International Journal of Computer Science and Engineering (IJCSE)
ISSN 2278-9960
Vol. 2, Issue 4, Sep 2013, 157-162

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

TEXT INDEPENDENT SPEAKER VERIFICATION IN REAL TIME ENVIRONMENT USING MULTIBAND SPECTRAL SUBTRACTION AND GMM ON MOBILE PHONES

PRANAB DAS & UTPAL BHATTACHARJEE

Rajiv Gandhi University, Doimukh, Arunachal Pradesh, India

ABSTRACT

In this paper we study the performance of speaker verification system by applying spectral subtraction to multiband speech in real environment. In real world environment noise from different sources could exist and which may interfere with the speech signal at different frequencies. Because of the colored nature of noise which does not spread uniformly over the spectrum of speech i.e. some of the frequencies may be most affected while some frequencies may be least affected, a multiband filter bank approach is proposed. In this approach a filter bank is designed which divides the speech signal into a number of frequency bands. Spectral subtraction is then applied to each of the bands and the result of all the subtraction are combined at the end. Results have shown quite a significant improvement in performance when spectral subtraction is applied to multi band than applied to the entire speech signal.

KEYWORDS: Speaker Verification, Spectral Subtraction, Filter Bank, Gaussian Mixture Model

INTRODUCTION

In the early ages of speaker recognition, researchers have faced the problem of enhancing speech degraded by additive noise in real world environment. Noise reduction is useful in many applications such as banking, telecommunication and automatic speaker recognition where efficient noise reduction techniques are required.

In the literature, various approaches that work at signal level, feature level and model level have been proposed to improve the performance of an Automatic Speaker Recognition system with respect to noise, such as Wiener filtering [2], spectral subtraction [1], RASTA [3], parallel model compensation (PMC) [4].

One of the most popular methods of reducing the effect of additive noise as proposed by [1] is Spectral Subtraction. Spectral subtraction is a method for restoration of the power spectrum or the magnitude spectrum of a signal observed in additive noise, through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. The noise spectrum is usually estimated, and updated, from the periods when the signal is absent and only the noise is present. The assumption is that the noise is a stationary or a slowly varying process, and that the noise spectrum does not change significantly in between the update periods.

As the real-world noise is not flat, the noise signal does not affect the speech signal uniformly over the entire spectrum. Some frequencies are most affected while some of the frequencies may be least affected. For instance, the low frequencies, components in a babble environment where most of the speech energy resides, are affected more than the high frequency component. Therefore it becomes imperative to apply spectral subtraction to each of the frequency bands. The approach involves design of a filter bank using low pass, band pass and high pass filter and apply the spectral subtraction method to reduce the above-mentioned distortions at different frequencies to a large extent while maintaining a high level of speech quality.

The structure of the paper is as follows: section I introduce the speaker verification system in noisy environment;

section II describes the proposed method. Section III describes noise reduction technique spectral subtraction. Section IV describes modeling using GMM. While section V gives describes the speaker recognition database. Section VI shows the results of the experiments and section VII gives the conclusion of the paper.

DESIGN OF FILTER BANK

As for a given input speech in real world environment, all the frequencies are not affected uniformly; the proposed method takes this into account and designs a filter bank as given in figure 1.

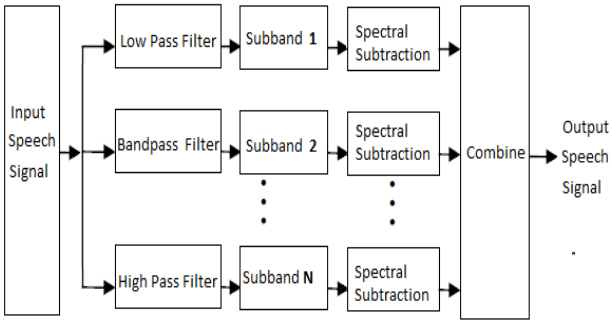

Figure 1: Filter Bank Design

Three types of filter are considered namely lowpass, highpass, bandpass filters. The lowpass filter passes signals with a frequency range of 0Hz to the corner frequency, f_c and blocks all signals operating at frequencies above f_c. The highpass filter blocks signals with a frequency range of 0Hz to the corner frequency, f_c and passes all signals operating at frequencies above f_c. And the bandpass filter blocks signals with a frequency range of 0Hz to the corner frequency, f_1 and all signals operating at frequencies above f_2. The signals operating at frequencies between f_1 and f_2 are passed.

