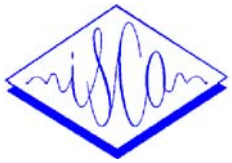

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

CHANNEL ESTIMATION AND NORMALIZATION BY COHERENT SPECTRAL AVERAGING FOR ROBUST SPEAKER VERIFICATION

Rajesh Balchandran, Vidhya Ramanujam, Richard J. Mammone

CAIP Center, Rutgers University,
Piscataway NJ 08854-8088, USA
brajesh, vidhyar, mammone@caip.rutgers.edu

10.21437/Eurospeech.1999-183

ABSTRACT

In real-world speech and speaker recognition systems, data is often recorded over commercial telephone lines. Consequently, differing transmission channels cause mismatch between training and testing conditions resulting in significant performance loss. This paper presents a new technique that uses complex spectral averaging to estimate the channel accurately. The estimated channel is used as an inverse filter for normalization. This technique being speech-in speech-out, can be used as the preprocessing stage in any automatic speech processing system. A refinement process is also presented that further improves the channel estimate. The combined technique is evaluated on a speaker verification task where the training and testing data were convolved with different telephone channels. The new technique provides excellent channel estimates and nearly restores performance back to that of clean conditions.

1. INTRODUCTION

Speech and speaker recognition systems undergo severe degradation due to mismatch between training and testing conditions caused by differing transmission channels. In order to reduce this mismatch it is necessary to compensate for these channel effects. This compensation is performed in the time domain or the feature domain. Time domain compensation consists of channel estimation followed by channel normalization.

Over the years a number of techniques [1],[4],[2] have been proposed to to reduce channel mismatch. We only review one of them – Cepstral Mean Normalization, which has proven to be the benchmark feature domain technique for channel normalization. We also discuss periodogram averaging as it forms the basis of our proposed technique.

This research effort was supported by the U.S. Air Force Research Laboratory at Rome, NY and the CAIP Center.

2. CEPSTRAL MEAN NORMALIZATION

Cepstral Mean Normalization (CMN) [1] is a channel normalization technique applied in the feature domain. It is based on the principle of *homomorphic deconvolution*. The cepstrum, $C_i(n)$, of a frame of speech with spectrum is defined as the inverse Fourier transform of the log magnitude spectrum and is given by,

$$C_i(n) = \mathcal{F}^{-1}\left(\ln|Y_i(\omega)|\right) \tag{1}$$

where $Y_i(\omega)$ is the spectrum of i^{th} speech frame. The channel corrupted speech can be represented as the convolution of clean speech, $x(n)$, with the channel impulse response $h(n)$. Therefore, in the spectral domain,

$$Y_i(\omega) = H(\omega)\ \ X_i(\omega) \tag{2}$$

where $H(\omega)$ is the channel spectrum and $X_i(\omega)$ is the clean speech spectrum of the i^{th} frame. Thus,

$$C_i(n) = \mathcal{F}^{-1}\left(\ \ln|H(\omega)|\ +\ \ln|X_i(\omega)|\ \right) \tag{3}$$

The mean cepstrum for a speech utterance having N frames is given by,

$$\overline{C}(n) = \mathcal{F}^{-1}\left(\ \ln|H(\omega)|\ \right) + \frac{1}{N}\sum_{i=1}^{N}\mathcal{F}^{-1}\left(\ \ln|X_i(\omega)|\ \right) \tag{4}$$

where, the channel is assumed to be time invariant, at least for the duration of the utterance. For long speech utterances, it is usually *assumed* that the clean speech energy is uniformly distributed over the entire range of the spectrum, making it flat. Therefore, the second term in the above expression is assumed to tend to zero and the cepstral mean of the channel corrupted utterance would represent the cepstrum of the channel itself. Thus, deconvolution of the channel from the signal is achieved by subtracting the the cepstral mean of the channel corrupted utterance from each cepstral vector. However, in most practical applications, the signal duration is not long enough for the above assumption to be valid and the subtraction removes the average clean speech cepstral mean in addition to the channel. In spite of this, CMN works remarkably well! This is because the subtraction *has*

to carried out on the training features *as well as* the test features, so that nearly similar average speech information gets removed from both there is minimal mismatch. This is particularly so for text-dependent applications.

3. PERIODOGRAM AVERAGING

Periodogram averaging has been used to estimate periodic signals buried in noise. For channel estimation one could consider the channel to be the desired signal, buried in "speech". Periodogram averaging consists of breaking up the signal into small segments and averaging their spectra, so that the variations due to noise are eliminated and the invariant spectral component is recovered.

