ABSTRACT

In this paper, we propose a method to remove the residual speech effects of the channel cepstrum for speaker recognition in the Cepstral Mean Subtraction framework. The proposed Formant-Broadened CMS (FBCMS) is based on the facts that the formants can be found easily in log spectrum which is transformed from the cepstrum and the formants correspond to the dominant poles of all-pole model which is usually modeled vocal tract. The FBCMS evaluates only poles to be broadening from the log spectrum without polynomial factorization and makes a formant-broadened cepstrum by broadening the bandwidths of formant poles. Using 8 simulated telephone channels, we compared the relative errors of estimating channel cepstrum, speaker identification and computational efficiency for CMS, PFCMS, and the proposed method respectively on two databases. The proposed method has shown to yield improved speaker recognition rates with lower computational burden.

1. INTRODUCTION

There are many channel normalization methods to compensate a mismatch between training and testing conditions for speech and speaker recognition[1]. CMS(Cepstral Mean Subtraction) or CMN(Cepstral Mean Normalization) is well known method to compensate the mismatch. Based on the assumption that an ensemble average of speech cepstra is zero, CMS removes channel effect which is estimated by averaging the cepstra of the speech through the channel.

The basic framework of CMS is composed of a cepstrum transformation, channel estimation and a subtraction of channel cepstrum from the corrupted cepstrum as shown in the Figure 1.

In practical condition, however, the amounts and categories of speech for training and testing are limited so that the assumption can hardly be true. Therefore improperly estimated channel cepstrum in CMS tends to attenuate a useful spectral information.

To overcome this manifest weak point of CMS, there has been several researches for estimating channel, such as MMCMNFW[2] which modifies the channel cepstral mean from log spectrum, PFCMS[3] which adopts pole-filtering in linear prediction model, CSAR[4] which is applicable when stereo data are available and the utterances for training and testing are similar, and DPCMS[5] which uses cepstral difference of stereo data.

Figure 1. The framework of CMS

The Naik’s PFCMS out of these researches is based on the simple framework of CMS as known as an effective method which estimate channel effect without a priori information of channel like stereo data. But the PFCMS requires much computational complexity to find dominant poles with narrow bandwidth among the all of the poles evaluated by factorization of linear prediction (LP) polynomial.

We describe new method to estimate channel accurately with less computational complexity, inherited from the simple framework of CMS and the pole-filtering approach of PFCMS.

2. POLE-FILTERING FOR CHANNEL NORMALIZATION [3]

In order to preserve a spectral distribution of speech when the cepstral subtraction is carried out for channel compensation by cepstral mean, we have to reduce a residual speech before averaging cepstrum in each frame.

Naik has proposed a pole-filtering method using the LP model which de-emphasizes the effect of the dominant modes on the all-pole model of vocal tract because the dominant poles with narrow bandwidth contribute to biased spectral content in average of cepstra.

The PFCMS is applied to the all-pole model based on LP analysis that is generally used in speech or speaker
recognition. All pole model of vocal tract with $p$ roots can be expressed as the following equation

$$V(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{\prod_{k=1}^{p} (1-z_k z^{-1})}$$  \hfill (1)

where $a_k$, $k = 1, 2, ..., p$ are the coefficients of the LP model and $z_k$ are the roots of the model. These roots form the dominant modes in modeling the speech segment.

The pole-filtering is accomplished by inflating the bandwidths of the narrow-band poles while their frequencies are left unchanged. The processing has the effect of selective bandwidth broadening, the narrow band poles are shifted inward away from the unit circle along the same radius, referred to as an $\alpha$-method.

To implement the $\alpha$-method, we evaluate the roots of the channel cepstrum properly. We transform the prediction coefficients of Figure 2(a) to cepstrum and present the log spectra correspondent to the cepstrum, referred to as a $\gamma$-method. Then corresponding cepstral transformation is $c_{\text{PC}}(n) = \gamma c_{\text{LPC}}(n)$, where $\gamma = e^{i \frac{\delta}{j B}}$, based on $\delta$ Hz, the frequency with which the bandwidth is to be broaden.