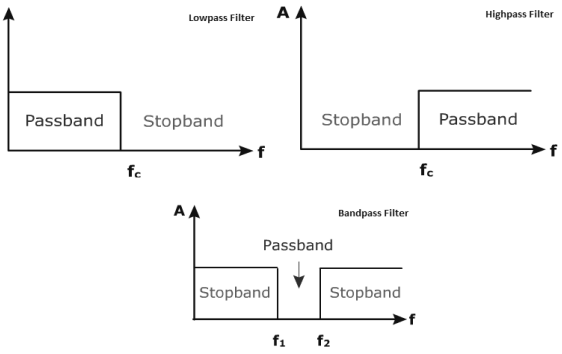

Figure 2: Lowpass, Highpass and Bandpass Filter

SPECTRAL SUBTRACTION

One of the most widely used methods of reducing noise from a speech signal is Spectral Subtraction.

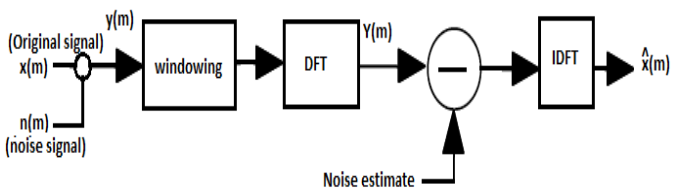

Figure 3: Basic Structure of Spectral Subtraction Using DFT

Assuming that the noise is a stationary random process .The noisy signal model of speech corrupted by background noise is expressed as:

$$y(m) = x(m) + d(m) \tag{1}$$

Where x(m), d(m) and y(m) are the signal, the additive noise signal and the noisy signal respectively, and m is the discrete time index. Windowing the signal of equation (1) using a Hamming window results in:

$$y_w(m) = x_w(m) + d_w(m) \tag{2}$$

Applying discrete Fourier transform (DFT) to both the sides of equation (2) gives:

$$Y_w(e^{j\omega}) = X_w(e^{j\omega}) + D_w(e^{j\omega}) \tag{3}$$

Dropping the subscript w for the windowed signal, the equation describing spectral subtraction can be expressed as:

$$\left|\hat{X}(e^{j\omega})\right|^b = \left|Y(e^{j\omega})\right|^b - \alpha\overline{|D(e^{j\omega})|^b} \tag{4}$$

Where $\left|\hat{X}(e^{j\omega})\right|^b$ an estimate of the original signal spectrum $|X(e^{j\omega})|^b$ and $\overline{|D(e^{j\omega})|^b}$ is the time-averaged noise spectra. The parameter α controls the amount of noise subtracted from the noisy speech signal, for full noise subtraction α equals one. For magnitude spectral subtraction, the exponent b=1, and for power spectral subtraction, b=2. For equation (4) assuming α to be unity and b=2, the power spectrum is given by

$$\left|\hat{X}(e^{j\omega})\right|^2 = \left|Y(e^{j\omega})\right|^2 - \overline{|D(e^{j\omega})|^2} \tag{5}$$

Taking expected value of both sides

$$E\left[\left|\hat{X}(e^{j\omega})\right|^2\right] = E\left[\left|X(e^{j\omega})\right|^2\right] \tag{6}$$

Again, for equation (4) assuming α to be unity and b=1, the magnitude spectrum is given by

$$\left|\hat{X}(e^{j\omega})\right| = \left|Y(e^{j\omega})\right| - \overline{|D(e^{j\omega})|} \tag{7}$$

Taking expectation of equation (7) for both sides we get

$$E\left[\left|\hat{X}(e^{j\omega})\right|\right] \approx E\left[\left|X(e^{j\omega})\right|\right] \tag{8}$$

