POLE-FILTERED CEPSTRAL MEAN SUBTRACTION

Devang Naik

CAIP Center, Rutgers University, Piscataway, New Jersey 08855.

ABSTRACT

This paper introduces a new methodology to remove the residual effects of speech from the cepstral mean used for channel normalization. The approach is based on filtering the eigenmodes of speech that are more susceptible to convolutional distortions caused by transmission channels. The filtering of Linear Prediction (LP) poles and their corresponding eigenmodes for a speech segment are investigated when there is a channel mismatch for speaker identification systems.

An algorithm based on Pole-filtering has been developed to improve the commonly employed Cepstral Mean Subtraction. Experiments are presented in speaker identification using speech in the TIMIT database and on the San Diego portion of the KING database. The new technique is shown to offer improved recognition accuracy under cross channel scenarios when compared to conventional methods.

1. INTRODUCTION

Channel normalization techniques implemented in the cepstral domain have been proposed in the past [1,9,6,11]. They modify or weight the cepstral coefficients to minimize the mismatch in the training and test data due to channel distortions caused by the acquisition of speech via different microphones, hand-sets or transmission channels. This paper introduces a new technique that offers a more accurate method of channel normalization.

The all-pole model based on LP analysis is frequently used in Speech/Speaker recognition [1]. For an all-pole filter of order p given by,

$$S(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{\prod_{k=1}^{p}(1 - z_k z^{-1})}, \quad (1)$$

with roots $z_k, k = 1, 2, \cdots, p$ of the model, the roots correspond to the modes of the linear system of speech. These roots form the dominant modes modeling the speech segment. Each root z_k has associated with it a center frequency, $\omega_k = \frac{1}{2\pi} \arctan \frac{\Im(z_k)}{\Re(z_k)}$, and a bandwidth, $B_k = -\frac{1}{\pi} ln(|z_k|)$, in units of π radians. Schroeder [4], expressed the cepstral coefficients as a root power sum formula,

$$c(n) = \frac{1}{n} \sum_{k=1}^{p} z_k^n = \frac{\tilde{c}(n)}{n} = \frac{1}{n} \sum_{k=1}^{p} e^{-n(B_k + j\omega_k)}, \quad (2)$$

where $c(n)$ is the n^{th} cepstral coefficient and $\tilde{c}(n)$ is the n^{th} liftered cepstral coefficient [2]. The roots generally occur in complex conjugate pairs or are real. A filter with p poles, may consist q pairs of complex poles and remaining $p - 2q$ real poles. The impulse response of complex conjugate pole pair, $[z_k, z_k^*]$, corresponds to a damped sinusoid represented by,

$$\frac{1}{(1 - z_k z^{-1})(1 - z_k^* z^{-1})} \xrightarrow{z^{-1}} |z_k|^n \cos(\omega_k n). \quad (3)$$

Each complex conjugate pole pair represents a component in the spectral domain (referred to as a **spectral component**) corresponding to a center frequency ω_k, and bandwidth B_k. The relationship between the cepstrum and the spectral components can be used to investigate the effect of channel variations. The modification of the components of speech under known convolutional distortions and their derived cepstra form the basis of the pole-filtering approach.

The outline of the paper is as follows. In Section 2, channel normalization in cepstral domain is discussed. Section 3 discusses the effect of all-pole parameters on estimates of convolutional distortions and the pole filtering methodology for extracting robust cepstral features. In Section 4, the results of speaker identification experiments are reported followed by summary and conclusions in Section 5.

2. CHANNEL NORMALIZATION USING CEPSTRAL MEAN SUBTRACTION

It is well known that a time-invariant distortion caused by a recording apparatus or the transmission channel, can be eliminated by *Cepstral Mean Subtraction* (CMS). This method of eliminating the distortion relies on the assumption that the ensemble average of the speech waveform is zero.

CMS has been widely used to equalize the channel mismatch between training and testing data for both, speech and speaker recognition systems [1,6,11]. Elimination of such cepstral bias is also implicit in most standard channel normalization techniques [6,9].