While the $\gamma$-method has less computational complexity than the $\alpha$-method, it highly depends on only the value of $\gamma$ to get effect of pole-filtering and has difficulty of accurate filtering. Hence we need more fast method which estimates the channel cepstrum properly.

3. FORMANT BROADENED CMS

We show visually the effect of pole-filtering using the all-pole model of vocal tract, the frequency response of LTI system in Figure 2.

To investigate the effect, the experiments are carried out by pole-filtering the voiced part of speech using the both $\alpha$ and $\gamma$ methods. We have determined the values of $\alpha$ and $\gamma$ as 0.85 for thresholding in both methods and the 12th order of prediction coefficient by LP analysis for the all-pole model. The results are shown in Figure 2(a). One can observe that the effect of pole-filtered poles results in the broadened formants because the formants of speech correspond to the poles of the transfer function $V(z)$ [6].

We can get the log spectrum of channeled speech, $X(e^{j\omega})$ from the complex cepstrum, $\hat{x}(n)$ that is transformed from the LPC coefficients by the following definition [6]

$$\hat{x}(n) = \alpha_n + \sum_{k=1}^{p} \left( \frac{n}{k} \right) \hat{x}(k) \alpha_{n-k} \quad 1 \leq n$$ \hfill (2)

$$\hat{X}(e^{j\omega}) = \log |X(e^{j\omega})| + j \arg[X(e^{j\omega})]$$ \hfill (3)

$$\hat{X}(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{x}(n) e^{j\omega n} d\omega$$ \hfill (4)

where $\alpha_n$, $k = 1, 2, ..., p$ are the coefficients of the LP model.

We transform the prediction coefficients of Figure 2(a) to cepstrum and present the log spectra correspondent to the cepstrum by the Fourier transform in Figure 2(b). One can also observe that the locations of formant in Figure 2(a) are coincident with one of smoothed formant in Figure 2(b) and the effect of the pole-filtering in Figure 2(a) can be discovered almost identically in Figure 2(b).

Exploited this relation a filtered dominant poles among the broadened formants, proposed strategy consists of formant-picking which finds the formant to be broadening from the log spectrum and formant broadening in cepstrum domain by transforming the roots to cepstrum.

3.1. Formant picking from log spectrum

The above experiments lead to the idea that we can easily find the formants without pitch information using a simple peak-picking process in log spectrum domain.

In formant picking, one process evaluates the derivatives of the spectrum envelope and selects the location where the direction changed finds the center frequency, $\omega_k$. And the other process searches the frequencies in the vicinity of center frequency to find where the gain is 3 dB lower than one at center frequency. The bandwidth $B_k$ is justified by the difference of 2 frequencies.

From the location of formant, the root $\hat{z}_k$ is given by

$$\hat{z}_k = \left\{ \begin{array}{ll} e^{-j\omega_k} & , B_k > B_{TH} \\ e^{-j\omega_k} & , B_k \leq B_{TH} \end{array} \right.$$ \hfill (5)
where $B_{TH}$ is the bandwidth to be broadening.

Consequently, in the proposed method, we can select the formants to be broadening by the threshold of bandwidth $B_{TH}$ and ignore the formants of narrower bandwidth than the value of $B_{TH}$. Thus the process is more efficient to select the dominant poles as compared with the factorization of LP model. When we perform the filtering of the dominant poles, we must find all of the poles to threshold dominant poles.

### 3.2. Formant-broadening in cepstrum

To broaden the formants with narrow bandwidth and to transform the broadened formant to cepstrum which has only not channel effect, we make use of the relation of the roots of LP polynomial, $z_{ik}$ between the LP derived cepstra, $c_{ik}$. The relation is given by [7]

$$c_{ik} = \frac{1}{n} \sum_{j=1}^{n} z_{ik}^j = \frac{1}{n} \left( z_{ik}^1 + z_{ik}^2 + z_{ik}^3 + \ldots + z_{ik}^n \right) \tag{6}$$

If we have $K$ roots, $z_{ik}$ to be broadening, we can show that the modified cepstrum, $\tilde{c}_{ik}$ become

$$\tilde{c}_{ik} = c_{ik} - \frac{1}{n} \sum_{j=1}^{n} z_{ik}^j + \frac{1}{n} \sum_{k=1}^{K} \tilde{z}_{ik} \tag{7}$$

where $\tilde{z}_{ik} = e^{-2\pi i \omega_k} z_{ik}$ and $\tilde{z}_{ik} = e^{-2\pi i \omega_k}$.

