CHANNEL-ROBUST SPEAKER IDENTIFICATION USING MODIFIED-MEAN CEPSTRAL MEAN NORMALIZATION WITH FREQUENCY WARPING

Alvin A. Garcia

SpeakEZ/T-NETIX, Inc.
371 Hoes Lane
Piscataway, NJ 08854
alvin@caip.rutgers.edu

Richard J. Mammone

CAIP Center
Rutgers University
Piscataway, NJ 08854
mammone@caip.rutgers.edu

ABSTRACT

The performance of automatic speaker recognition systems is significantly degraded by acoustic mismatches between training and testing conditions. Such acoustic mismatches are commonly encountered in systems that operate on speech collected over telephone networks, where different handsets and different network routes impose varying convolutional distortions on the speech signal.

A new algorithm, the *Modified-Mean Cepstral Mean Normalization with Frequency Warping* (MMCMNFW) method, which improves upon the commonly-employed Cepstral Mean Subtraction method, has been developed. Experimental results on closed-set speaker identification tasks on a channel-corrupted subset of the TIMIT database and on a subset of the NTIMIT database are presented. The new algorithm is shown to offer improved recognition rates over other existing channel normalization methods on these databases.

1. CHANNEL MISMATCH COMPENSATION

The channel normalization method presented in this paper extends upon two existing channel normalization techniques: *cepstral mean subtraction* (CMS), also known as *cepstral mean normalization* (CMN), and *frequency warping*. A brief overview of both methods is presented first.

1.1. Cepstral Mean Subtraction

Cepstral Mean Subtraction is one of the earliest and most popular methods employed to ameliorate the effects of channel variability in speaker and speech recognition systems. CMS feature vectors are computed from the N cepstral vectors $\vec{c}_{y;i}, \quad i = 1, \ldots, N$ from a channel-corrupted speech utterance $y(n)$ by subtracting the cepstral mean, or average of all N cepstral vectors, from each of the original cepstral vectors $\vec{c}_{y;i}$:

$$\vec{c}_{cms;i} = \vec{c}_{y;i} - \vec{c}_{y;avg}, \qquad i = 1, \ldots, N, \tag{1}$$

where $\vec{c}_{y;avg} \triangleq \frac{1}{N}\sum_{i=1}^{N} \vec{c}_{y;i}$. The principle behind this approach is based upon the behavior of the cepstrum under convolutional distortions, and the assumption that the channel filter $h(n)$ does not vary significantly over the duration of the utterance, i.e. $h(n)$ is a linear, time-invariant filter.[1] As is well known, a convolutional distortion in the time domain, such as that introduced by a channel, corresponds to an additive bias component in the cepstral domain. That is, if we denote the clean speech signal, prior to corruption by the channel $h(n)$, by $s(n)$, and the channel-corrupted speech signal by $y(n)$, then

$$y(n) = s(n) * h(n) \Longleftrightarrow \vec{c}_y = \vec{c}_s + \vec{c}_h, \tag{2}$$

with $*$ denoting linear convolution and $\vec{c}_y$, $\vec{c}_s$, and $\vec{c}_h$ denoting the corresponding cepstral features. Now, taking the utterance-long time averages of both sides of this cepstral relation, we have

$$\begin{aligned} \vec{c}_{y;avg} &\triangleq \frac{1}{N}\sum_{i=1}^{N} \vec{c}_{y;i} \\ &= \frac{1}{N}\sum_{i=1}^{N} \vec{c}_{s;i} + \frac{1}{N}\sum_{i=1}^{N} \vec{c}_{h;i} \\ &\triangleq \vec{c}_{s;avg} + \vec{c}_{h;avg}. \end{aligned} \tag{3}$$

Since it is assumed that the channel does not vary over the duration of the utterance, the last summation becomes simply $\frac{1}{N}\sum_{i=1}^{N} \vec{c}_h$, or just $\vec{c}_h$, the cepstrum of the channel. The middle summation corresponds to the cepstral mean of the *clean* (not channel-corrupted) speech signal $s(n)$. If the distribution and variety of sounds in $s(n)$ is such that the average spectrum over the utterance is relatively flat, then the corresponding cepstral mean vector will go to zero, i.e.

