction (DPCMS). The main idea is that the cepstral mean of clean speech is not necessarily zero and that the cepstral difference between clean and channel-corrupted speech is mainly contributed by the channel effects on LP poles within a certain frequency range. A speaker verification system based on radial basis function networks was used to evaluate the proposed approach. Clean speech was used to train the networks and telephone speech was used to evaluate their performance. Experimental results show that the proposed method reduces verification error rate significantly.

1. INTRODUCTION

Speaker verification is to verify the identity of a speaker based on his/her own voice. While today's speaker verification systems perform reasonably well under controlled conditions, their performance often suffers in real-world, unpredictable acoustic environment. For example, variations in handset characteristics could result in a mismatch between the speech gathered during enrollment and verification.

There are two main schools of thought to address the above problem. The first, known as intra-frame processing, looks at the local spectral characteristics of a given frame of speech. One of the recent proposals is the adaptive component weighting (ACW) [1], which minimizes the variations due to the channel effect by normalizing the residues of the LP poles. The second school of thought, known as inter-frame processing, exploits the temporal variability of a sequence of feature vectors. Typical examples include cepstral mean subtraction (CMS) [2] and pole-filtered cepstral mean subtraction [3], where the channel characteristics are represented by the cepstral mean of a segment of channel distorted speech.

In this work, we propose a novel approach, which uses both intra-frame and inter-frame cepstral processing to overcome the limitations of CMS and ACW. For the intra-frame processing, the approach assumes that high-frequency poles are more sensitive to channel effects and that they have less contribution to speech formants as compared to low-frequency poles. For the inter-frame processing, the channel is represented by the cepstral difference between clean speech and channel-corrupted speech at certain frequency range. We demonstrate the superiority of this approach by training a speaker verification system using TIMIT speech and evaluating its performance with NTIMIT speech [5]. The results show that this approach can reduce the equal error rate by more than 50%.

2. ANALYSIS OF CHANNEL EFFECTS BY USING TIMIT AND NTIMIT SPEECH

The short-term transfer function of a vocal tract-filter $H(z)$ can be expressed in a parallel form by means of partial expansion

$$H(z)=\frac{1}{A(z)}=\frac{1}{1+\sum_{i=1}^{P}a_i z^{-i}}=\sum_{i=1}^{P}\frac{r_i}{\left(1-z_i z^{-1}\right)}. \qquad (1)$$

In (1), r_i is the residue of the poles and z_i represents the center frequency and the bandwidth of the i^{th} component of the LP model. It was observed in [1] that the residue r_i is highly sensitive to channel effects. The Adaptive Component Weighting (ACW) cepstrum was proposed to reduce channel mismatches by normalizing the residue r_i, thereby minimizing their variations. It was shown that the ACW cepstrum can reduce the channel effects caused by a channel simulator formed by a first-order FIR filter [1]. However, this method can only address part of the channel mismatch problem, as it does not account for the non-linear characteristics of communication channels [4].

In order to study the distortion caused by a telephone channel, the spectra of individual LP poles of a frame of voiced speech derived from the TIMIT and NTIMIT [5] corpuses are plotted in Fig. 1. A comparison of the spectra reveals that not only the residue but also the bandwidth and center frequency of the poles are perturbed by the channel. As a result, the ACW approach, which only normalizes the residues of the poles as shown in Fig. 2, cannot compensate for the changes in center frequencies and bandwidths.

The LP and ACW cepstra of a frame of voiced speech are plotted in Figures 3 and 4 in order to show their robustness to channel effects. These figures show that the LP spectrum and the cepstrum of NTIMIT speech

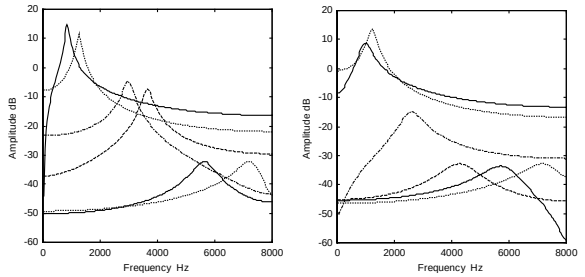
differ significantly from those of TIMIT. This could result in poor verification performance when TIMIT speech is used for enrollment and NTIMIT speech is used for verification. A new method, which consists of both intra-frame and inter-frame cepstral processing, is proposed here to tackle the problem.

