

# Statistical And Model Based Approach To Unvoiced Speech Detection

Krithika Giridharan, Brett Y. Smolenski and Robert E. Yantorno  
Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, Pa 19122-6077, USA  
Krithika.giridharan@temple.edu, bsmolens@temple.edu, robert.yantorno@temple.edu  
[http://www.temple.edu/speech\\_lab](http://www.temple.edu/speech_lab)

**Abstract:** The detection of unvoiced speech in the presence of additive background noise is complicated by the fact that unvoiced speech is very similar to white noise. The mechanism of production of unvoiced speech is known to be due to turbulent airflow in the constrictions of the vocal tract. Three approaches of detecting unvoiced speech from additive background noise have been developed. Two of them which are very effective in the presence of additive white noise, are model based and autocorrelation based respectively. The probability of correct detections, on an average being 74%. A statistical approach is however developed that works both for additive white and pink noise. Further research on this statistical measure is being attempted to use it in a simple threshold based detector of unvoiced speech.

## 1. Background

Unvoiced speech detection in the presence of additive background noise is important in speech segmentation applications, where the incoming speech is classified either as voiced, unvoiced or background noise. The type of classification considered in this research is tree type, where the voiced sections of speech are detected and removed and we are left with the task of classifying unvoiced speech and background noise.

Many features have been used in the past in discriminating between voiced, unvoiced and silence. The chief among them are energy and zero crossings, autocorrelation coefficient at unit sample delay, first predictor co-efficient and energy of prediction error which have been used in a pattern recognition type classification [1]. The disadvantage with this type of classification is that it is strongly dependent on changing recording conditions, especially in the classification of silence in the presence additive background noise.

Both statistical and model based approaches have been looked at, to see if new features can be derived, that are robust to changing background noise conditions.

## 2. Introduction

The goal of this research has been to identify the unvoiced portions in speech after all the voiced portions are removed. The task is complicated by the fact that unvoiced speech resembles white noise in many respects. Three features that can be employed in a simple threshold based classifier are presented.

The measure based on the autocorrelation of the noisy unvoiced speech is the Autocorrelation Distance Measure (ADM). The Normalized Residual Energy Measure (NREM) is however model based. The ADM is computed by observing the autocorrelation for the first ten lags of a frame of input speech; whereas the NREM is the normalized residual energy obtained by modeling unvoiced speech with added white noise as an ARMA (5, 5) process. Both these measures work well when used in the classification of unvoiced speech with added white noise but perform poorly in the case of added pink noise.

A statistical measure called the Quantile Slope Measure is presented that has been used to make a classification in background white as well as pink noise conditions. This method is based on observing the first and third quantiles of a frame of input speech. The ratio of the difference between the first and third quantiles of the input frame to that of the first and third standard normal quantiles is found. This measure is based on the fact that although unvoiced speech is similar to white noise, its distribution deviates from being Gaussian when considered over frames that are long enough. The same applies to background pink noise conditions, as pink noise which has  $1/f$  dependence is Gaussian distributed.

The rest of this paper is organized as follows. Section 2 deals in detail about the three measures. Section 3 describes the experiments and results and section 4 has the conclusions and future research scope.

## 3. Features

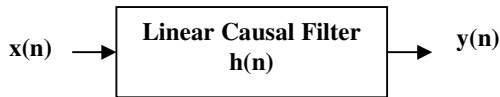
### 3.1 ARMA modeling and the NREM

Extraction of information in the short time speech spectra by approximation of the speech production process as an autoregressive process has been done in the past and has achieved good performance in digital transmission, synthesis and recognition of speech.[3-6]. However for a more accurate description of speech, especially nasalized sounds the ARMA modeling is proposed [7-9].

The term 'model' is used to describe any process or system used to generate data of interest. The data of interest, in our case, is the unvoiced speech corrupted with white noise. It is known that unvoiced speech is a quasi-stationary autoregressive moving average process with five poles and three zeroes ARMA (5, 3) [6]. This is because unvoiced speech is the result of turbulent airflow at a constriction in the vocal tract, which produces a corresponding number of

resonances (poles) and anti-resonances (zeros). The addition of noise to the process introduces zeroes so it becomes an ARMA (5, 5) process.

