

A Survey and Evaluation of Voice Activity Detection Algorithms



Seshashyama Sameeraj Meduri

(ssme09@student.bth.se, 861003-7577)

Rufus Ananth (anru09@student.bth.se, 861129-5018)

Examiner:

Dr. Sven Johansson

Department of Electrical Engineering School of Engineering Blekinge Tekniska Högskola SE 37175 Karlskrona Sweden Email: <u>sven.johansson@bth.se</u> Supervisor:

Dr. Benny Sällberg

Department of Electrical Engineering School of Engineering Blekinge Tekniska Högskola SE 37175 Karlskrona Sweden Email: <u>benny.sallberg@bth.se</u>

ACKNOWLEDGEMENT

This thesis work was carried out at the Department of Electrical Engineering, Blekinge Institute of Technology, Karlskrona, Sweden under the supervision of Dr. Benny Sällberg.

We would like to express our gratitude to our supervisor, Dr. Benny Sällberg for his guidance, valuable suggestion and important discussions without whose help this thesis would not have been accomplished.

We would also like to appreciate the support and encouragement from family and friends.

Karlskrona, June 2011

Seshashyama Sameeraj Meduri Rufus Ananth

ABSTRACT

The term Voice Activity Detector (VAD) refers to a class of signal processing methods that detects if short segments of a speech signal contain voiced or unvoiced signal data. A VAD is normally using decision rules based on selected estimated signal features. VADs play a major role as a preprocessing block in a variety of speech processing applications such as speech enhancement, speech coding and speech recognition where it is desirable to classify voiced signal parts from unvoiced. This thesis presents a thorough investigation of modern VAD algorithms that are based on energy threshold, zero crossing and other statistical measures. The selected VAD algorithms are implemented in MATLAB and evaluated using objective parameters in different noise environments. The simulation results indicate that the selected methods produce favorable results in the noise environments with SNR above 5dB. VAD based on pattern recognition approach method proved effective when compared to those based on energy threshold, zero crossing measures and statistical measures.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	ii
ABSTRACT	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vi
1. INTRODUCTION	1
1.1. Overview	1
1.2. Introduction	1
1.3. Objective	2
1.4. Framework	2
1.5. Thesis Outline	3
2. VAD METHODS	5
2.1. VAD based on zero crossing rate and energy	5
2.1.1. Zero Crossing Measurement	5
2.1.2. Short-Time Energy	6
2.1.3. Implementation	6
2.2. LED: Linear Energy-Based VAD	7
2.2.1. Full-Band Energy	8
2.2.2. Implementation	8
2.3. ALED: Adaptive Linear Energy-Based Detector	9
2.3.1. Implementation	9
2.4. A Pattern Recognition Approach to Voiced-Unvoiced Classification	10
2.4.1. Zero Crossing Count	10
2.4.2. Log-Energy	12
2.4.3. Normalized Autocorrelation Coefficient	12

2.4	.4.	First Predictor Coefficient	13				
2.4	.5.	Normalized Prediction Error	14				
2.4	.6.	Distance Computation	15				
2.4	2.4.7. Implementation						
2.5.	VAD	Based on Statistical Measures	17				
2.5	.1.	Signal-to-Noise Measure	17				
2	2.5.2.	Variance of SNR Measure	18				
2.5	.3.	Threshold Adaptation and Decision	18				
2.5	.4.	Implementation	19				
3. EV	VAL	UATION OF METHODS	23				
3.1.	Obje	ective Parameters	23				
3.1	.1.	Front End Clipping (FEC)	23				
3.1	.2.	Mid-Speech Clipping (MSC)	23				
3.1	.3.	Over Hang (OVER)	23				
3.1	.4.	Noise Detected as Speech (NDS)	24				
3.2.	NO	ZEUS: A Noisy Speech Corpus	25				
4. RF	ESUI	LTS AND ANALYSIS	26				
4.1.	VAD	Based on Zero Crossing Rate and Energy Measure	26				
4.2.	LED	: Linear Energy-Based VAD	28				
4.3.	ALE	D: Adaptive Linear Energy-Based Detector	30				
4.4.	A Pa	ttern Recognition Approach to Voiced-Unvoiced Classification	31				
4.4	.1.	VAD based on Statistical Measures	33				
4.5.	Sum	Imary	35				
5. CC	DNC	LUSION	36				
REFE	REN	CES	37				