However, in most practical situations, where the amount of speech data available is limited for training and for testing, the assumption that the average cepstrum due to speech is zero-mean does not hold. In general, the long term cepstral mean tends to represent the gross spectral distribution of the speech in addition to an estimate of the time-invariant distortion. With CMS, an improper estimate of the channel cepstrum tends to attenuate useful spectral information from every frame. Hence, although CMS helps normalize the channel mismatch, it tends to eliminate useful spectral information which reduces the classification accuracy. The effect of CMS can be understood by studying the spectra

Work was supported by the Air Force/Rome Laboratories contract number F30602-91-C-0120.

0-7803-2431-5/95 $4.00 © 1995 IEEE

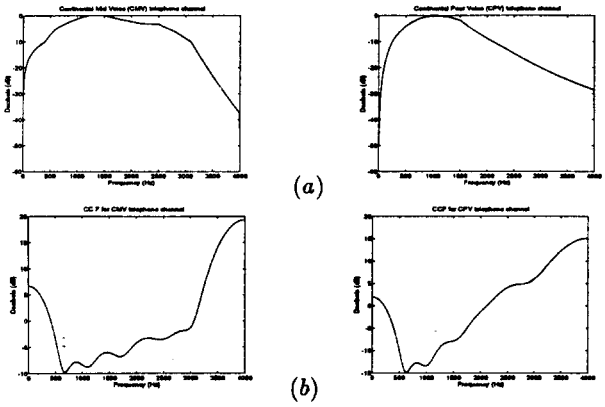


Figure 1: (a) Original Channels [Wire] (b) Estimated inverse filter responses.

of the cepstral mean, also called the channel compensation filter (CCF) [10].

The spectra of cepstral means of a speech utterance from the TIMIT database, processed through two simulated telephone channels (Continental Mid Voice (CMV) and Continental Poor Voice (CPV)) in figure (1(a)), is shown in figure (1(b)). One can observe from the frequency responses of the filters that they have the characteristic response of a corresponding inverse (or deconvolution) filter. The spectral contents of the cepstral mean, $\mathbf{c}_S$, for a sentence, S, can be categorized in the cepstral domain as corresponding to,

- a spectral roll-off mainly due to the channel, $\mathbf{h}_S$, and,
- variations in the spectra which are due to the gross spectral distribution of speech, $\mathbf{s}_S$.

The speech information present in the cepstral mean is important and should not be eliminated when CMS is carried out for channel compensation. In order to achieve proper channel normalization a methodology needs to be developed that **decouples** the speech information in the cepstral mean from the channel information. A more accurate channel normalization would be achieved by de-emphasizing the component, $\mathbf{s}_S$, to effectively eliminate a cepstral mean, $\mathbf{c}_S \rightarrow \mathbf{h}_S$. An reasonably accurate estimate of the channel cepstrum, $\mathbf{c}_S$, could be obtained if the cepstral mean solely due to clean speech, $\mathbf{s}_S$, with which the channel were convolved was available, by computing,

$$\hat{\mathbf{h}_S} \approx \mathbf{c}_S - \mathbf{s}_S. \tag{4}$$

However, $\mathbf{s}_S$ is never available in practice and hence it is impossible to entirely decouple the cepstral component due to speech from the cepstral component that corresponds to the channel.

The residual speech in the cepstral mean also attenuates the spectral content in some regions of the spectra [10]. Such attenuation would typically occur for all speech frames from which the cepstral mean is being subtracted. This has a degrading effect on the accumulated spectral distortion over the entire speech utterance when used for classification. These observations motivate the pole filtering approach to channel normalization. The following section outlines the approach in detail.