To generate an ARMA process  $y(n)$  we use a discrete time linear filter with a transfer function that contains both poles and zeroes. Accordingly, with a white noise process  $x(n)$  as the input, the filter would generate an ARMA process at the output which is as shown below.[10]



**Figure 3.1:** ARMA process  $y(n)$  produced by white noise excitation  $x(n)$  to a linear causal filter

The order of the ARMA process is  $(M, K)$  where  $M$  denotes the order of the denominator of the transfer function of the discrete-time linear filter (or the number of poles) and  $K$  denotes the order of the numerator of the transfer function (or the number of zeroes). The **Figure 3.1** above describes an ARMA generator.

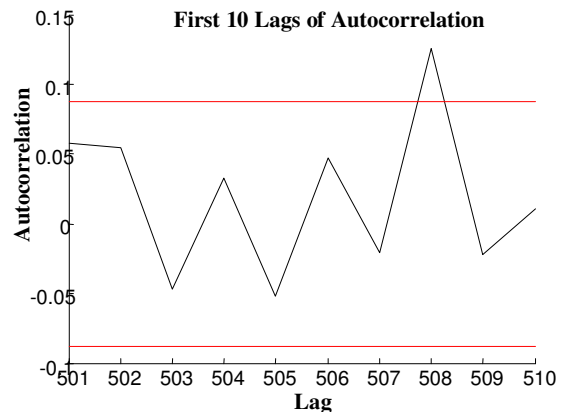
The transfer function of an ARMA generator contains both poles and zeroes. Similarly an ARMA analyzer which is nothing but the inverse of the ARMA generator is characterized by a transfer function containing both poles and zeroes. The function of the ARMA analyzer is to produce white noise  $x(n)$  given  $y(n)$  as the input. The normalized energy at the output of this inverse filter forms the basis of the Normalized Residual Energy Measure (NREM).[10]

The Residual Energy Measure is defined as the ratio of the variance of the output of the inverse filter (or the ARMA analyzer) to that of the variance of the input speech signal. An inverse filter is a system that accepts the ARMA process as the input and generates white noise at the output. This normalized variance is high for portions of background noise (nearing unity) and is considerably less for unvoiced portions enabling us to use this measure to differentiate between unvoiced and background.

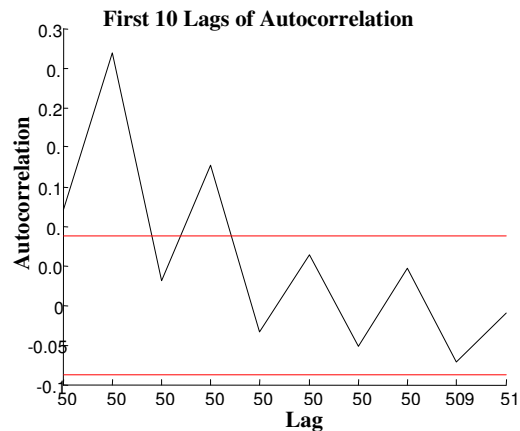
### 3.2 Autocorrelation Distance Measure (ADM)

The ADM sums the distances of those values of the normalized autocorrelation that is outside a 95% confidence interval. The autocorrelation distance measure is expected to have a greater value for unvoiced speech than for background noise. This is because, background noise is uncorrelated and hence the autocorrelation is expected to be zero for all lags other than lag zero. However, unvoiced speech is expected to be more correlated, and thus have large values within the first ten samples of the autocorrelation sequence. Hence, for the autocorrelation

distance measure, only the first ten samples near the maximum peak for zero lag are considered, since, the later samples have less correlation. Peaks of the autocorrelation sequence do sometimes occur for background noise portions at lags other than zero. So, a 95% confidence interval is assumed. By using a 95% confidence interval, we assume that any peaks greater than 1.96 times the standard deviation of the signal is produced by correlation in the unvoiced speech. One might have chosen a higher confidence interval, but 95% confidence appears to be optimal for our purposes.



**Figure 3.2.1:** Autocorrelation of white noise at 5 dB SNR



**Figure 3.2.2:** Autocorrelation of noisy unvoiced speech at 5 dB SNR

As can be seen from the **Figures 3.2.1** and **3.2.2** above, white noise has less correlation as compared with unvoiced speech. This is evident from the number of peaks that are greater than the 95% confidence interval (within the first ten lags of the autocorrelation sequence), which are greater for unvoiced speech than for white noise.

Both the above measures are expected to work poorly for high values of SNR (when corrupted speech tends towards clean speech), since our method assumes substantial background noise is present. Also, they will work poorly for low SNR values, because the unvoiced portions plus background noise will appear much more noise-like, and therefore, produce results very similar to background noise.