The periodogram spectrum [5] for a frame i of speech containing M samples is given by,

$$S_i(\omega) = \frac{1}{M}|Y_i(\omega)|^2 \tag{5}$$

where $Y_i(\omega)$ is the spectrum of the i^{th} frame. M being a normalization factor will be ignored in subsequent calculations. The average periodogram $\overline{S}(\omega)$ for an utterance having N frames is given by,

$$\overline{S}(\omega) = |H(\omega)|^2 \ \frac{1}{N}\sum_{i=1}^{N}|X_i(\omega)|^2 \tag{6}$$

where, $Y_i(\omega)$ has been replaced using Eq. 2 and is once again assumed to be time-invariant.

For $\overline{S}(\omega)$ to represent the channel, we require the second term (the average of the clean speech component) to be constant at all frequencies. In practice this average is not constant and can be represented by an invariant component μ, and a variable component $\delta_i(\omega_j)$ that represents the deviation of the spectrum of the i^{th} frame from μ at frequency (ω_j). Thus, for the i^{th} frame, at any frequency ω,

$$X_i(\omega) = \mu + \delta_i(\omega) \tag{7}$$

$$\tfrac{1}{N}\textstyle\sum_{i=1}^{N}|X_i(\omega)|^2 = \tfrac{1}{N}\sum_{i=1}^{N}|\mu + \delta_i(\omega)|^2 \tag{8}$$

$$= |\mu|^2 + \tfrac{1}{N}\textstyle\sum_{i=1}^{N}|\delta_i(\omega)|^2 + 2\,|\mu|\,\tfrac{1}{N}\sum_{i=1}^{N}|\delta_i(\omega)| \tag{9}$$

In the above expression, if the two $\delta(\omega)$ terms were to go to zero, then the periodogram estimate $\overline{S}(\omega)$ in Eq.(6) would indeed represent the true channel spectrum $H(\omega)$, within a gain factor. However, as magnitudes are used in the average, the contribution due to $\delta(\omega)$ terms can never go to zero (except in the trivial case $\delta(\omega) = 0$). Therefore, the periodogram estimate *cannot* truly represent the channel.

4. PROPOSED METHOD: COHERENT SPECTRAL AVERAGING

From the above discussion, it is clear that magnitude averaging cannot yield good channel estimates. Therefore, a new averaging method called *Coherent Spectral Averaging* is proposed, that uses the complex spectrum instead of the magnitude spectrum. This allows the phase information to be preserved. Previously, we have used the coherent averaging scheme for blind channel estimation in a speaker identification task under channel mismatch with significant success [6]. In this paper, we present a modification of the original algorithm, that is applicable when a *reference* or training utterance is available in the testing phase. The use of the reference utterance overcomes some of the problems described in [6] and results in a better channel estimate.

4.1. Technique

Consider Eq. 2 once again. Representing the spectra by their real and imaginary parts,

$$Y_i^{Re}(\omega) + jY_i^{Im}(\omega) = \left[H^{Re}(\omega) + jH^{Im}(\omega)\right] \cdot \left[X_i^{Re}(\omega) + jX_i^{Im}(\omega)\right] \tag{10}$$

Taking the *coherent* mean over all frames, that is, computing the mean of the real and imaginary parts separately we get,

$$\overline{Y}^{Re}(\omega) + j\overline{Y}^{Im}(\omega) = \left[H^{Re}(\omega) + jH^{Im}(\omega)\right] \cdot \left[\ \overline{X}^{Re}(\omega) + j\overline{X}^{Im}(\omega)\ \right] \tag{11}$$

where the bar represents the average over all speech frames. Once again the channel has been assumed time-invariant, for the duration of the utterance.

We can solve for $H^{Re}(\omega)$ and $H^{Im}(\omega)$ by equating the real and imaginary parts of Eq. 11 at each frequency ω. In practice, the X terms, which correspond to the clean version of the channel corrupted test utterance, are not available, so we use a training or reference utterance instead. Therefore, Eq. 11 is modified to,

$$\overline{Y}^{Re}(\omega) + j\overline{Y}^{Im}(\omega) \approx \left[H^{Re}(\omega) + jH^{Im}(\omega)\right] \cdot \left[\overline{X}^{Re}_{Train}(\omega) + j\overline{X}^{Im}_{Train}(\omega)\right] \tag{12}$$

The approximation in Eq. 12 (and hence the channel estimate) becomes more accurate when the reference utterance used has the same text as the test utterance. The channel estimate obtained in this manner is usually quite noisy, so it needs to be smoothed. The smoothed estimate is converted to an inverse FIR filter which is then used for deconvolution of the channel corrupted test utterance.