$$\vec{c}_{s;avg} \Rightarrow \vec{0}. \tag{4}$$

The cepstral mean of the channel-corrupted utterance $y(n)$, as given by Eq. 3, then becomes

$$\begin{aligned} \vec{c}_{y;avg} &= \vec{c}_{s;avg} + \vec{c}_{h;avg} \\ &= \vec{0} + \vec{c}_h \\ &= \vec{c}_h, \end{aligned} \tag{5}$$

which is just the cepstrum of the channel. Thus removing the time average of all the cepstra from the cepstrum of each frame, as suggested by Eq. 1, corresponds to subtracting the cepstrum of the channel. We can rewrite Eq. 1 as

$$\begin{aligned} \vec{c}_{cms;i} &= \vec{c}_{y;i} - \vec{c}_{y;avg} \\ &= (\vec{c}_{s;i} + \vec{c}_h) - \vec{c}_h \\ &= \vec{c}_{s;i}. \end{aligned} \tag{6}$$

[1] This is generally a fair assumption for utterances lasting several seconds, which is typical in many speaker recognition scenarios.

The resulting CMS cepstral vectors, $\vec{c}_{cms;i}$, no longer dependent on $\vec{c}_h$, are thus invariant to any channel present. This can also be seen by noting that the subtraction of $\vec{c}_h$ in the cepstral domain corresponds to deconvolution with $h(n)$ in the time domain, or multiplication by $\frac{1}{|H(e^{j\omega})|}$ in the frequency domain.

1.2. Frequency Warping

Frequency warping is a standard frequency domain signal processing technique used to limit subsequent processing to a particular range of frequencies within the Nyquist interval. The basic principle in frequency warping is to map the desired frequency range of interest, $[\omega_{min}, \omega_{max}]$, to the entire Nyquist interval, $[0, \pi]$ radians.[2] Such a mapping can be accomplished trivially by the following affine transform:

$$\omega' = \frac{\omega - \omega_{min}}{\omega_{max} - \omega_{min}}\pi, \tag{7}$$

where ω denotes the original radian frequency, and ω' denotes the new radian frequency.

In the context of robust speaker recognition over bandwidth-limited channels, such as with telephone speech, frequency warping is used prior to feature extraction to map the passband of the channel to the entire Nyquist interval. For instance, in telephone speech, the nominal bandwidth of the channel is approximately [300, 3200] Hz. Frequency warping is used to map this interval to $[0, f_s/2]$ Hz, the Nyquist interval. In this manner, signal energy outside the passband, which is bound to be relatively low in energy and highly sensitive to noise, is completely ignored. In [7], it was found that frequency warping alone offered significant performance gains on a speaker identification task on telephone speech.

2. MODIFIED-MEAN CEPSTRAL MEAN NORMALIZATION WITH FREQUENCY WARPING

In spite of its apparent simplicity, Cepstral Mean Subtraction has been found to be a very effective method for combating the effects of utterance-to-utterance channel variation [1], [7]. However, it is not without its drawbacks. In practice, it has been found that the clean speech cepstral mean, given by $\vec{c}_{s;avg}$ in Eq. 3, does not go to zero, as implied by Eq. 4. Recall that since the cepstrum is the inverse Fourier transform of the log magnitude spectrum, a zero-vector cepstral mean corresponds to a "white", or flat, average spectrum. Fig 1 shows the cepstral mean and corresponding log magnitude spectrum of a 15 second long utterance of clean (not channel-corrupted), read speech. As is evident from this figure, the cepstral mean is non-zero, and the corresponding log magnitude spectrum is not flat. Taking this fact into account, Eq. 5 becomes

$$\vec{c}_{y;avg} = \vec{c}_{s;avg} + \vec{c}_h \tag{8}$$

and the CMS vectors given by Eq. 6 become

$$\begin{aligned} \vec{c}_{cms;i} &= \vec{c}_{y;i} - \vec{c}_{y;avg} \\ &= (\vec{c}_{s;i} + \vec{c}_h) - (\vec{c}_{s;avg} + \vec{c}_h) \\ &= \vec{c}_{s;i} - \vec{c}_{s;avg}. \end{aligned} \tag{9}$$