### 3.3 Quantile Slope Measure

Unvoiced speech looks like white noise, but statistically is different from white noise. White noise has a Gaussian or normal distribution; whereas unvoiced speech or noise corrupted unvoiced speech will have a distribution different from white noise. It should be noted that all speech is Laplacian distributed [11].

The quantile quantile plot (Q-Q plot) is a graphical tool used to verify the Gaussianity of a signal [12]. The QQ plot is the plot of the standard normal quantiles versus the quantiles of the dataset under consideration and is a straight line with slope unity for a Gaussian signal. Any deviations from the straight line is a sign of non-Gaussianity.[13]

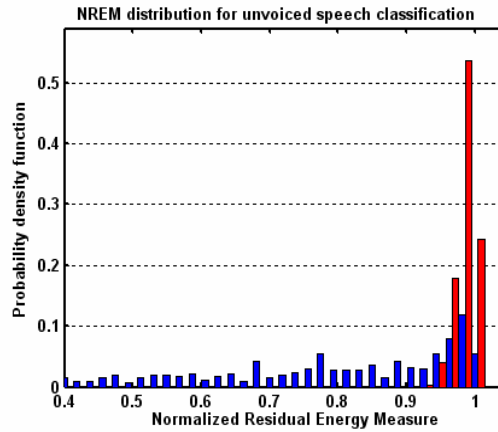
The measure presented here is statistical and based on the first and third quantiles of the unvoiced and background noise data set. The idea is to use the slope of the Quantile Quantile (Q-Q) plot as the measure to differentiate between unvoiced and background noise. The slope of this plot is different for unvoiced speech as well as for unvoiced speech with added noise as compared with just noise.

Because there is a statistical difference between unvoiced speech and noise the slope value of the QQplot is expected to be different for unvoiced frames as compared to noise only frames. This is because background noise either white or pink is Gaussian distributed, whereas, the unvoiced portions are only approximately Gaussian distributed and hence expected to have outliers from the otherwise straight line obtained for background noise frames.

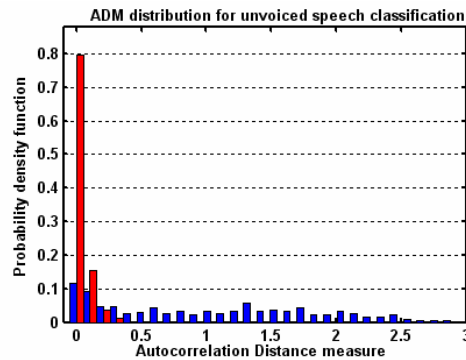
It should work for clean speech, because, background noise portions are now replaced with silence (noise at very high SNR), and hence the slope of the QQ plot should be different for unvoiced and silence frames.

## 4. Experiments and Results

White noise was added at 10 dB SNR to the speech file of a female speaker. Unvoiced and silence portions were extracted and the histograms for the background noise and unvoiced speech were plotted (See **Figures 4.1** and **4.2** below). The histogram shows that there is separation between the two classes. Based on information of **Figures 4.1** and **4.2** below, thresholds of 0.95 for the Normalized Residual Energy Measure and 0.2 for Autocorrelation Distance Measure were chosen.



**Figure 4.1: Histogram of the Normalized Residual Energy Measure for background noise (red) and unvoiced speech (blue) for a single speech file of a female speaker with added white noise at 10 dB SNR.**



**Figure 4.2: Histogram for the Autocorrelation Distance Measure for background noise (red) and unvoiced speech (blue) for a single speech file of a female speaker with added white noise 10 dB SNR.**

The Air Force speech data, which had added White noise at varying SNR values, from 5dB to 25dB, was used. The speech files were sampled at a rate of 16000 samples per second. Both male, female and Spanish speech files of durations on an average 90 seconds were used and the measure values were computed for each frame having a length of 512, which is equal to 32msec speech frame. The frame length of 512 was chosen because, experimentation by changing the frame length values gave good results for the hits and false alarms for a frame length of 512. The hits and false alarms were based on ground truth air force data. The thresholds chosen from the information in the histograms were adjusted so that an optimum value of the hits and false alarms was obtained for the entire range of SNR values 5dB to 25dB. These values are given in **Figures 4.3** and **4.4**. The thresholds that produced the optimal results were 0.95 for the Normalized Residual Energy Measure and 0.2 for the Auto-correlation Distance Measure. From the figures below it can be seen that the

measures work well for added white noise at SNR values in the range 5 dB to 25 dB.

