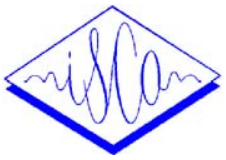

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ROBUST SPEAKER VERIFICATION IN NOISY CONDITIONS BY MODIFICATION OF SPECTRAL TIME TRAJECTORIES

Vidhya Ramanujam, Rajesh Balchandran and Richard J. Mammone

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

10.21437/Eurospeech.1999-192

ABSTRACT

Real-world speech and speaker recognition systems are often subject to ambient noise which results in significant performance loss, more so when the noise types and noise levels are different between training and testing. This paper presents a new preprocessing technique, *Coherent Spectral Modification*, that aims to reduce distortion due to noise by modifying the complex speech spectrum using information from non-speech regions of the spectral time trajectories. A refinement process is also proposed that further reduces noise mismatch. This combined technique was evaluated on a speaker verification task where the test data was corrupted with varying levels of white noise and pink noise. The new method yielded significant reduction in error rates and performed better than conventional spectral subtraction, particularly at moderate SNRs.

1. INTRODUCTION

In speech and speaker recognition applications, training data is usually recorded in pristine and noise-free conditions, however test recordings are often subject to background noise and/or channel noise. The resulting mismatch in noise types and signal to noise ratios (SNR) between training and testing leads to severe performance loss.

This paper focuses on improving the performance of *short utterance* speaker recognition systems in the presence of additive noise.

2. CONVENTIONAL APPROACHES TO NOISE ROBUSTNESS

The conventional approaches to improve recognition accuracy in the in the presence of noise [4] include 1) improving SNR at the transducer, 2) preprocessing and 3) adapting model parameters to account for noise.

This research effort was supported by the Air Force Research Laboratory (AFRL), Rome, NY.

2.1. Spectral Subtraction

Spectral subtraction proposed by Boll [1] is a preprocessing technique for reducing the spectral effects of additive noise. When speech is corrupted with additive noise, the spectra of speech and noise are also added, that is,

$$y(n) = x(n) + \eta(n) \quad (1)$$

$$Y(\omega) = X(\omega) + N(\omega) \quad (2)$$

where, $x(n)$, $y(n)$ are time samples of the clean and noisy speech and $X(\omega), Y(\omega)$ are their corresponding spectra. $\eta(n)$ and $N(\omega)$ are the time and frequency domain representations of the corrupting noise source.

The spectral subtraction technique estimates the noise spectrum from periods of silence or non-speech activity and uses that estimate to subtract from the speech spectrum. This method assumes that the background noise environment remains locally stationary so that estimates made during non-speech regions can be used for subtraction during speech activity. The estimate used for spectral subtraction is the average *magnitude* spectrum of all the non speech frames. The magnitude spectrum of a frame i of the noise suppressed utterance, $|\hat{X}_i(\omega)|$ is given by,

$$|\hat{X}_i(\omega)| = |Y_i(\omega)| - |\overline{N}(\omega)| \quad (3)$$

where $|\overline{N}(\omega)|$ is the average noise magnitude spectrum estimated over M non-speech frames,

$$|\overline{N}(\omega)| = \frac{1}{M}\sum_{i=1}^{M}|N_i(\omega)| \quad (4)$$

These corrected magnitude spectra can then be used to reconstruct the noise reduced speech waveform. When $|\overline{N}(\omega)|$ is greater than $|Y_i(\omega)|$ at any frequency ω, $|\hat{X}_i(\omega)|$ becomes negative. Ad-hoc schemes such as spectral flooring (setting $|\hat{X}_i(\omega)|$ to some floor) have been employed to overcome this problem resulting in the characteristic musical effect in the reconstructed speech. Spectral subtraction provides a significant degree of speech enhancement thus enabling better speech/silence detection [3]. However, it does not always improve speaker recognition performance because it distorts the speech waveform thereby increasing mismatch between training and testing.

2.2. Model Adapatation

Experiments have shown that the performance degradation due to noise is more pronounced when the training and testing SNR levels are *mismatched* and is less severe when the SNRs are *matched*. Hence, one obvious approach to improve performance is to add noise to the less noisy signal so as to simulate matched conditions. This often involves re-training of models making it computationally expensive and impractical. Several approaches [4][2] have been proposed to adapt trained models (by re-estimating model parameters) to account for noise present in the test utterance. These methods require appreciable amounts of speech and noise data for meaningful parameter re-estimation. As most real world speaker recognition systems operate with small amounts of speech data (typically 3-4 seconds), these methods are difficult to implement.