LIST OF FIGURES

Figure 1-1.Block Diagram of a VAD	2
Figure 1-2. Framework for implementation, comparison and evaluation of VAD	
algorithms	3
Figure 2-1.Block diagram for VAD based on Zero Crossing Rate and Energy	
Measurements [11]	7
Figure 2-2.Probability density function for the zero crossing measurement[5]	11
Figure 2-3.Probability density function for the energy measure [5]	12
Figure 2-4. Probability density function for the normalized autocorrelation	
coefficient [5]	13
Figure 2-5.Probability density function for the first LPC coefficient measure	14
Figure 2-6.Probability density function for LPC error measurement [5]	15
Figure 2-7.Block diagram for VAD based on pattern recognition approach [5]	16
Figure 2-8.Block diagram for the VAD method based on statistical measures	20
Figure 3-1.Objective Parameters [1]	24
Figure 4-1.Energy and ZCR Measurements of VAD based on Energy and Zero	
Crossing Rate	26
Figure 4-2.Energy Measurement for LED method	28
Figure 4-3.Energy Measurement of ALED Method	30
Figure 4-4.Extracted Features from Speech Signal using VAD based on Pattern	
Recognition Approach	32
Figure 4-5.SNR Measure of VAD based on Statistical Measures	34

1. INTRODUCTION

1.1. Overview

With the recent advances in speech signal processing techniques, the need to detect the presence of speech accurately in the incoming signal under different noise environments has become a major concern of the industry. The separation of speech segment from the non-speech segment in an audio signal is achieved using a Voice Activity Detectors (VAD). VAD's are a class signal processing methods that detects the presence or absence of speech in short segments of audio signal.

A VAD has a pivotal role as a preprocessing block in wide range of speech applications. An integrated VAD in speech communication system, improves channel capacity, reduces co-channel interference and power consumption in portable electronic devices in cellular radio systems and allows simultaneous voice and data applications in multimedia communications [1], [2]. In slowly varying non-stationary environments where speech is corrupted by noise, a VAD is used to learn noise characteristics and estimate the noise spectrum [3]. Furthermore, the output from the VAD is helpful in improving the performance of the speech recognition systems which applies a technique called non-speech frame dropping (FD) to reduce the number of insertion errors caused by the noise [4].

1.2. Introduction

A basic VAD works on the principle of extracting measured features from the incoming audio signal which is divided into frames of 5-40 ms duration. These extracted features from the audio signal are then compared to a threshold limit usually estimated from the noise only periods of the input signal and a VAD decision is computed. If the feature of the input frame exceed the estimated threshold value, a VAD decision (VAD = 1) is computed which declare that speech is present. Otherwise, a VAD decision (VAD = 0) is computed which declares the absence of speech in the input frame. The block diagram of a basic VAD is shown in fig 1.1.



Figure 1-1.Block Diagram of a VAD

1.3. Objective

The goal of this thesis is to carry out a thorough investigation of modern VAD algorithms based on energy threshold, zero crossing rate, and statistical measures and thereby implement the same in MATLAB. These algorithms are later compared for their correct classification of the input signal data into voiced and unvoiced classes in different noise environments such as airport, babble, restaurant and train environments with SNR values ranging from 0-15dB. These selected VAD algorithms are analyzed and evaluated using the four objective parameters [1] which are,

- (i) FEC (Front-End Clipping)
- (ii) MSC (Mid-Speech Clipping)
- (iii) OVER (Over Hang)
- (iv) NDS (Noise Detected as Speech)

1.4. Framework

Framework 1.2 outlines the framework for the implementation, comparison and evaluation of the VAD algorithms. The framework is mainly divided into two blocks. In the first block, the VAD algorithms are implemented in MATLAB and the signal data is classified into voiced and unvoiced segments and the decisions are computed. The VAD decisions obtained from the first block are passed into the second block, where a reference VAD decision data recorded in

a quiet environment is used to calculate the performance attributes FEC, MSC, OVER and NDS. Using these parameters, the VAD algorithms are compared and evaluated.