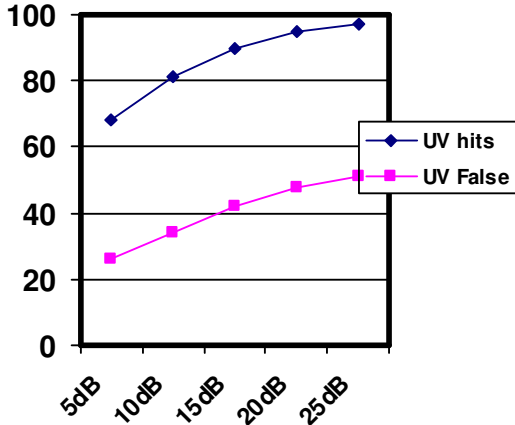


Figure 4.3: Results for Autocorrelation Distance Measure, Threshold = 0.2

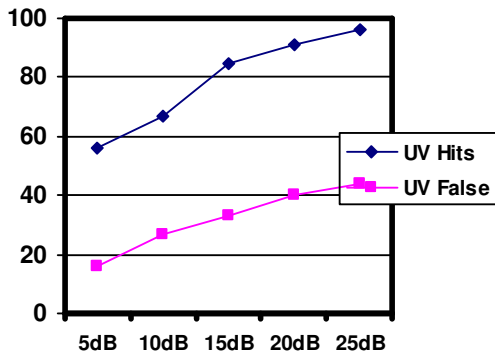


Figure 4.4: Results for Normalized Residual Energy Measure, Threshold = 0.95

However the results for added pink noise were not found to be good and hence a statistical approach that would work both for added pink noise and white noise conditions was attempted.

As mentioned before the slope of the Q-Q plot for a frame of background noise and a frame of unvoiced speech with added background noise is noted to be different from the Figures 4.5 and 4.6 shown.

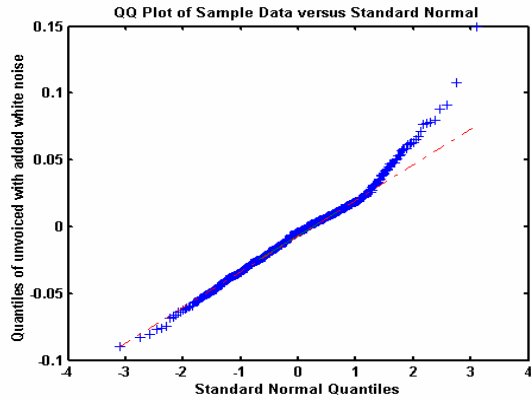


Figure 4.5: QQplots for a data length of 512 for ‘unvoiced speech+white noise’ against standard normal distribution (bandwidth 8 kHz).

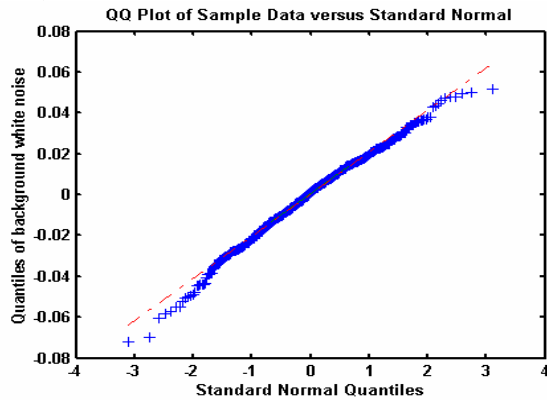


Figure 4.6: QQplots for a data length of 512 for ‘white noise’ against standard normal distribution (bandwidth 8 kHz).

It is observed that the QQplot of the background noise data lies along the reference forty five degree line (Figure 4.6), whereas the plot for noisy unvoiced frame deviates from the reference line (Figure 4.5), which was expected. The same is true for unvoiced speech corrupted with added Pink noise as shown in Figures 4.7 and 4.8 below shown for a data length of 512.