3. INTRA-FRAME CEPSTRAL PROCESSING

A new intra-frame cepstral processing method, called pole-removed cepstral processing, is introduced to minimize channel mismatches. Fig. 1 shows that narrowband low-frequency poles are less sensitive to channel effects and that they have a larger contribution to speech formants. In light of this observation, we propose to remove the poles with frequencies above a certain threshold. For each frame of voiced speech, $H(z)$ is calculated, and any LP poles z_i with frequency above f_t are removed from the partial fraction expansion of $H(z)$. A new cepstrum $c_{pr}(n)$ is then calculated based on the modified $H(z)$

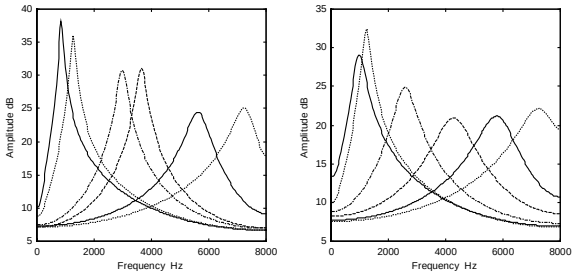
Figures 5 and 6 show the modified LP spectra and the resulting pole-removed cepstra $c_{pr}(n)$ of a frame of TIMIT and NTIMIT speech, with the threshold frequency f_t being set to 3.5 kHz and 2.5 kHz, respectively. It is evident that the lower the threshold frequency, the smaller the difference between the pole-removed LP spectra of TIMIT speech and that of NTIMIT speech. Similar situation occurs in the cepstra. Although the pole-removed cepstra are not as sensitive as the ordinary cepstra to channel effects, they become less capable of modeling speakers when a large number of poles has been removed. Therefore, a compromise must be made when the threshold frequency is selected.

Figures 7(c) and 7(d) show the average pole-removed cepstra, at f_t = 2.5 kHz and 3.5 kHz respectively, of different voiced speech derived from the TIMIT (solid line) and NTIMIT (dotted line) corpuses. They demonstrate that pole-removed cepstra with a low threshold frequency are more capable of reducing channel mismatches. However, the removal of high-frequency poles will reduce the speaker-specific information contained in the high-order cepstral coefficients and cause performance degradation, as will be shown later. The next section presents an inter-frame



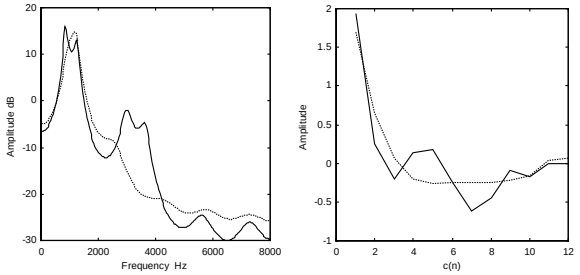
(a) (b)

Figure 1 Components of the LP spectrum of a frame of voiced speech derived from (a) TIMIT and (b) NTIMIT.



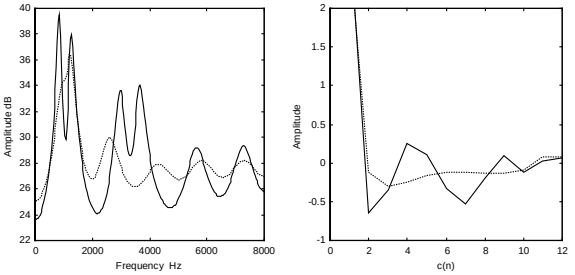
(a) (b)

Figure 2 Components of the ACW spectrum of a frame of voiced speech derived from (a) TIMIT and (b) NTIMIT.



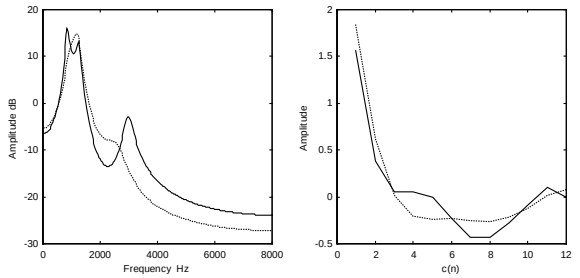
(a) (b)

Figure 3 (a) LP spectra and (b) LP cepstra of a frame of voiced speech derived from TIMIT (solid line) and NTIMIT (dotted line).