Figure 1-2. Framework for implementation, comparison and evaluation of VAD algorithms

1.5. Thesis Outline

This thesis report is outlined in five chapters. The first chapter introduces the Voice Activity Detectors (VADs), explains the working principle of a basic VAD, objective of this thesis work and framework to implement these VAD methods in MATLAB.

In chapter 2 describes five VAD methods are described which are VAD based on Zero Crossing Rate and Energy measurement, Linear based Energy Detector (LED), Adaptive Linear based Energy Detector (ALED), VAD based on pattern recognition approach and VAD based on Statistical Measures.

Chapter 3 explains the evaluation of the methods using four objective parameters, front end clipping (FEC), mid speech clipping (MSC), OVER (Over

Hang) and NDS (Noise Detected as Speech) which are described with formulas and figure. This is followed by a brief description about the test database used for evaluating the methods.

In chapter 4, the analysis and results for each method are presented using tables containing calculations obtained from objective parameters.

Chapter 5 ends with the conclusion of the thesis work.

2. VAD METHODS

Over the years, different approaches have been proposed for the detection of speech segments in the input signal data. The early VAD algorithms were based on extracting features such as short-time energy, zero crossing rate, linear prediction [5] and pitch analysis [6]. In the recent years, classification of voiced and unvoiced segments was based on cepstral coefficients [7], wavelet transform [8], periodicity measure [9] and statistical models [10].

In this thesis, five different VAD algorithms based on short-time energy, zero crossing and statistical measures are presented.

2.1. VAD based on zero crossing rate and energy [11]

This method is a simple and fast approach method to divide the given speech signals into voiced and unvoiced classes. The method works on the combination of zero crossing rate and energy calculations.

2.1.1.Zero Crossing Measurement

Zero crossing rate can be defined as the number of times the successive samples in a speech signal have different algebraic signs or the amplitude of signal crosses the value of zero. Equation 2.1 defines the zero crossing count, Z_n as,

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)]| - |sgn[x(m-1)]|w(n-m)$$
(2.1)

Where

$$sgn[x(m)] = \begin{cases} 1 & x(m) \ge 0\\ -1 & x(m) < 0 \end{cases}$$
$$w(n) = \begin{cases} \frac{1}{2N}, & 0 \le n \le N-1\\ 0, & otherwise \end{cases}$$

N is the duration of the window used in the method.

Zero crossing rate indicates the presence or absence of speech in the input signal. If the zero crossing rate is high, the frame is considered to be unvoiced and if it is low, the frame is considered to be voiced frame.

2.1.2.Short-Time Energy

Short-time energy calculation is another parameter used in the classification of voiced and unvoiced segments. If the energy of the incoming frame is high, the frame is classified into voiced frame and if the energy of the incoming frame is low, it is classified into unvoiced frame. The short-time energy of the frame, x(m) denoted by E_n is defined according to the equation (2.2) as

$$\sum_{m=-\infty}^{\infty} [x(m)h(n-m)]^2$$
(2.2)

Where,

$$h(n) = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), & 0 \le n \le N-1 \\ 0, & Otherwise \end{cases}$$

In this method hamming window is used which give much attenuation outside the band pass when compared to the rectangular window.

2.1.3. Implementation

The data flow for classifying the input signal into voiced or unvoiced segments is done as shown in the block diagram of figure 2.1. The method begins with end point detection which is a process of detecting the starting and ending point of a speech utterance. Following the detection of end points, small sample of silence interval prior to the commencement of speech signal is taken and shorttime energy and zeros-crossing rates are calculated. These measures are used as thresholds for energy and zero crossing rate.