The summary of the FBCMS process to evaluate formant-broadened cepstrum is as follows

a. Find the formant to be broadening in log spectrum domain (where, the center frequency of formant is $\omega_k$ and its bandwidth is $B_k$).

b. Determine whether the formant is to be broadening or not by $B_k$.

c. Transform the broadened formant to cepstrum by substituting the roots $z_{ik}$ with $\tilde{z}_{ik}$.

In the implementation, the number of determined formants is considered as less than the half of order of LPC due to their conjugates. And the experiments show that 2 to 4 formants can be found from the cepstrum of 12th order LPC.

### 4. EXPERIMENTAL RESULTS

The experiments for the relative errors of estimating channel cepstrum and closed-set speaker identification task were conducted on two databases.

The first database was collected from 6 men and 6 women in a soundproof booth. Each of 12 speakers was assigned with the utterances of a combination lock phrases, three 2-place numbers, such as YOHO. The speech data were digitized by 8Khz rate, 16-bit resolution. And the other database, clean speech utterances were selected from the 38 speakers in the TRAIN section of the DR1 subset of the TIMIT.

To simulate utterance passed over a telephone line, utterances were filtered by 8 simulated telephone channels, such as CMV, CPV, EMV, EPV, LC1, LC2, LC4 and LC30, each with unique characteristics, such as pass bandwidth, pass band attenuation, transition band fall off and so on [8].

In experiments, the 12th order LPC analysis was done on the 30msec frame with 10 msec overlapping. The log spectrum was transformed from LP based cepstrum by 256 size FFT. The eigen value method that evaluates complex roots from real coefficients of LP polynomial has been applied to implement the $\alpha$-method for LP polynomial factorization.

#### 4.1. Channel Estimation

In the first experiment, the relative errors of estimating channel cepstrum from the speech cepstrum filtered by 8 channels were evaluated for CMS, $\alpha$, $\gamma$ and FBCMS respectively. Relative error can be represented as the following equation (10).

$$\text{Relative error} = \frac{\| \tilde{c}_{ik} - c_{ik} \|}{\| c_{ik} \|} \tag{8}$$

where $c_{ik}$ is the cepstrum of real channel and the $\tilde{c}_{ik}$ is the cepstrum of estimated channel, that is, the average cepstrum of channel passed speech.

Table 1 shows the average relative errors of 10 utterances of 12 speakers for each channel. We can see that proposed method is superior to $\gamma$-method and similar to $\alpha$-method for all channels. As can be seen from Table 1, the channel estimation of the FBCMS has good performance in channels with broad pass band, while the performance was poor in channels with narrow pass band and slow fall off in transition band such as CPV, EPV and LC2.

#### Table 1: Relative Error for channel estimation method

<table>
<thead>
<tr>
<th>Channel</th>
<th>CMS</th>
<th>PFCMS $\gamma$</th>
<th>PFCMS $\alpha$</th>
<th>FBCMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMV</td>
<td>0.797</td>
<td>0.778</td>
<td>0.734</td>
<td>0.663</td>
</tr>
<tr>
<td>CPV</td>
<td>0.686</td>
<td>0.622</td>
<td>0.602</td>
<td>0.591</td>
</tr>
<tr>
<td>EMV</td>
<td>0.992</td>
<td>0.955</td>
<td>0.904</td>
<td>0.751</td>
</tr>
<tr>
<td>EPV</td>
<td>1.138</td>
<td>1.028</td>
<td>0.984</td>
<td>0.877</td>
</tr>
<tr>
<td>LC1</td>
<td>0.643</td>
<td>0.678</td>
<td>0.642</td>
<td>0.632</td>
</tr>
<tr>
<td>LC2</td>
<td>0.804</td>
<td>0.761</td>
<td>0.714</td>
<td>0.683</td>
</tr>
<tr>
<td>LC4</td>
<td>0.601</td>
<td>0.844</td>
<td>0.783</td>
<td>0.697</td>
</tr>
<tr>
<td>LC30</td>
<td>2.183</td>
<td>0.661</td>
<td>0.616</td>
<td>0.592</td>
</tr>
<tr>
<td>Average</td>
<td>0.818</td>
<td>0.791</td>
<td>0.747</td>
<td>0.686</td>
</tr>
</tbody>
</table>