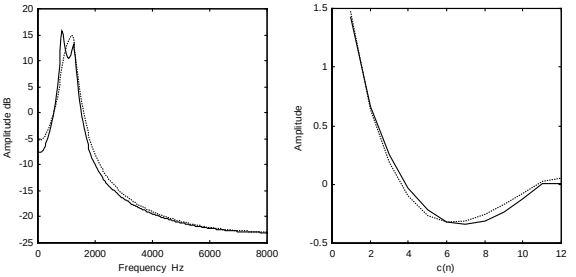
(a) (b)

Figure 4 (a) ACW spectra and (b) ACW cepstra of a frame of voiced speech derived from TIMIT (solid line) and NTIMIT (dotted line).



(a) (b)

Figure 5 (a) Pole-removed LP spectra and (b) cepstra of a frame of voiced speech derived from TIMIT (solid line) and NTIMIT (dotted line). The threshold frequency is set to 3.5 kHz.



(a) (b)

Figure 6 (a) Pole-removed LP spectra and (b) cepstra of a frame of voiced speech derived from TIMIT (solid line) and NTIMIT (dotted line). The threshold frequency is set to 2.5 kHz.

processing method, which not only reduces the channel, mismatches, but also able to preserve speaker-specific information in the higher-order cepstral coefficients.

4. INTER-FRAME CEPSTRAL PROCESSING

It is commonly believed that linear distortions due to the filtering effect of the channel can be removed by subtracting the cepstral mean of the distorted speech (which represents the channel) from the cepstrum of the distorted speech. This technique is known as cepstral mean subtraction (CMS) [2]. The method assumes that the channel is linear and that the clean speech is phonetically balanced, giving zero-mean cepstrum for clean speech. However, the assumptions are hardly valid in most practical situations [4]. Fig. 8 shows the mean of LP cepstrum of clean speech derived from 20 speakers in the TIMIT corpus. It clearly shows that the assumption of zero-mean cepstrum is invalid in this case. Fig. 7(b) shows the effect of CMS on the NTIMIT (dotted line) speech. Clearly, CMS is not very effective in eliminating the discrepancy between the cepstra of TIMIT and NTIMIT speech.

Here, we introduce a new way of CMS called differential-partial cepstral mean subtraction (DPCMS). The main idea is based on the observations that the cepstral mean of clean speech is not zero and that the cepstral difference between clean and channel-corrupted speech is mainly contributed by the channel effects on LP poles within a certain frequency range, as shown in Figures 1 and 3.

Assuming LP poles with frequencies above f_t can be neglected for the purpose of speaker recognition, the pole-removed cepstrum $c_{pr}(n)$ is calculated by removing poles with frequencies higher than f_t. Similarly, a base-frequency f_b is chosen, and any poles with frequencies smaller than f_b are assumed to be unaffected by channel effects. A base-cepstrum $c_b(n)$ is calculated by removing the poles with frequencies above f_b. Hence, the relation between $c_{pr}(n)$ and $c_b(n)$ is

$$c_{pr}(n) = c_b(n) + c_h(n) \tag{4}$$

where $c_h(n)$ is the cepstrum corresponding to poles having frequencies between f_b and f_t. $c_h(n)$ can be found simply by $c_{pr}(n) - c_b(n)$. The relation between the pole-removed cepstrum of clean speech and channel-corrupted speech is

$$c_{pr,corrupted}(n) = c_{pr,clean}(n) + c_{channel}(n) \tag{5}$$

where $c_{pr,corrupted}(n)$ is the pole-removed cepstrum of channel-corrupted speech, $c_{pr,clean}(n)$ is the pole-removed cepstrum of clean speech and $c_{channel}(n)$ is the channel cepstrum. According to (4), (5) can be written as

$$\begin{aligned} &c_{b,corrupted}(n) + c_{h,corrupted}(n) \\ &= c_{b,clean}(n) + c_{h,clean}(n) + c_{channel}(n)\,. \end{aligned} \tag{6}$$

Since the base-frequency f_b is chosen such that the poles with frequencies below it are insensitive to channel distortion, we can assume that $c_{b,corrupted}(n) \approx c_{b,clean}(n)$.