In the frame by frame block, the speech signal is divided into non-overlapping frames of 400 samples at 8 KHz sampling frequency which is equivalent to 50ms time duration. Short time energy and average zero crossing rate measures of these frames are compared with their threshold values. The frames are classified as voiced segments if the short time energy of the frame is greater than its calculated threshold and the average zero crossing rate is less than the zero crossing threshold. Else, the frames are classified as unvoiced segments. If the decision is unclear, the frame is sub-divided into half the size of the original frame, that is, it is sub-divided into two sub-frames of 200 samples each which is equivalent to 25ms time duration. The energy and zero crossing measures from these sub-divided frames is calculated and compared with the threshold values to classify the subdivided frames into voiced and unvoiced classes. This process is repeated until all frames are classified into the two classes.



Figure 2-1.Block diagram for VAD based on Zero Crossing Rate and Energy Measurements [11]

2.2. LED: Linear Energy-Based VAD [12]

In the previous method, the threshold remained as a constant through the entire process. This method works on the principle of updating the threshold value adaptively.

2.2.1.Full-Band Energy

The full-band energy measure calculates the energy of the incoming frames. This energy, E_i is given by the equation (2.3)

$$E_{j} = \frac{1}{N} \sum_{i=(j-1)N+1}^{jN} x^{2}(i)$$
(2.3)

Where, E_j is the energy of the j-th frame and if x(i) is the i-th sample of speech and the length of the frame is N samples, then frame j, f_j is represented by equation (2.4) as

$$f_j = \{x(i)\}_{i=(j-1)N+1}^{jN}$$
(2.4)

2.2.2.Implementation

Calculating the threshold value is very important as it estimates the background noise. In this method, it is assumed that the initial 100ms does not contain any speech. Therefore, the mean energy of the initial 100ms is calculated according to the equation (2.5)

$$E_r = \frac{1}{\nu} \sum_{m=0}^{\nu} E_m$$
 (2.5)

Where, E_r is the initial threshold, v is the number of frames whose individual size is 80 samples which is equivalent to 10ms sampled at 8 KHz frequency.

The speech signal is divided into frames of 10ms duration at 8 KHz sampling frequency. This corresponds to 80 samples per frame.

The energy of the incoming frame is calculated according to the equation (2.3) and compared to the estimated threshold. If the energy of the frame is greater than the threshold, the frame is judged as a voiced frame. Otherwise, the frame is considered to be an unvoiced frame and the new threshold is calculated as per the equation (2.6)

$$E_{r,new} = (1-p).E_{r,old} + p.E_{silence}$$
 (2.6)

Where, $E_{r,new}$ is the updated threshold value, $E_{r,old}$ is the previous threshold value, $E_{silence}$ is the energy of the recent unvoiced frame and, 0 . In this method, the coefficient*p*takes the value of 0.2 [13].

2.3. ALED: Adaptive Linear Energy-Based Detector [12]

This method is an improvement of the previous method of linear energy-based detector (LED). The coefficient 'p' in the equation (2.6) is limited to constant value which is insensitive to the varying noise statistics. To overcome this limitation, E_r , the energy threshold is computed using the second order statistics of the unvoiced frames.

2.3.1.Implementation

A buffer of '*m*' silence frames is used in this method. When a new silence frame is detected, it is added to the buffer by discarding the oldest frame. The variance of this buffer is calculated in terms of its energy according to the equation (2.7)

$$\sigma^2 = var[E_{silence}] \tag{2.7}$$

The background noise in the speech signal is detected by comparing the variance of the buffer before the addition of the new silence frame with the variance of the buffer after a new silence frame has been added to the buffer. If σ_{old}^2 denotes the variance of the buffer before the addition and σ_{new}^2 denotes the buffer after the addition, a change in background as in equation (2.8) indicates that

$$\sigma_{new}^2 > \sigma_{old}^2 \tag{2.8}$$

Hence, a new rule is formulated to vary p in equation (2.6) by the table (2.1)