GAUSSIAN MIXTURE MODELING

A Gaussian Mixture Model is a parametric probability density function represented as a weighted sum of M Gaussian component densities given by the equation.

$$p(X|\lambda) = \sum_{k=1}^{M} w_k \; g(X|\mu_k, \Sigma_k) \tag{9}$$

where X is a D-dimensional continuous-valued data vector, w_k , k = 1, . . . ,M, are the mixture weights, and, $g(X|\mu_k, \Sigma_k)$, k =1, . . . ,M, are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form,

$$g\left(X\middle|\mu_k, \Sigma_k\right) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(x - \mu_k\right)'\Sigma_k^{-1}\left(X - \mu_k\right)\right\}, \tag{10}$$

With mean vector μ_k and covariance matrix Σ_k .

GMM parameters, λ, are estimated by the most popular and well-established method is maximum likelihood estimation. ML parameter estimates are obtained iteratively using a special case of the expectation-maximization (EM)

algorithm [5]. The basic idea of the EM algorithm is to begin with an initial model λ, and to estimate a new model,$\bar{\lambda}$, such that p(X|, $\bar{\lambda}$) ≥ p(X| λ).

On each EM iteration, the following re-estimation formulas are used which guarantee a monotonic increase in the model's likelihood value,

$$\overline{w_k} = \frac{1}{T}\sum_{t=1}^{T} \Pr(k|X_t, \lambda) \quad (11)$$

$$\overline{\mu_k} = \frac{\sum_{t=1}^{T} Pr(k|X_t, \lambda)}{\sum_{t=1}^{T} Pr(k|X_t, \lambda)} \quad (12)$$

$$\overline{\sigma_k}^2 = \frac{\sum_{t=1}^{T} Pr(k|X_t, \lambda) x_t^2}{\sum_{t=1}^{T} Pr(k|X_t, \lambda)} - \overline{\mu_k}^2 \quad (13)$$

where $\overline{w_k}$, $\overline{\mu_k}$ and $\overline{\sigma_k}^2$ are mixture wights, means and diagonal covariance[6].

SPEAKER RECOGNITION DATABASE

To carry out the experiments, a speaker verification database was developed and all the testing and evaluation of the speaker recognition system was done with respect to that database. Recording was done for 22 male and 22 female speakers.

There were two enrolment sessions and two verification sessions for the same subject and a gap of 20 days between two consecutive sessions. Each recording for the training phase is of 3 minutes duration while it was of 30 seconds duration for the testing phase. Data were recorded in parallel across two recording devices, which are listed in table 1.

Table 1: Device Type and Recording Specifications

Device Sl. No.	Device Type	Sampling Rate	File Format
1	Mobile 1	16 kHz	wav
2	Mobile 2	16 kHz	wav

The speakers were recorded for reading style of conversation. The speech data collection was done in real time roadside environment. The speech data was chosen from the age group 16-25 years. During recording, the subject was asked to read a paragraph of phonetically rich sentences of duration 4 minutes in English language for twice and the second reading was considered for recording.

EXPERIMENTS AND RESULTS

A speaker verification system was developed using Gaussian Mixture Model (GMM) based modeling approach. In the first set of experiment training and test data were taken from mobile 1 recording. At the acoustic noise removal stage each of the noise corrupted speech signals is cleaned by using spectral subtraction in both the training and testing phase.

Also a second set of experiment were done by applying the proposed system i.e. filter bank and spectral subtraction, at the noise removal stage for the training and testing phase of mobile 1 recording. Similarly both the experiments were conducted for mobile 2 training and test data first using spectral subtraction and then by using filter bank and spectral subtraction at the acoustic noise removal stage.

Detection Error Tradeoff curve were plotted for both the experiments. Figure 4 and Table 2 shows the DET curve and Equal error rate (EER) respectively for both mobile 1 and mobile 2 recordings.