#### 4.2. Speaker Identification

In the second set of experiments, the closed-set speaker identification task, we used a GMM based classifiers with the 46 Gaussian mixtures for each speaker.[9] All utterances were down-sampled to 8Khz and filtered by one of 8 channels from the simulated channels. Five utterances (SX)
for each speaker were used for training while the rest were used for testing.

According to our results in Table 2, the speaker recognition rates of the CMS were lower than those of the proposed method except EPV/CPV cross channel condition. Both PFCMS and FBCMS used the value of the threshold as 0.85 for pole-filtering and formant-broadening. Meanwhile it is interesting us that PFCMS $\gamma$-method showed the higher rate than that of $\alpha$-method in some condition.

Table 2 Results of closed-set speaker identification

<table>
<thead>
<tr>
<th>Train/Test</th>
<th>CMS</th>
<th>PFCMS $\gamma$</th>
<th>PFCMS $\alpha$</th>
<th>FBCMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMV/CPV</td>
<td>46.8%</td>
<td>50.5%</td>
<td>50.0%</td>
<td>52.1%</td>
</tr>
<tr>
<td>CPV/CMV</td>
<td>40.5%</td>
<td>43.2%</td>
<td>45.8%</td>
<td>41.1%</td>
</tr>
<tr>
<td>EMV/EPV</td>
<td>52.6%</td>
<td>54.2%</td>
<td>47.4%</td>
<td>52.6%</td>
</tr>
<tr>
<td>EPV/EMV</td>
<td>48.4%</td>
<td>47.9%</td>
<td>49.5%</td>
<td>51.6%</td>
</tr>
<tr>
<td>CPV/EPV</td>
<td>40.0%</td>
<td>43.7%</td>
<td>41.6%</td>
<td>45.8%</td>
</tr>
<tr>
<td>EPV/CPV</td>
<td>48.4%</td>
<td>39.5%</td>
<td>41.1%</td>
<td>42.6%</td>
</tr>
</tbody>
</table>

4.3. Computational Complexity

In another experiment, the relative errors of the estimated channel using the FBCMS method have been compared with the several different sizes of FFT. Figure 3 shows the relative errors in FFT with the sizes 64, 128, 256 and 512, respectively. The results show that the channel estimation performance is improved by increasing the FFT size because the resolution of frequency axis becomes higher with the increase in the size of FFT. (figure 4)

Finally we compared the computational complexity of the proposed method, which performs the FFT, with that of $\alpha$-method, which evaluates LP polynomial factorization. The $\alpha$-method has the complexity of $O(n^3)$, where $n$ is the order of LP coefficients, and the complexity for the proposed method that has performed Fourier transform by FFT is $O(N \log N)$, where $N$ is the size of FFT.

We evaluated the computational complexity in MATLAB. Using the 12th order of LP coefficients and 256 size FFT, the FBCMS occupies 62.5% less FLOPS than that of PFCMS.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we have proposed a method to remove the residual speech effects of the cepstral mean to estimate channel effect properly with lower computational complexity.

Exploited the relation a filtered dominant poles between a broadened formants, we have proposed the strategy consist of formant picking and formant broadening in cepstrum domain. The experiments have shown that the proposed FBCMS is superior to conventional CMS and PFCMS for performance of channel estimation and speaker identification rates.

Our future works will carry out the experiments to evaluate the performance of proposed method for speaker verification and apply the methodology of this approach to the other cepstrum like as MFCC.

6. REFERENCES


Figure 3. Relative Errors for FFT sizes