$\frac{\sigma^2_{\text{new}}}{\sigma^2_{\text{old}}} \ge 1.25$	0.25
$1.25 \geq \frac{\sigma^2_{\text{new}}}{\sigma^2_{\text{old}}} \geq 1.10$	0.20
$1.10 \ge \frac{\sigma^2_{\text{new}}}{\sigma^2_{\text{old}}} \ge 1.0$	0.15
$1.0 \geq \frac{\sigma^2_{\text{new}}}{\sigma^2_{\text{old}}}$	0.10

Table 2.1. Value of 'p' depending on $\frac{\sigma^2_{\text{new}}}{\sigma^2_{\text{old}}}$ [12]

2.4. A Pattern Recognition Approach to Voiced-Unvoiced Classification [5]

In this method, the concept of pattern recognition is applied to classify the given speech signal into two classes which are voiced and unvoiced. This method employs the measurement of five different parameters. These features extracted from the speech signal are zero crossing count, speech energy, and correlation between adjacent speech samples, first predictor coefficient from linear predictive coding analysis and the energy in the prediction error. The five parameters are simple and highly effective for the classification.

The classification of speech segment into voiced or unvoiced classes is achieved by computing the weighted Euclidean distance measure with the parameters extracted from the speech segment and assigning to class with minimum distance.

2.4.1.Zero Crossing Count

If the successive samples in the speech signal have different algebraic signs, then a zero crossing is said to occur. Zero crossing rate can be defined as the rate of occurrence of these zero crossings in a frame which is a measure of the frequency content of a signal. The zero crossing rate for speech is given by the equation (2.9) [14] and is similar to equation (2.1)

$$N_{z} = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|w(n-m)$$
(2.9)

Where

$$sgn[x(n)] = \begin{cases} 1 & x(n) \ge 0\\ -1 & x(n) < 0 \end{cases}$$

And

$$w(n) = \begin{cases} \frac{1}{2N} & 0 \le n \le N-1 \\ 0 & otherwise \end{cases}$$

The energy is concentrated at low frequencies for voiced speech and for the unvoiced speech; the energy concentration is at the high frequencies. Thus, the zero crossing count, N_z for voiced speech is lower and is typically in the range of 0-30 and for unvoiced speech, it has higher rate ranging from 10-100. The probability density function for the zero crossing measurement is shown in figure (2.2).



Figure 2-2. Probability density function for the zero crossing measurement[5]

2.4.2.Log-Energy

The log-energy, E_s is defined by the equation (2.10) as

$$E_{s} = 10 * \log_{10} \left(\epsilon + \frac{1}{N} \sum_{n=1}^{N} x^{2}(n) \right)$$
(2.10)

Where ϵ is a small positive constant which has a value of 10^{-5} .

The energy of the voiced signal is considered to be higher than the energy of the unvoiced signal. The distribution function of the voiced and unvoiced signal for the log-energy measure is shown in figure (2.3)



Figure 2-3. Probability density function for the energy measure [5]

2.4.3.Normalized Autocorrelation Coefficient

The normalized autocorrelation coefficient, C_1 gives the correlation between the adjacent samples of the signal which usually varies between -1 and +1. This value of C_1 for voiced signal is close to unity because of the frequency concentration in the low frequencies and for unvoiced signal, it is close to zero. The normalized correlation coefficient at unit delay is defined by equation (2.11) as

$$C_1 = \frac{\sum_{n=1}^N s(n)s(n-1)}{\sqrt{(\sum_{n=1}^N s^2(n))(\sum_{n=0}^N s^2(n))}}$$
(2.11)

The probability density function of the normalized autocorrelation coefficient, C_1 is shown in figure (2.4)



Figure 2-4. Probability density function for the normalized autocorrelation coefficient [5]

2.4.4.First Predictor Coefficient

The first predictor coefficient of a p-pole is a number obtained from the linear predictive coding (LPC) analysis. Its value varies from -5 for voiced signal to 1 for unvoiced signal. The first predictor coefficient is obtained by minimizing the equation (2.13).