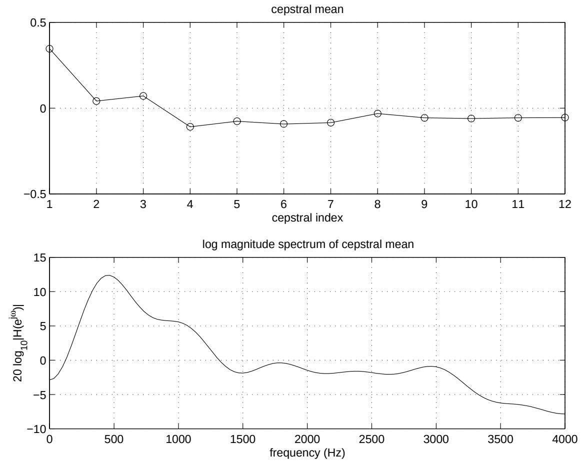

Figure 1: Top: cepstral mean of an utterance of clean speech; Bottom: log magnitude spectrum corresponding to same cepstral mean

From this expression, we see that in the CMS vectors $\vec{c}_{cms;i}$, the dependence on the channel cepstrum, $\vec{c}_h$, is still removed. As the channel may vary from utterance to utterance, and contains no information about the identity of the speaker, this quality is desirable.[3] However, additional information, namely the clean speech cepstral mean, $\vec{c}_{s;avg}$, is also removed. This term represents the long-term average spectrum of the clean speech signal, prior to corruption by the channel. It can also be interpreted as the cepstrum corresponding to the average vocal tract shape of the speaker [8]. This component is what is shown in Fig 1. Unlike the channel cepstrum, this term offers information which *is* useful for speaker recognition. In fact, some early efforts in speaker recognition used these long-term spectral averages as features [2], [5]. Thus, subtraction of this information in the CMS feature vectors results in the discarding of speaker-dependent information, which would otherwise be useful in identifying the speaker. As a result, overall speaker recognition performance falls short of that which could be obtained without the removal of such information. This observation is documented in [3], for instance.

The basic principle of the *Modified-Mean Cepstral Mean Normalization with Frequency Warping* (MMCMNFW) method is to modify the cepstral mean $\vec{c}_{y;avg}$ in such a way as to reduce the component due to the clean speech cepstral mean, $\vec{c}_{s;avg}$, while leaving mostly the component due to the channel, $\vec{c}_h$, an idea suggested in [6]. That is, we would like to create a modified cepstral mean $\vec{c}_{y;MMCM} \approx \vec{c}_h$. Then, performing Cepstral Mean Subtraction with the modified cepstral mean should remove the channel bias from the resulting CMS features, while leaving intact the component due to the spectral average of the clean speech. In the MMCMNFW method, such manipulation of the cepstral mean is performed in the log magnitude spectral domain on $Y_{avg}(e^{j\omega})$, the Fourier transform of the cepstral mean $\vec{c}_{y;avg}$.

An analysis was performed comparing the graphical relationship between the log magnitude spectra $Y_{avg}(e^{j\omega})$ of the cepstral means of clean speech utterances which had been passed through

[2] Note that for simplicity of presentation, we are only considering the positive frequency half of the Nyquist interval. The same operations apply symmetrically to the negative frequency half.

[3] It should be noted that in certain applications, such as platform identification in military communications, it *is* desirable to preserve information about the channel.

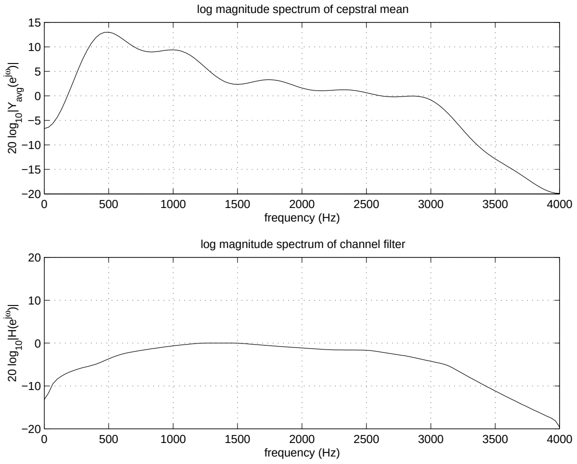

Figure 2: Top: log magnitude spectrum of cepstral mean of an utterance of channel-corrupted speech. Bottom: log magnitude spectrum of the corrupting channel filter