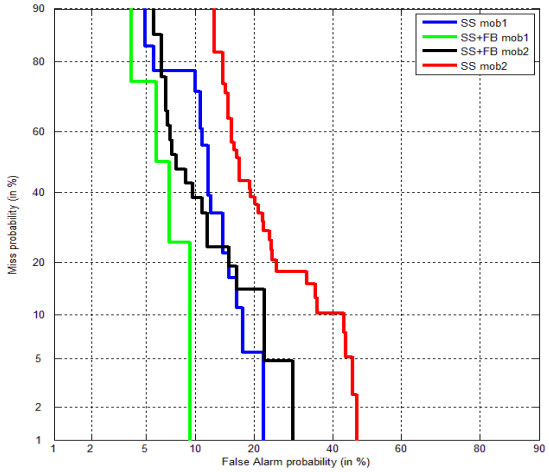

Figure 4: DET Curve Showing the Results when Training and Testing Data were Taken from Same Device

Table 2: Equal Error Rate for Train and Test Data from Same Device

	Train -Mobile1 Test -Mobile1	**Train –Mobile2 Test –Mobile2**
Spectral subtraction(SS)	EER=16.67	EER=20.51
Filter bank(FB) and spectral subtraction(SS)	EER=9.09	EER=14.28

Two more experiments were conducted by first taking mobile 1 data for training and mobile 2 data for testing and applying spectral subtraction at the noise removal stage and plotting the DET curve. Similarly the same set of data was again taken and the proposed system is applied at the noise removal stage and DET curve is plotted for the verification system.

Next the training and test data were taken from mobile 2 and mobile 1 respectively and the above mentioned experiments were conducted and DET curves were plotted. Figure 5 and Table 3 shows the DET curve and Equal error rate (EER) respectively for both the mobile recordings.

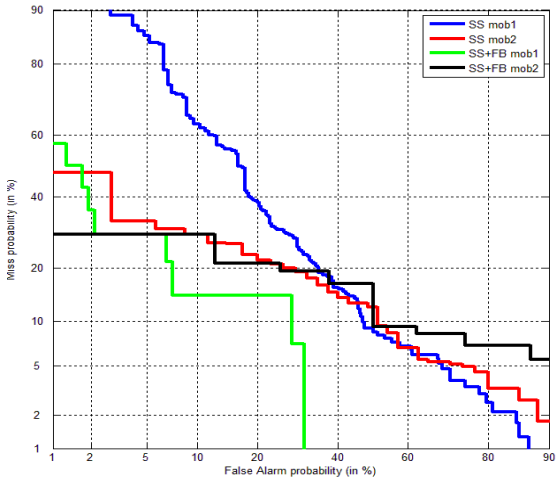

Figure 5: DET Curve Showing the Results when Training and Testing Data were from Different Device

Table 3: Equal Error Rate for Train and Test Data from Different Device

	Train -Mobile1 Test –Mobile2	**Train –Mobile2 Test –Mobile1**
Spectral subtraction(SS)	EER=27.89	EER=22.27
Filter bank(FB) and spectral subtraction(SS)	EER=14.28	EER=20.73

CONCLUSIONS

We have proposed a multiband spectral subtraction method for speaker verification system. The above mentioned method has shown improvement in equal error rate when the proposed system is applied to the Speaker verification system at the acoustic noise removal stage then when spectral subtraction is applied to the entire speech spectrum in the noise removal stage of the system. Also it is observed that when the training and testing data are from same or different devices the proposed method tends to outperform the existing method.

REFERENCES

1. S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process. vol.27, pp. 113-120, Apr. 1979.
2. Vaseghi, S.V., Milner, and B.P.: Noise compensation methods for hidden Markov model speech recognition in adverse environments. IEEE Trans. Speech Audio Process. 5 (1), 11–21 (1997).
3. Hermansky, H., Morgan, N.: RASTA of processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589 (1994).
4. Gales, M.J.F., Young, S.J.: Robust speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4 (5), 352–359 (1996).
5. Dempster, A., Laird, N., Rubin, D.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society 39(1) (1977) 1–38
6. Gaussian Mixture Models ,Douglas Reynolds MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA, dar@ll.mit.edu