Figure (2.5) shows the distribution function for the first LPC coefficient measure.



Figure 2-5. Probability density function for the first LPC coefficient measure

2.4.5.Normalized Prediction Error

The normalized prediction error, E_p is defined by the equation (2.11)

$$E_p = E_s - 10 * \log_{10} \left(10^{-6} + \left| \sum_{k=1}^p \propto_k \emptyset(0,k) + \emptyset(0,0) \right| \right)$$
(2.11)

Where

$$\phi(i,k) = \frac{1}{N} \sum_{n=1}^{N} s(n-i)s(n-k)$$
(2.12)

Equation (2.12) is the (i, k) term of the covariance matrix, E_s is the log-energy defined in equation (2.10) and \propto_k is the predictor coefficient obtained by minimizing equation (2.13)

$$E = \frac{1}{N} \sum_{n=1}^{N} \left[s(n) + \sum_{k=1}^{p} \propto_{k} s(n-k) \right]^{2}$$
(2.13)

The normalized prediction error gives the measure of the non-uniformity of the spectrum. The prediction error is higher for voiced signal compared to the

unvoiced signal. The parameter E_p varies between 0 and 40 dB. Figure (2.6) shows the distribution function of the prediction error parameter.



Figure 2-6. Probability density function for LPC error measurement [5]

2.4.6.Distance Computation

A training set is created by manually marking a clean speech recorded in a quiet environment for the speech periods and segmenting the signal into regions of voiced and unvoiced signal. These segments each are then divided into blocks of size 10ms duration and the five measurements as explained in section 2.4.1 are calculated for each block and saved in a test file.

Let $x_i(n)$ be the measurement vector for the nth block belonging to class i (i = 1 for voiced decision and 2 for unvoiced decision) and N_i be the total number of blocks in class i, we have from equations (2.14) and (2.15), the mean vector \boldsymbol{m}_i and the covariance matrix \boldsymbol{W}_i for each class of i.

$$\boldsymbol{m}_{i} = \frac{1}{N_{i}} \sum_{n=1}^{N_{i}} x_{i}(n)$$
 (2.14)

$$\boldsymbol{W}_{i} = \frac{1}{N_{i}} \sum_{n=1}^{N_{i}} x_{i}(n) x_{i}^{t}(n) - \boldsymbol{m}_{i} \boldsymbol{m}_{i}^{t}$$
(2.15)

The distance measure \hat{d}_i is therefore formulated using the equation (2.16)

$$\widehat{d}_i = (\boldsymbol{x} - \boldsymbol{m}_i)^t \boldsymbol{W}_i^{-1} (\boldsymbol{x} - \boldsymbol{m}_i)$$
(2.16)

Where x is the measurement vector for the incoming speech blocks which are to be classified into voiced and unvoiced classes.

2.4.7.Implementation

The practical implementation of the algorithm is shown with the help of a block diagram in figure (2.7)



Figure 2-7.Block diagram for VAD based on pattern recognition approach [5]

A low pass filter with cut off frequency 4 KHz sampling frequency of 10 KHz is applied at the beginning of the process. The output is then high pass filtered at 200 Hz to remove any dc or low-frequency hum from the signal. The signal is now divided into blocks of 10ms duration with 100 samples each.

Following the filtering, the five measurements are computed on each block of size 10ms duration and stored in a vector \mathbf{x} . This vector is used to estimate the distances for each class with their respective mean vectors and covariance

matrices obtained from equations (2.14) and (2.15). The distance is computed by equation (2.16) and the blocks are classified into voiced and unvoiced classes using minimum probability-of-error decision. Based on the distance measure, \hat{d}_i the blocks are classified into class *i* such that the distance is minimized. This process is continued till all the blocks are classified into voiced and unvoiced classes.