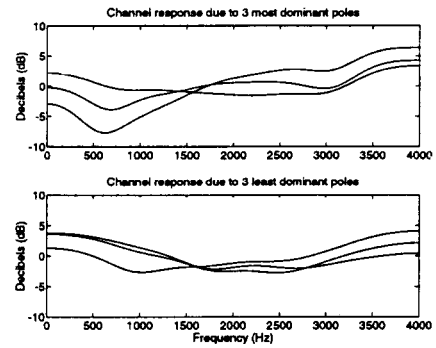


Figure 2: Responses for partial cepstral means for speech degraded by CMV channel.

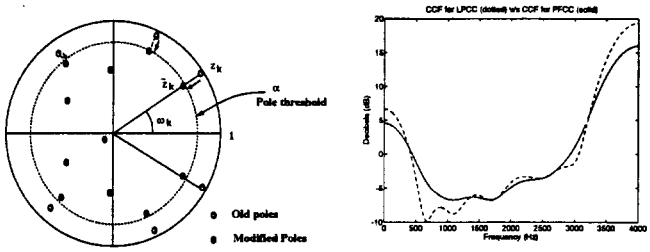
3. THE POLE-FILTERING APPROACH

The pole-filtering approach makes use of the effect of individual poles from the all-pole model of the vocal tract on *apriori* known channel distortions. Since the cepstrum is a weighted combination of poles or spectral components, the effect of the individual components on the cepstral mean can be investigated. By studying this effect, algorithms can be developed that reduce the speech content in the cepstral mean and thereby improve the channel estimate.

A simple experiment to study the effect was carried out by evaluating the partial long-term cepstral means corresponding to the dominance of poles (based on bandwidth) in every frame of a speech utterance. The contribution to the long-term cepstral mean due to most the dominant spectral component (pole pair closest to the unit circle) was first evaluated. Next, the contribution due to the second most dominant spectral component was found and so on for the rest of the LP spectral components. The responses were investigated for each of the partial means for a clean utterance degraded by the CMV channel as shown figure (2).

One can observe from the individual frequency responses due to the partial cepstral means, that the contribution to the overall long-term cepstral mean due to the more dominant poles (or the narrow band poles), is more biased by the spectral content relating to speech represented by high-Q regions. In fact, the inverse filter due to the narrow-band poles exhibits characteristics that would attenuate spectral information when subtracted in the cepstral domain. The contribution to the long-term mean by the broad-band poles however, tends to exhibit smoother inverse filter characteristics and compensates only for the roll-off in the spectra due to the channel.

Pole filtering algorithms exploit this observation to improve the channel estimate by modifying the dominant modes in the speech frame. The strategy is to de-emphasize the effect of the dominant modes on the cepstral mean estimate. One technique of improving the estimate of the channel is to use Pole filtered cepstral coefficients (PFCC). The PFCC are LP-based cepstral coefficients derived by inflating the



(a) (b)

Figure 3: (a) Pole thresholding process on the unit circle, (b) Effect of pole filtering on spectra of cepstral mean.

bandwidths of the narrow-band poles while their frequencies are left unchanged. The bandwidth broadening approach can be carried out by selectively inflating the pole bandwidths.

In selective bandwidth broadening, the narrow band poles are shifted inward away from the unit circle along the same radius, thus keeping the frequency unchanged but broadening the bandwidths. The procedure is illustrated in the figure (3(a)). The resulting cepstrum computed after manipulation of the poles is called pole–filtered cepstrum and is averaged to calculate a modified cepstral mean.

For each frame:

Evaluate the roots z_k of the LP polynomial.
if $abs(z_k) \geq \alpha$, (α= pole bandwidth threshold),
$abs(z_k) = \alpha$;
Modify z_k to $\tilde{z_k}$ by,
$\tilde{z_k} = \alpha \angle z_k$;
endif
Evaluate **LPCC** using z_k.
Evaluate **PFCC** using $\tilde{z_k}$.