3. PROPOSED METHOD: COHERENT SPECTRAL MODIFICATION

The proposed method, *Coherent Spectral Modification* (CSM), attempts to reduce the distortion due to noise by modifying the *complex* speech spectrum (hence the name *coherent*) using information from the spectral time trajectories.

3.1. Technique

Noise added to speech in the time domain manifests as an additive component in the frequency domain as well. If $N_i(\omega)$, $X_i(\omega)$ and $Y_i(\omega)$ represent the noise, clean speech and noisy speech spectra for the i^{th} frame of speech,

$$Y_i(\omega) = X_i(\omega) + N_i(\omega) \tag{5}$$

Expressing this more precisely in terms of real and imaginary components,

$$\begin{aligned} Y_i^{Re}(\omega) + j\, Y_i^{Im}(\omega) \quad &= \quad \left[X_i^{Re}(\omega) + j\, X_i^{Im}(\omega)\right] + \\ &\left[N_i^{Re}(\omega) + j\, N_i^{Im}(\omega)\right] \end{aligned} \tag{6}$$

The time trajectories of the real (X^{Re}, Y^{Re}) and imaginary (X^{Im}, Y^{Im}) spectral components (at 3000 Hz) of the clean and noisy versions of the same utterance (at 10 dB SNR) are shown in Fig. 1. This was obtained by framing the speech, computing the spectrum for each frame and plotting the spectral variation across frames at 3000 Hz.

From Fig. 1, it can be observed that the addition of noise causes a significant increase in the clean speech spectral amplitudes. This is true for both the real and imaginary components and for the most part, noise causes the positive spectral amplitudes to go more positive and the negative values to go more

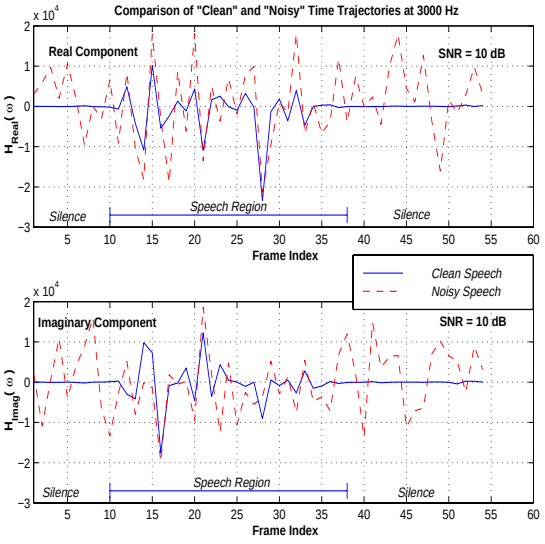

Figure 1: Clean and Noisy time trajectories for the real and imaginary components at 3000 Hz.

negative. This can be understood more clearly by examining the distributions of the time trajectories of clean and noisy speech shown in Fig. 2.

The addition of two random independent random processes (speech and noise in this case) causes their probability density functions to be convolved. This convolution causes the distribution of the resulting random process to have a larger variance. This increase in variance manifests as an increase in spectral amplitude of the time trajectories observed above. The CSM technique aims to reduce this noise induced increase in spectral amplitudes in the speech regions. This will effectively reduce the variance of the noisy spectral distribution.

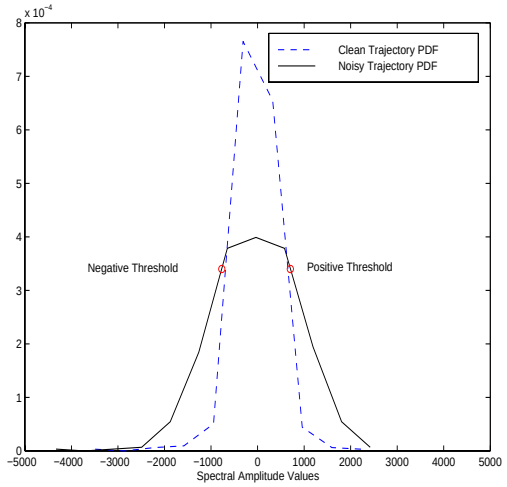

Figure 2: Normalized Distributions of the real part of the Clean and Noisy time trajectories.