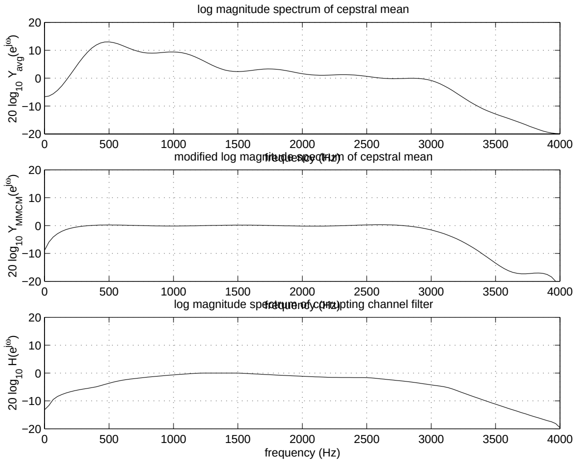

Figure 3: Top: log magnitude spectrum of cepstral mean; Middle: above with "Mean Modifications"; Bottom: log magnitude spectrum of actual corrupting channel filter

various linear time-invariant channels, $h(n)$, and the log magnitude spectra $H(e^{j\omega})$ of those channels. Speech utterances from the TIMIT speech database and telephone channel impulse responses from the Wireline telephone channel simulator [4] were used. Fig. 2 illustrates a typical case: The top frame shows the log magnitude spectrum $Y_{avg}(e^{j\omega})$ of the cepstral mean of the channel-corrupted speech utterance, and the bottom frame shows the log magnitude spectrum $H(e^{j\omega})$ of the channel used to corrupt the speech. Based upon the analysis of many such plots, the following steps were empirically determined to improve the approximation of the channel spectrum $H(e^{j\omega})$ by the cepstral mean log spectrum $Y_{avg}(e^{j\omega})$:

1. Clip any portion of the curve $Y_{avg}(e^{j\omega})$ above the value 0 dB to 0 dB.[4]
2. Translate the portion of the curve in the range [0, 500] Hz up or down such that $Y_{avg}(e^{j2\pi 500}) = 0$ dB; i.e. $Y_{avg}(e^{j\omega}) =$ 0 dB at 500 Hz.
3. Replace the resulting modified curve by an order-11 polynomial fit.

The resulting modified $Y_{avg}(e^{j\omega})$ is designated as $Y_{MMCM}(e^{j\omega})$. Fig 3 shows the result of these processing steps on a typical cepstral mean. The top panel shows the log magnitude spectrum $Y_{avg}(e^{j\omega})$ of the cepstral mean of a channel-corrupted speech utterance, the middle panel shows the modified spectrum $Y_{MMCM}(e^{j\omega})$, which results after applying the modifications detailed above, and the bottom panel shows the response of the actual channel filter which was used to corrupt the original, clean speech signal. Note the closer resemblance to the channel filter after employing the modifications.

To perform Cepstral Mean Subtraction with $Y_{MMCM}(e^{j\omega})$, we would compute its Fourier transform to yield $\vec{c}_{y;MMCM}$, the corresponding modified "cepstral mean", and subtract $\vec{c}_{y;MMCM}$ from each frame's original cepstral vector. However, since such nonlinear spectrum modifications may introduce components which may no longer be accurately modeled by a finite length cepstral vector, as is used in CMS to model the average spectrum, the cepstral mean "subtraction" with $\vec{c}_{y;MMCM}$ is actually implemented as a time-domain filter representing the inverse of $Y_{MMCM}(e^{j\omega})$. Hence the choice of the term *normalization* over *subtraction* in the name MMCM*N*FW.

Finally, after performing Cepstral Mean Normalization with the modified cepstral mean, frequency warping is used to map the average passband of all the channels to the entire Nyquist interval. Note that this average passband frequency range can be determined empirically, as was done in the experiments presented in the following section, by visual inspection of the log magnitude spectra of the cepstral means of the channel-corrupted speech. For the databases used in these experiments, the passband of the channels was empirically determined to be [300, 3200] Hz. The motivation for such frequency warping was the observation that in the log magnitude spectra of the channel-corrupted cepstral means, the signal energy outside this frequency range dropped to very low levels relative to that in the passband. Even if the channel could be accurately estimated outside the passband, the inversion of such low-energy frequency regions of the signal spectra would result in significant amplification of noise in those frequency regions, an undesirable side effect. The use of frequency warping circumvents this problem by eliminating such out-of-passband components.