Figure (3(b)) illustrates the effect of pole-filtering on spectra of cepstral mean. Broadening of bandwidths of the poles can also be achieved by weighting the prediction coefficients to compute the spectrum. This can be accomplished using $A(\gamma z) = 1 + \sum_{k=1}^{p} a_k(\gamma z)^{-k}$ and the corresponding cepstral transformation $c_{PFCC}(n) = \gamma^n c_{LPCC}(n)$ where γ with a value between 0 and 1, is a bandwidth broadening factor [8]. The value of $\gamma = e^{-(\pi \frac{\delta}{f_s})}$, based on δ Hz, which is the frequency with which the pole bandwidths can be broadened.

The modified long-term cepstral mean, $\mathbf{c}_S^{pf}$, is subtracted from the LP cepstrum of every speech frame instead of subtracting the ordinary long-term cepstral mean. The choice of the broadening factor α or γ for selective pole modification is justified by empirically observing and choosing the range of bandwidths of the poles of an all-pole fit to the impulse responses of the actual simulated channels. It can be observed that the poles are sufficient more broad-band compared to the poles of a typical voice speech frame, when spectrally fitting the bandpass effect of a channel.

The relative error in the cepstral mean estimate with respect to $\hat{\mathbf{h}}_S$, in equation (4) for ordinary cepstral mean,

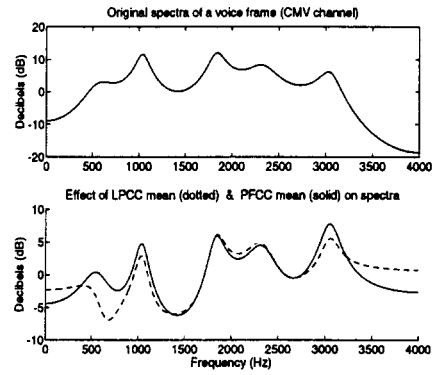


Figure 4: Channel normalization using ordinary mean v/s pole filtered mean.

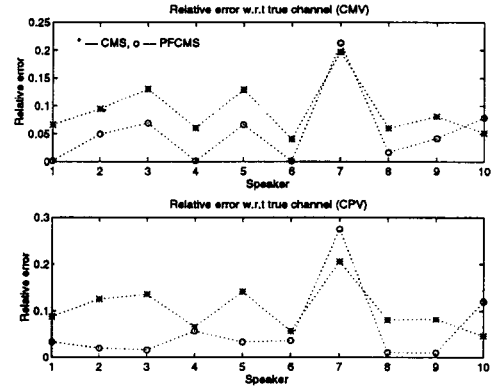


Figure 5: Relative error due to ordinary cepstral mean and pole-filtered cepstral mean.

$\mathbf{c}_S$, can be formulated as,

$$Rel.Error(CMS) = \frac{\|\mathbf{c}_S - \hat{\mathbf{h}_{ch}}\|}{\|\hat{\mathbf{h}_{ch}}\|}, \quad (5)$$

whereas, for the pole-filtered cepstral mean $\mathbf{c}_S^{pf}$,

$$Rel.Error(PFCMS) = \frac{\|\mathbf{c}_S^{pf} - \hat{\mathbf{h}_{ch}}\|}{\|\hat{\mathbf{h}_{ch}}\|}. \quad (6)$$

The relative errors have been plotted in Figure (5) for training utterances for ten speakers from the training set chosen for the simulated channel experiment. One can observe that the relative error due to the pole-filtered channel estimate is smaller than ordinary channel estimate for two simulated channels, CMV and CPV.

4. EXPERIMENTAL RESULTS

In this section experiments on closed set text-independent speaker identification have been presented on two standard speech corpuses, the TIMIT database and the KING database. Speech in the TIMIT was first downsampled and passed through a telephone channel simulator [5]. A VQ-based classifier is used for classification.

Method	Training	Testing	Accuracy(%)
LPCC-MR	CMV	CMV	63.1
PFCC-MR	CMV	CMV	69.5
LPCC-MR	CPV	CPV	62.1
PFCC-MR	CPV	CPV	68.9
LPCC-MR	CMV	CPV	59.4
PFCC-MR	CMV	CPV	64.7
LPCC-MR	CPV	CMV	56.8
PFCC-MR	CPV	CPV	62.6

Table 1: TIMIT experiments.