3.1.1. Modification Process

From Fig. 1, it can be seen that the amplitude variation in the silence regions of the noisy speech trajectory is entirely due to noise. Therefore, the CSM algorithm uses information present in the "noise-only" or non-speech regions to estimate the extent of modification required. Independent correction factors are estimated for the positive and negative components of the real and imaginary parts. This results in four correction factors which are applied to their corresponding spectral trajectory amplitudes. The correction procedure (for the real part[1]) is given by the following equation:

$$\hat{X}_i^{Re}(\omega) = \begin{cases} Y_i^{Re}(\omega) - \delta_+^{Re}(\omega), & \text{if } Y_i^{Re}(\omega) > \delta_+^{Re}(\omega) \\ Y_i^{Re}(\omega) + \delta_-^{Re}(\omega), & \text{if } Y_i^{Re}(\omega) < \delta_-^{Re}(\omega) \\ Y_i^{Re}(\omega), & \text{otherwise} \end{cases} \tag{7}$$

where $\delta_+^{Re}(\omega)$ and $\delta_-^{Re}(\omega)$ are correction factors for the real positive and real negative spectral values.

In order to completely restore the clean spectrum from its noisy version, distinct correction factors would have to be applied at each point, but it is not possible to estimate these precise corrections owing to the random nature of noise. In the current implementation of the CSM technique, these multiple correction factors are approximated by an average obtained from the non-speech regions. More specifically, $\delta_+^{Re}(\omega)$ is the average of all positive spectral amplitude values in the silence regions of the time trajectory at frequency ω. The other three correction factors are found in a similar manner.

3.1.2. Correction Criterion

The four correction factors estimated using the procedure given above are also used as thresholds that determine the regions in the distribution where correction is to be performed. These thresholds are shown in Fig. 2. The magnitude values of the points that lie in the non-overlapping regions of the clean and noisy distributions definitely needs to be *reduced* in order to correct for the spread in the distribution. This reduction is achieved by using the correction factors described previously. However, in the region between the positive and negative thresholds, it is difficult to determine the extent and direction of correction required for restoration. As we have only one fixed correction factor, we leave the points in this region untouched.

After the spectral modifications, the speech signal is reconstructed from the spectrum and used for further processing.

[1]The imaginary component is corrected similarly

3.2. Comparison with Spectral Subtraction

This approach although similar in principle to spectral subtraction differs significantly from it as follows:

- Whereas spectral subtraction (SS) subtracts the average noise *magnitude* spectrum, CSM operates coherently on the real and imaginary components. Hence the problem of negative magnitudes and the corresponding musical effects prevalent in SS do not arise.
- Independent correction factors are estimated for the positive and negative values of the real and imaginary spectra providing four degrees of freedom for correction. On the other hand, SS uses a single correction factor for each frequency.

4. REFINEMENT

In speaker recognition systems, reduction in mismatch is more important than speech enhancement. When training utterances are available, a further reduction in mismatch between the training and testing is possible in addition to the noise reduction described above. Eq. 5 can be rewritten as,

$$|Y_i(\omega)| = \left| \frac{X_i(\omega) + N_i(\omega)}{X_i(\omega)} \right| |X_i(\omega)| \tag{8}$$

$$= |\mathcal{R}_i(\omega)\, X_i(\omega)| \tag{9}$$

where, $\mathcal{R}_i(\omega)$ can be considered as a filter that depends on the noise as well as the clean speech signal for that frame. If we could construct this *frame by frame* adaptive filter, we could use it as an inverse filter to recover the clean signal frames. But in order to do so, we need the clean signal itself. In a *text-dependent* speaker verification scenario that has been trained in a clean environment, the training signal spectrum could be used in place of the clean test signal spectrum, i.e. $X_i(\omega)$ can be replaced with $T_i(\omega)$. However, we still need precise time alignment between the training and testing utterances to construct the adaptive filter. Time alignment is very difficult particularly in the case of low SNRs. In order to circumvent this problem we construct a time-invariant filter using the mean spectrum of the training and testing utterances. That is,