After performing Cepstral Mean Normalization with the modified cepstral mean, and then frequency warping the resulting speech signal, the standard cepstral feature vectors are computed on the signal output of the frequency warping procedure to yield the MMCMNFW feature vectors.

3. EXPERIMENTS

Closed-set speaker identification experiments were conducted on two databases. The first database, designated as W-TIMIT, consists of clean read speech utterances taken from the 38 speakers in the TRAIN section of the DR1 (New England Dialect) subset of the standard TIMIT database. These utterances are downsampled to

[4] Note than in practice, the 0th cepstral coefficient, representing log energy, is not retained. In converting the cepstral mean $\vec{c}_{y;avg}$ to $Y_{avg}(e^{j\omega})$, the value of the cepstral mean's 0th coefficient was set to zero.

method	recognition accuracy
baseline	44.7%
CMS	63.7%
frequency warping	65.2%
CMS + frequency warping	64.7%
MMCMNFW	82.1%

Table 1: Results on W-TIMIT database

method	recognition accuracy
baseline	50.5%
CMS	43.2%
frequency warping	64.7%
CMS + frequency warping	56.8%
MMCMNFW	67.9%

Table 2: Results on NTIMIT database

8 kHz, and then filtered by one of eight randomly-selected channel filters from the Wireline channel simulator previously cited. Five of each speaker's utterances were used to train a vector quantizer (VQ) codebook, and the remaining five were used for testing.[5] The utterances averaged approximately 3 seconds in duration (including silence). Results on W-TIMIT are summarized in Table 1, which compares the recognition rates achieved using no channel compensation (baseline), Cepstral Mean Subtraction (CMS), frequency warping to $[300, 3200]$ Hz (frequency warping), Cepstral Mean Subtraction followed by frequency warping to $[300, 3200]$ Hz (CMS+frequency warping) and Modified-Mean Cepstral Mean Normalization with Frequency Warping (MMCMNFW). The features used in all cases are FFT-derived cepstra, computed after the channel normalization method has been applied.

The second database consists of those utterances in the NTIMIT database directly corresponding to those used in the W-TIMIT database just described. NTIMIT consists of the original TIMIT utterances played back and resampled after transmission along various routes through the public telephone network. The results achieved on this database with the aforementioned channel normalization methods are summarized in Table 2.

As can be seen from Table 1, on the W-TIMIT database, all channel normalization methods yielded better recognition rates than the baseline result, where no normalization was used. The new MMCMNFW method offered the highest accuracy on this database. Examining Table 2, we observe the same pattern on the NTIMIT database, with the curious exception of Cepstral Mean Subtraction, which would be expected to outperform the baseline on such a channel-corrupted database. Again, MMCMNFW offered the highest recognition accuracy on this database.

4. CONCLUSION

A new method for performing channel normalization for automatic speaker recognition systems has been presented. The new method improves upon the standard Cepstral Mean Subtraction approach by 1) refining the log spectrum of the cepstral mean to be shaped more like a channel response, and 2) not attempting to invert low-energy frequency ranges in the average spectrum. Future work will address adaptation of the MMCMNFW method to communication channels other than those types encountered in telephone networks.

[5] Note that all five training utterances used to train a given speaker's VQ codebook were filtered with the same, randomly-selected channel filter.

5. REFERENCES

[1] Bishnu S. Atal. Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 64:460–475, April 1976.

[2] S. Furui, F. Itakura, and S. Saito. Talker recogntiion by long-time averaged speech spectrum. *Electronic Communications in Japan*, 55-A(10):54–61, 1972.

[3] H. Gish, K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, and J. Wolf. Investigation of text-independent speaker identification over telephone channels. *Proceedings ICASSP–1985*, 8:379–382, 1985.

[4] J. Kupin. Wire — a wireline simulator. CCR-P, April 1993. [software].

[5] J. Markel, B. Oshika, and Jr. A. Gray. Long-term feature averaging for speaker reognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 54–61, August 1977.

[6] D. Naik. Pole-filtered cepstral mean subtraction. *Proceedings ICASSP–1995*, 1:157–160, 1995.

[7] Douglas A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643, October 1994.

[8] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.