Experiments on TIMIT database

The Speaker Identification experiment on TIMIT consists of 38 speakers from the New England Dialect. For each speaker there are 10 sentences, five of which are concatenated and used for training, while the remaining five are used for testing. The training data is typically 8–10 seconds for every speaker and the testing data varies from 0.7–3 seconds. The downsampled speech is filtered through the telephone channel simulator [5] by either the CMV or CPV channels. Two sets of experiments are conducted on the TIMIT database by training and testing on the same telephone channel and across telephone channels. The results have been tabulated in Table 1. Ordinary mean removal has been abbreviated as LPCC-MR and Pole-filtered cepstral mean removal as PFCC-MR in the tables.

Experiments on KING database

Results have been reported on the San Diego portion of the KING database. Sessions 1–5 form one group and 6–10 form the second. Experiments within a group are experiments *within the great divide*, and across the group as experiments *across the great divide*, which imply considerable channel mismatch across the groups. A pole-based frame selection (FS) process [11] is used to eliminate these undesirable frames for experiments *across the divide* after standard energy-based silence removal. The results have been compared to ordinary mean removal in Table 2 and 3. Pole bandwidth thresholds in all experiments were chosen in the range, $\alpha \in [0.85, 0.9]$. Comparable results are also obtained by broadening pole bandwidths using weighted predictor coefficients.

Method	Identification rate	Accuracy(%)
LPCC-MR	383/520	73.6
PFCC-MR	404/520	77.7

Table 2: KING experiments, within the great divide.

5. CONCLUSION AND FUTURE WORK

A new method for normalizing channel distortions has been presented. By studying the effect of poles on channel estimates, a new algorithm is proposed to improve the channel normalization using a refined Cepstral Mean Subtraction. The ceptral mean estimate is improved by introducing the concept of pole-filtered cepstral coefficients. The ordinary long-term mean removal when replaced by long-term mean of pole-filtered cepstral coefficients, is shown to improve the performance of speaker identification systems. Future work will focus on adaptive pole thresholds so as to optimally decouple the channel information and speech information from the cepstral mean estimate.

Method	Identification rate	Accuracy(%)
LPCC-MR	314/650	48.2
PFCC-MR	346/650	53.2
FS + PFCC-MR	366/650	56.3

Table 3: KING experiments, across the great divide.

6. REFERENCES

1. B. Atal. "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *JASA*, 55:1304–1312, June 1974.
2. K. Paliwal. "On the performance of the quefrency weighted cepstral coefficients in vowel recognition", *Speech Commun.* , 1:151-154, 1982.
3. D. Naik, K. Assaleh and R. Mammone. "Robust speaker identification using pole filtering", *Proc. ESCA Workshop on Speaker Recognition*, Martigny, Switzerland, April 1994.
4. M. Schroeder. "Direct (nonrecursive) relations between cepstrum and and predictor coefficients", *IEEE ASSP*, 29:297–301, April 1981.
5. J. Kupin, "A wireline Simulator [Software]", CCR-P, April 1993.
6. S. Furui. "Cepstral analysis technique for automatic speaker verification", *IEEE ASSP*, 29:254–272, April 1981.
7. B. Juang, L. Rabiner and J. Wilpon. "On the use of Bandpass Liftering in Speech recognition", *IEEE ASSP*, 35:947-954, July 1987.
8. B. Atal, J. Remde. "A new model of LPC excitation for producing natural sounding speech at low bit rates", *Proc. ICASSP*, pp.614–617, Paris, May 1982.
9. H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. " RASTA-PLP Speech Analysis Technique", *Proc. ICASSP*, pp.121–124, San Francisco, 1992.
10. D. Naik, R. Mammone. "Channel normalization using Pole-filtered Cepstral Mean Subtraction", *Proc. SPIE* , Vol. 2277, July 1994.
11. K. Assaleh, R. Mammone. "Robust features for Speaker Recognition", *Proc. ICASSP*, April 1994, Australia.