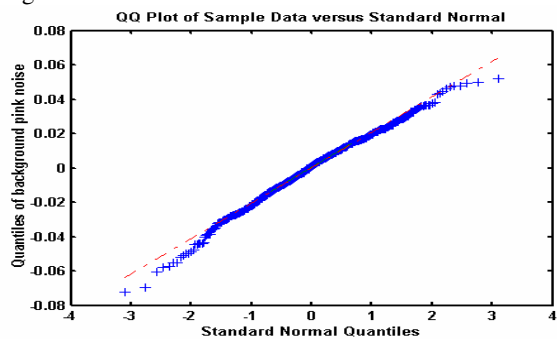


Figure 4.7: QQplots for a data length of 512 for ‘pink noise’ against standard normal distribution (bandwidth 8 kHz).

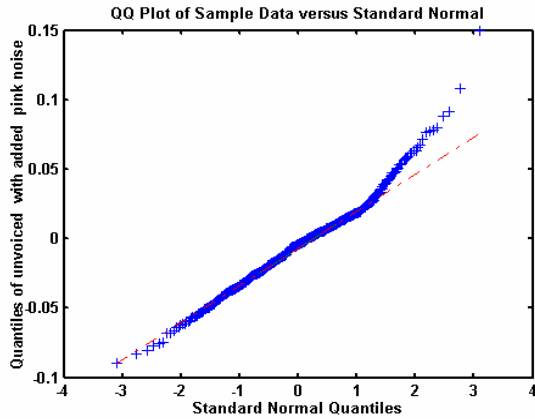


Figure 4.8: QQplots for a data length of 512 for 'unvoiced speech+pinknoise' against standard normal distribution (bandwidth 8 kHz).

Hence, the slope of the plot measured as the ratio of the difference between first and third quantiles of the speech frame to the difference between first and third standard normal quantiles can be used as a measure to differentiate between unvoiced and Background noise. For all the figures shown above a female file with 15 dB added background noise was used. However the experiments were performed over both male and female files at different SNR's.

Figures 4.9 and 4.10 show that the histograms of the measure for the case added white noise and pink noise respectively.

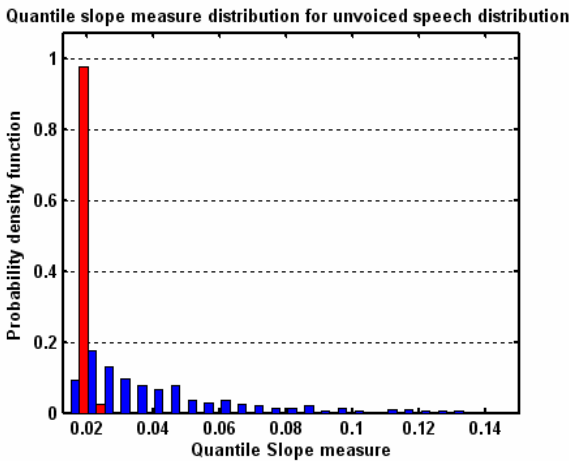


Figure 4.9: Histogram for the Quantile Slope Measure for background noise (red) and unvoiced speech (blue) for a single speech file of a female speaker with added white noise 15 dB SNR.

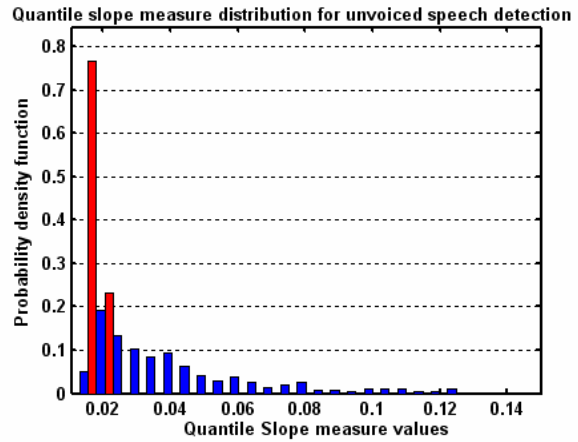


Figure 4.10: Histogram for the Quantile Slope Measure for background noise (red) and unvoiced speech (blue) for a single speech file of a female speaker with added pink noise 15 dB SNR.

The Receiver Operating Curves (ROC) were plotted as shown in Figures 4.11 and 4.12 by changing the threshold values for the measure and noting the percentage of hits and false alarms in each case. The male and female speech files were used at 10 and 20 dB added background noise.