$$|\mathcal{R}_i(\omega)| \approx |\overline{\mathcal{R}(\omega)}| = |\overline{Y(\omega)}|/|\overline{T(\omega)}| \tag{10}$$

where $\overline{T(\omega)}$ is the mean training spectrum. This ratio, computed at each frequency is often very noisy. A level of smoothing is performed with low order moving average and median filters to obtain a less noisy estimate. The inverse of this smoothed spectrum defines the filter that can be applied on the noisy test to minimize mismatch. As this technique uses the average training and testing spectra, it captures

	Equal Error Rates							
	White Noise					Pink Noise		
SNR →	**5 db**	**10 dB**	**15 dB**	**20dB**	**25 dB**	**5 dB**	**15 dB**	**25 dB**
Noise on Test	50.0 %	44.28 %	35.17 %	24.19 %	15.27 %	40.90 %	20.10 %	6.96 %
Spectral Subtraction	41.25 %	40.47 %	34.44 %	30.43 %	25.55 %	30.91 %	17.96 %	14.05 %
CSM only	40.23 %	34.01 %	26.33 %	15.90 %	11.06 %	30.39 %	15.82 %	8.99 %
CSM and Refinement	*39.45 %*	*29.30 %*	*21.32 %*	*14.71 %*	*9.49 %*	*25.18 %*	*12.56 %*	*8.24 %*

Table 1: Comparison of different noise reduction schemes for speaker verification at different test SNRs

the average spectral difference in environments. In cases where the training speech is corrupted (either by channel or noise or both), the ratio would still capture the gross spectral difference between the training and testing environments and is therefore still effective.

5. EXPERIMENTS

Speaker verification experiments were performed to evaluate the performance of the Coherent Spectral Modification technique (before and after refinement) and compare it with Spectral Subtraction.

The experiments were performed on a 51 speaker text-dependent database[2]. Each speaker has 12 repetitions of the phrase "Rome Laboratory" (about 1 second in length). All these utterances have been recorded through the same local telephone network at 8k Hz. The recordings are relatively "clean" with the average SNR being above 30dB. 4 utterances were used for training and 8 for testing. The experiments were performed by maintaining a clean training environment and corrupting the test environment different levels of additive white Gaussian noise and pink noise to simulate various test SNRs.

Speaker verification was performed using 12^{th} order Linear Prediction (LP) derived cepstrum. Vector Quantization (VQ) with 64 codebooks per speaker was used for classification. The baseline system (clean training, clean testing) yielded an Equal Error Rate (EER) of 2.36%. The results are shown in Table 1. The following observations can be made from the results:

- The performance drops drastically when noise levels are mismatched. Even reasonable SNRs like 25 dB with white noise yield equal error rates as high as 15%.
- The spectral subtraction noise suppression technique improves performance only at signal to noise ratios of 15 dB or lower (for both white and pink noise).
- Coherent Spectral Modification (CSM) almost always improves performance quite substantially and always performs better than Spectral Subtraction, particularly at moderate SNRs. The refinement when applied after CSM improves performance even further.

[2] Obtained from U.S. Air Force

6. CONCLUSIONS

In this paper, we presented a new speech-in speech-out technique that reduces the noise induced mismatch between training and testing data. Although similar in principle to spectral subtraction, this method employs a non-linear subtraction scheme that operates independently on the real and imaginary parts of the spectrum using different correction factors for their positive and negative components. We also developed a refinement process that is applicable when training utterances are available. Speaker verification results showed that the combined technique improved performance substantially, regardless of the noise type, and outperformed spectral subtraction, especially at moderate SNRs.

7. REFERENCES

[1] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics Speech and Signal Processing, Vol ASSP-27*, 1979.

[2] M. J. F. Gales and S. J. Young. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, September 1996.

[3] Vijay Raman and Vidhya Ramanujam. Incorporation of noise preprocessing into an entrenched speech recognition system. *Proceedings, ECSA Workshop*, May 1996.

[4] R. C. Rose, E .M. Hofstetter, and D. A. Reynolds. Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Speech and Audio Processing*, April 1994.