5. REFINEMENT

An additional refinement step is proposed that improves the channel estimate even further. Let $H_{Initial}(\omega)$ represent the initial channel estimate obtained above, and let $H_{True}(\omega)$ represent the true channel. Their ratio,

$$\mathcal{R}(\omega) = \frac{H_{True}(\omega)}{H_{Initial}(\omega)} \tag{13}$$

represents the correction factor that maps the initial channel estimate to the true channel. Obviously, $H_{True}(\omega)$ cannot be used to compute $\mathcal{R}(\omega)$. However, it was empirically found that the ratio of the spectra obtained from the cepstral means of the restored test utterance (using the initial channel estimate), $\overline{Y}_{Restored}(\omega)$, to that of the reference utterance, $\overline{X}_{Reference}(\omega)$, closely followed the ideal $\mathcal{R}(\omega)$. That is,

$$\mathcal{R}(\omega) \approx \frac{\overline{Y}_{Restored}(\omega)}{\overline{X}_{Reference}(\omega)} \tag{14}$$

Therefore, the refined channel estimate is given by,

$$H_{Refined}(\omega) = H_{Initial}(\omega) \cdot \frac{\overline{Y}_{Restored}(\omega)}{\overline{X}_{Reference}(\omega)} \tag{15}$$

The combined technique is called *Coherent Spectral Averaging with Refinement* (CSAR).

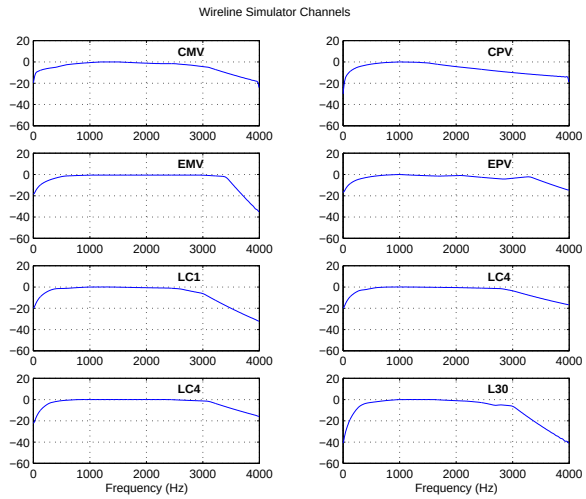

Figure 1: **Wireline Channels**

6. EXPERIMENTS

The CSAR approach was evaluated on a Speaker verification task using a 51 speaker text-dependent database [1]. Each speaker has 12 repetitions of the phrase "Rome Laboratory" (about 1 second in length). All the utterances were recorded through the same local telephone network at 8 kHz. The data can therefore be considered free of telephone channel effects.

[1] Obtained from U.S. Air Force

In order to obtain channel corrupted speech a set of 8 *Wireline* channels [3] that simulate different telephone channels were used. These channels have bandpass characteristics with varying cut-offs and roll-offs and are quite severe. They are shown in Fig. 1

6.1. Channel Estimation

Numerous experiments were carried out to study the effectiveness of the CSAR technique for channel estimation. Figure 2 shows the estimates for four of the wireline channels. In each case the channel was applied on the test utterance and a clean reference utterance (different from test) was used. As can be seen the estimate captures the overall profile of the channel and the roll-offs very effectively.