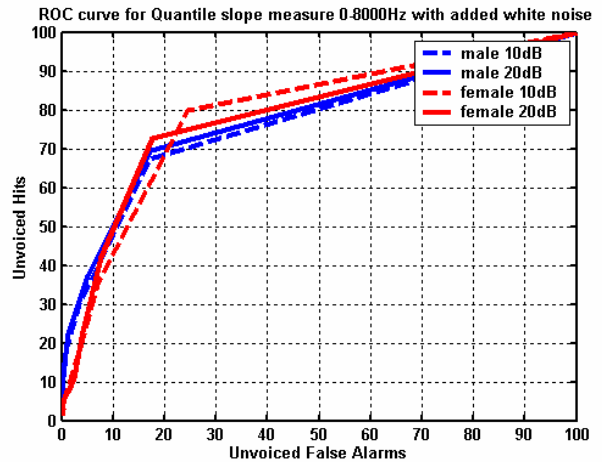
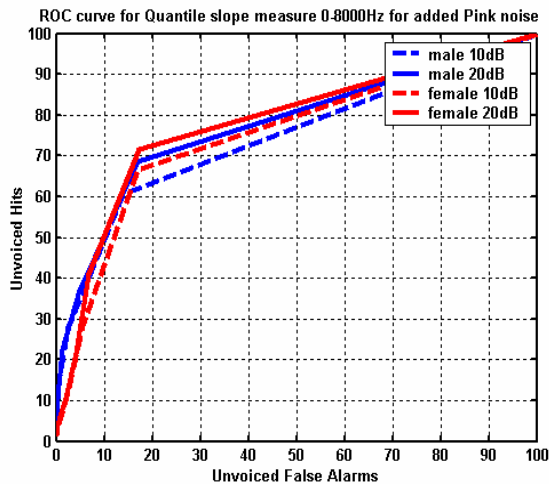


Figure 4.11: ROC curves for speech files corrupted with pink noise



**Figure 4.12: ROC curves for speech files corrupted with pink noise**

However on further experimentation it was found that the threshold for unvoiced detection kept changing with the changing level of SNR in dB. Further experimentation is being carried out to make this measure robust to changes in the signal to noise ratio

## 5. Conclusion

The goal of this research effort has been to develop measures for detection of unvoiced speech in the presence of additive white and pink noise in varying signal to noise conditions. It is seen that the NREM and the ADM measures perform well for added white noise conditions. The Quantile slope measure is robust to varying background noise conditions but has the problem that a single threshold of the measure value cannot be found that can reliably detect unvoiced speech for varying signal to noise conditions. Research is on to overcome this problem and to see how these measures perform for voiced speech detection.

## References

[1]Atal, B. S. and L. R. Rabiner (1976). "A Pattern Recognition Approach to Voiced - Unvoiced - Silence Classification with Applications to Speech Recognition." IEEE Transaction on Acoustics, Speech, and Signal Processing ASSP - 24(3): 201-212.

[2]O'Shaughnessy, "Speech Communication Human and machine", Addison-Wesley, 1987.

[3]F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," Trans. IECE Japan, vol. 53-A, no. 1, pp. 35-42, 1970.

[4]B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. SOC. Amer., vol. 50, pp. 637-655, Aug. 1971.

[5]F. Itakura, "Minimum prediction residual principle applied to speech recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp. 67-72, Feb. 1975.

[6]Hiroyoshi Morikawa, Hiroya Fujisaki, "Adaptive Analysis, of Speech Based on a Pole-Zero Representation", IEEE transactions on acoustic speech and signal processing, Vol. ASSP-30, no. 1, February 1982

[7]G. E. Kopec, A. V. Oppenheim, and J. M. Tribolet, "Speech analysis by homomorphic prediction," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, pp. 40-49, Feb. 1977.

[8]K. Steiglitz, "On the simultaneous estimation of poles and zeros," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, pp. 229-234, June 1977.

[9]B. S. Atal and M. R. Schroeder, "Linear prediction analysis of speech based on a pole-zero representation," J. Acoust. Soc. Amer., vol. 64, pp. 1310-1318, Nov. 1978.

[10]Simon Haykin, "Adaptive Filter Theory", Pearson Education Inc, 2002

[11]Saeed Gazor, Wei Zhang "Speech Probability Distribution," IEEE Signal Processing letters, Vol. 10, no. 7, pp.204-206, July 2003.

[12]A.C Rencher, "Methods of Multivariate Analysis", John Wiley & sons 1995

[13]Khaled Helmi El-Maleh, "Classification-based Techniques for digital coding of speech-plus-noise", thesis submitted to McGill University Montreal, Canada, January 2004