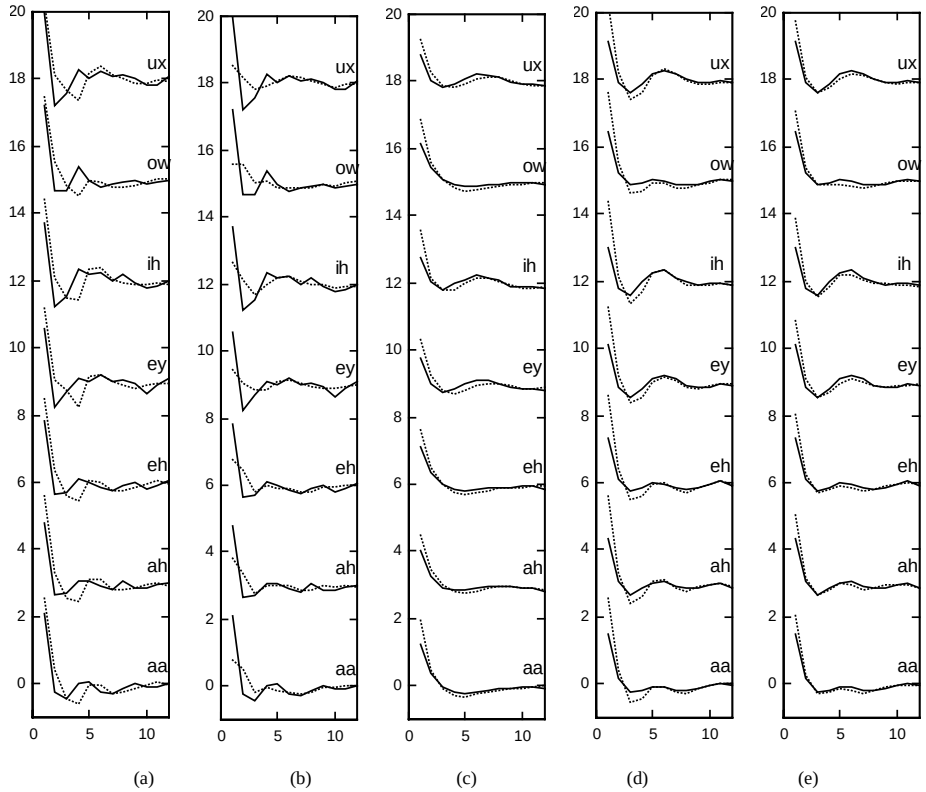


Figure 7 Graphs showing (a) the cepstra, (b) cepstra with CMS being applied to NTIMIT speech, (c) pole-removed cepstra with f_t = 2.5 kHz, (d) pole-removed cepstra with f_t = 3.5 kHz, and (e) pole-removed cepstra with f_t = 3.5 kHz and DPCMS being applied to NTIMIT speech of seven vowels derived from TIMIT (solid line) and NTMIT (dotted line). The horizontal axis represents the order of the cepstral coefficients and the vertical axis represents the relative amplitude.

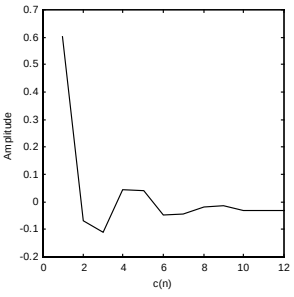


Figure 8 Mean of LP cepstrum of clean speech frames derived from 20 speakers in the TIMIT corpus

Hence, the channel cepstrum can be approximated by

$$c_{channel}(n) = E[c_{h,corrupted}(n)] - E[c_{h,clean}(n)] \quad (7)$$

where $E[\cdot]$ denotes the expectation operator. Finally, the pole-removed cepstrum with DPCMS can be computed as

$$c_{pr,DPCMS}(n) = c_{pr,corrupted}(n) - c_{channel}(n) . \quad (8)$$

The mean of $c_{h,clean}(n)$ can be obtained from a clean speech database, whereas the mean of $c_{h,corrupted}(n)$ can be obtained from the channel-corrupted speech during verification. Figures 7(c) and (d) plot the pole-removed cepstrum with f_t being set to 2.5 kHz and 3.5 kHz respectively. The results show that at high threshold frequency, the high-order cepstral coefficients preserve more spectral features. However, the cepstra of TIMIT and NTIMIT speech exhibit a greater difference at high threshold frequency. This situation can be remedied by applying DPCMS to the pole-removed cepstrum of NTIMIT speech, as illustrated in Fig. 7(e).