2.5. VAD Based on Statistical Measures [15], [16]

This method describes a statistical method which makes use of signal to noise ratio measure for the detection of speech segment in the input signal. The method incorporates estimation of low-variance spectrum and adaptive threshold mechanism for the detection of voiced segments in the input signal. The expected noise power spectral density and the variance of 'signal to noise ratio' measure are estimated from the non-speech periods. The adaptive threshold computation improves the performance of the VAD. The method is described in detail in the following section.

2.5.1.Signal-to-Noise Measure

Consider a signal corrupted by additive noise which is modeled using equation (2.17) as

$$x_k(n) = s_k(n) + v_k(n)$$
 (2.17)

where $s_k(n)$ is the clean speech and $v_k(n)$ is the additive noise of the kth frame.

It is assumed that speech and noise are independent and that the noise is longterm stationary and the speech is short-term stationary.

Spectrum estimation techniques are the common methods to analyze the signal. As it is known that periodogram is considered as an inconsistent spectral estimator, a low-variance spectrum estimation technique is used in this method to evaluate the spectral content of the signal. The Welch method of overlapping windows was used to generate reduced variance, reduces resolution power spectral density (PSD) estimate, $P_{xx,k}(f_l)$. M sub-frames overlapped by 50% and each sub-frame of length L are windowed with a hanning window.

The signal-to-noise ratio (SNR) measure is defined by equation (2.18)

$$\psi_k(f_l) = \frac{P_{xx,k}(f_l)}{\hat{P}_{vv}(f_l)} - 1$$
(2.18)

Where $\hat{P}_{vv}(f_l)$ is the expected value of the noise PSD and $P_{xx,k}(f_l)$ is the PSD of the current frame k for a particular spectral bin f_l .

$$\hat{P}_{\nu\nu}(f_l) = \frac{1}{K} \sum_{k=0}^{K-1} P_{xx,k}(f_l)$$
(2.19)

Equation (2.19) gives the expected value of the noise PSD which is the sample mean calculated over an initial period of non-speech activity and k is the total number of frames during the initial period of non-speech activity.

For the periods of non-speech activity, when x=v, expected value of SNR measure given by equation (2.18) is modified to equation (2.20)

$$\psi_k(f_l) = \frac{P_{\nu\nu,k}(f_l)}{\hat{P}_{\nu\nu}(f_l)} - 1$$
(2.20)

2.5.2.Variance of SNR Measure

The variance of SNR measure is determined for the non-speech activity and is given by the equation (2.21)

$$\sigma_{\nu,k}{}^{2} = E[\psi_{k}{}^{2}(f_{l})]$$
(2.21)

Where $\sigma_{v,k}^2$, is the variance of SNR measure during non-speech activity and is estimated by calculating the average square of the SNR.

2.5.3.Threshold Adaptation and Decision

For the decision process, two hypotheses, null and alternative hypotheses are considered. The null and alternative hypotheses represent the non-speech and speech cases respectively. They are represented as follows

$$H_0: \psi_k(f_l) = \frac{P_{\nu\nu,k}(f_l)}{\hat{P}_{\nu\nu}(f_l)} - 1$$

$$H_1: \psi_k(f_l) = \frac{P_{vv,k}(f_l) + P_{ss,k}(f_l)}{\hat{P}_{vv}(f_l)} - 1$$

Where H_0 and H_1 represent the null and alternative hypotheses and $P_{ss,k}(f_l)$ is a PSD estimate of the speech in the f_l^{th} spectral bin.

The threshold $\eta_k'(f_l)$ is determined from the noise statistics and false-alarm probability by the equation (2.22)

$$\eta_{k}'(f_{l}) = \sqrt{2\sigma_{\nu,k}^{2}(f_{l})} \cdot erfc^{-1}(2P_{FA})$$
(2.22)

Where $\sigma_{v,k}{}^2(f_l)$, is the variance of the SNR measure during non-speech activity in the $f_l{}^{th}$ spectral bin, P_{FA} is the probability of false alarm and erfc(u) is the complementary error function [18].