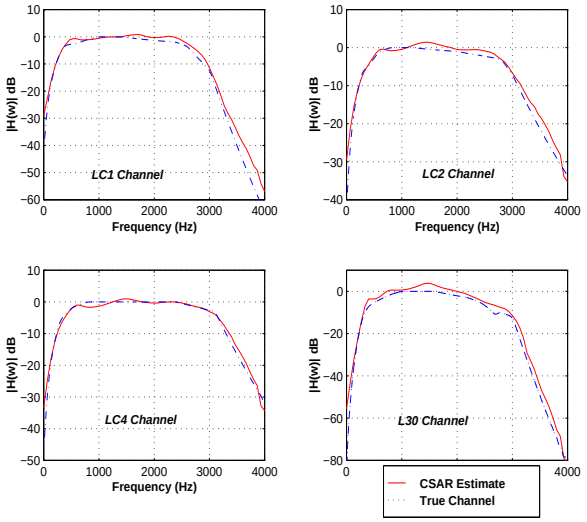

Figure 2: **Channel Estimates Using The CSAR Technique**

6.2. Speaker Verification

6.2.1. Clean Training

The first set of speaker verification experiments were performed by keeping the training environment clean and corrupting the test environment with different wireline channels. Twelfth order Linear Prediction (LP) derived cepstral features were used with Vector Quantization (64 codebooks per speaker) for classification. Four utterances were used for training and eight were used for testing.

The channel was estimated from each test utterance using the CSAR technique and then used to inverse filter it. This restored utterance was then used for classification. The base system (clean training, clean testing) yielded an Equal Error Rate (EER) of 2.36%. Table 1 compares the CSAR results with

those obtained using CMN for each of the wireline channels.

Test Channel	Equal Error Rate (EER)		
	No Norm.	CMN	CSAR
CMV	29.2 %	6.8 %	*3.4 %*
CPV	34.7 %	6.5 %	*3.2 %*
EMV	21.5 %	6.5 %	*4.6 %*
EPV	17.1 %	5.2 %	*3.2 %*
LC1	34.5 %	13.3 %	*3.5 %*
LC2	27.9 %	6.21 %	*2.9 %*
LC4	23.9 %	5.11 %	*3.0 %*
L30	37.1 %	14.8 %	*3.2 %*

Table 1: Comparison of Speaker Verification results on clean training data and channel corrupted test

6.2.2. Cross-Channel Experiments

In the second set of experiments one channel was applied on each training utterance and a different channel was applied on the test utterance. During testing a *difference* channel was estimated using the CSAR technique and used to inverse filter the test. Results for three channel combinations are shown below. Results obtained on data collected over a telephone network are also shown. The utterances were recorded using carbon button and/or electret handsets. This telephone database consisted of 13 speakers and training and testing comprised 4 utterances each.

Channel on		Equal Error Rate (EER)		
Train	Test	No Norm.	CMN	CSAR
CMV	CPV	27.4 %	4.2 %	*3.5 %*
EMV	CMV	14.6 %	3.5 %	*4.3 %*
CPV	EMV	41.5 %	13.3 %	*7.5 %*
Tel.	Tel.	17.9 %	6.9 %	*5.8 %*

Table 2: Comparison of Speaker Verification results under cross-channel conditions

From all the verification experiments, we see that,

- Channel mismatch results in significant performance degradation
- Cepstral Mean Normalization (CMN) results in appreciable improvement in performance only when applied on *both* training and testing.
- The Coherent Spectral Averaging Technique with Refinement (CSAR) technique yields a very good channel estimate and also results in significant improvement in performance – it almost restores performance to clean-clean values.

7. CONCLUSION

A new spectral domain technique for channel estimation and normalization has been presented. This coherent spectral averaging scheme was shown to provide excellent channel estimates. It also improved speaker recognition accuracy tremendously in a channel corrupted speaker verification task and almost restored performance to baseline values. It also consistently performed better than the feature domain approach of cepstral mean normalization. Unlike CMN, it is enough to apply this technique on the test only.

Being speech-in speech-out this technique can be used as a *preprocessing* stage in any automatic speech processing system. Unlike any feature domain approach, this spectral estimation technique is not restricted to any feature set or application. This approach is found to be effective for channel estimation in both short ($\approx$ 1 sec) and long duration utterances. Therefore it can also be used for channel estimation in a *Channel Identification* task.

8. REFERENCES

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

[2] Alvin A. Garcia and Richard J. Mammone. Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping. *Proceedings, ICASSP*, 1999.

[3] J. Kupin. Wire - a wireline simulator. Technical report, CCR-P, April, 1993.

[4] Devang Naik. Pole-filtered cepstral mean subtraction. *Proceedings, ICASSP*, 1995.

[5] Sophocles J. Orfanidis. *Optimum Signal Processing: An Introduction.* McGraw-Hill Publishing Company, 1988.

[6] Vidhya Ramanujam, Rajesh Balchandran, and Richard J. Mammone. Robust speaker identification under channel mismatch conditions: A new approach to blind channel estimation. *Proceedings of Audio-Video-Based Biometric Person Authentication(AVBPA)*, 1999.