5. EXPERIMENTAL RESULTS

In this section, experiments on closed set text-independent speaker verification are presented. The speech from the TIMIT corpus was used for enrollment while the speech from the NTIMIT corpus was used for verification. The aim is to evaluate the channel compensation methods for robust telephone speaker verification.

Radial Basis Function (RBF) networks [6] were used as the pattern classifiers. Each speaker was assigned a network characterizing his/her own voice. Conventional approach was used to estimate the RBF parameters. More specifically, the K-means algorithm was applied to determine the center positions, followed by the K-nearest neighbor algorithm, which determines the spread of each RBF unit. Then the output weights are estimated by a least-squares method.

Seventy six speakers from dialect region 2 of the TIMIT and NTIMIT corpuses were used in the experiments. In an enrollment session, feature vectors derived from the SA and SX sentence sets of a speaker and 20 anti-speakers were used to train a speaker-specific RBF network. Each network was trained to distinguish the voice of the speaker and those of the anti-speakers. In a verification session, feature vectors derived from the SI sentence set of the speaker and 36 impostors from the NTIMIT corpus were applied to the speaker-specific network.

For each sentence, the unvoiced speech was removed since the cepstral features we proposed are based on the channel effect on voiced speech. Table I shows the false acceptance rate (FAR), false rejection rate (FRR) and equal error rate (ERR) based on different cepstral features.

Table I shows that the conventional CMS cannot reduce the error rate since the mean cepstrum is not usually zero in practical situations. The ACW cepstrum also cannot improve the accuracy. However, our proposed pole-removed cepstrum with DPCMS can significantly reduce the channel mismatch between TIMIT and NTIMIT speech, resulting in a much smaller ERR.

6. CONCLUSION

This paper has presented a new cepstrum-based channel compensation method to reduce the error rate of telephone-based speaker verification systems. The method consists of both intra-frame and inter-frame cepstral processing. It has been shown that the method can reduce the mismatches between TIMIT and NTIMIT speech. Experimental results show that the new method can reduce the error rate by more than 50%.

ACKNOWLEDGEMENT

This work was supported by The Hong Kong Polytechnic University under Grant No. G-S725.

REFERENCES

1. K.T. Assaleh and R.J. Mammone, "New LP-derived features for speaker identification," *IEEE Trans. on Speech, Audio Processing*, vol. 2, no. 4, pp. 630-638, October 1994.
2. B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, pp. 1304-1312, 1974.
3. D. Naik, "Pole-filtered Cepstral Mean Subtraction," *Proc. ICASSP'95*, vol. 1, pp. 157-160, 1995.
4. D.A. Reynolds, M.A. Zissman, T.F. Quatieri, G.C. O'Leary, and B.A. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," *Proc. ICASSP'95*, pp. 329-332, May 1995.
5. C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech telephone bandwidth speech database," *Proc. ICASSP'90*, pp. 109-112, April 1990.
6. J. Moody and C.J. Darken, "Fast learning in networks of locally tuned processing units," *Neural Computation*, vol. 1, pp. 281-194, 1989.

Features	Threshold frequency f_t / kHz	Unvoiced speech removed?	Inter-frame processing on NTIMIT	FAR / %	FRR / %	ERR / %
LP cepstrum	-	No	-	75.35	23.44	46.68
LP cepstrum	-	No	CMS	75.18	22.02	47.58
LP cepstrum	-	Yes	-	40.39	58.00	44.35
ACW cepstrum	-	No	-	16.31	85.00	47.38
ACW cepstrum	-	Yes	-	49.56	46.67	48.33
Pole-removed cepstrum	2.5	Yes	-	37.59	52.92	25.43
Pole-removed cepstrum	3.5	Yes	-	38.79	44.25	31.48
Pole-removed cepstrum	3.5	Yes	DPCMS with base-frequency f_b = 2.5 kHz	40.35	33.67	20.50

Table I The FAR's, FRR's and ERR's obtained by the verification system. TIMIT speech was used as the training set and NTIMIT speech was used as the test set.