The divided frames are classified into speech and non-speech classes based on the comparison between the average SNR and average threshold according to the equation (2.23)

$$\frac{1}{L}\sum_{f_l=0}^{L-1}\psi_k(f_l) \stackrel{>H_1}{<} \frac{1}{L}\sum_{f_l=0}^{L-1}\eta_k'(f_l)$$
(2.23)

If the average SNR is greater than or equal to the average threshold, H_1 is decided. Otherwise, H_0 is decided.

2.5.4.Implementation

The method is described by the block diagram shown in figure 2.8.



Figure 2-8.Block diagram for the VAD method based on statistical measures

The procedure starts by dividing the input signal into frames of 20ms time duration (160 samples, sampled at 8 KHz) with 50% overlap between the frames. Welch method of overlapping sub-frames of length L equal to 16 samples is used. The frame results in M=19 overlapping sub-frames. Following the process of dividing frames, SNR $\psi_k(f_l)$ is calculated using the equations

(2.18) and (2.20). The short exponential average over time for the SNR $\psi_k(f_l)$ is calculated according to the equation (2.24) and compared to the threshold, η' .

$$\hat{\psi}_{k}(f_{l}) = (1 - \alpha_{\psi_{k}(f_{l})})\psi_{k}(f_{l}) + \alpha_{\psi_{k}(f_{l})}.\hat{\psi}_{k-1}(f_{l}) \quad (2.24)$$

$$\alpha_{\psi_{k}(f_{l})} = \begin{cases} \alpha_{C,\psi_{k}(f_{l})}, & \hat{\psi}_{k}(f_{l}) \leq \hat{\psi}_{k-1}(f_{l}) \\ 0, & \hat{\psi}_{k}(f_{l}) > \hat{\psi}_{k-1}(f_{l}) \end{cases}$$

Where $\propto_{\psi_k(f_l)}$ is the averaging coefficient and $\propto_{C,\psi_k(f_l)}$ is a constant value and is given in table (2.2).

The threshold $\eta_k'(f_l)$ is found by calculating the variance of SNR measure $\sigma_{v,k}^2(f_l)$ for non-speech periods and is exponentially averaged over time according to the equation (2.25). The threshold $\eta_k'(f_l)$ is calculated from the equation (2.22) and is exponentially averaged over time according to equation (2.26).

$$\hat{\sigma}_{v,k}^{2}(f_{l}) = (1 - \alpha_{\sigma_{v}^{2}}) \sigma_{v,k}^{2}(f_{l}) + \alpha_{\sigma_{v}^{2}} \cdot \hat{\sigma}_{v,k-1}^{2}(f_{l})$$
(2.25)

$$\widehat{\eta'}_{k}(f_{l}) = (1 - \alpha_{\eta'}) \eta_{k'}(f_{l}) + \alpha_{\eta'} \cdot \widehat{\eta'}_{k-1}(f_{l})$$
(2.26)

Measure	Value
$\eta_k'(f_l)_{MAX}$	1.5
$\eta_k'(f_l)_{MIN}$	0.45
$\widehat{P}_{vv,MIN}(f_l)$	0.001
$\propto_{\mathcal{C},\psi_k(f_l)}$	0.75
$\propto_{P_{vv}}$	0.999
$\propto_{\sigma_v^2}$	0.35
$\propto_{\eta'}$	0.75
L	16
М	19
P _{FA}	5%

Table 2.2 Parameters for VAD implementation [15]

The smoothing coefficients, $\propto_{\sigma_v^2}$, $\propto_{\eta'}$, the constant for probability of false alarm P_{FA} , the upper limit $\eta_k'(f_l)_{MAX}$, lower limit $\eta_k'(f_l)_{MIN}$ for $\eta_k'(f_l)$, $\hat{P}_{vv,MIN}(f_l)$ and $\propto_{P_{vv}}$ are presented in table (2.2). The arithmetic mean over frequency of threshold is calculated and compared to the SNR and based on the decision rule given in equation (2.23) the frames are classified into speech and non-speech classes.

The limiting of the threshold is applied to limit the estimated variance of the background noise effectively. The upper limit trades the false rejection for false alarms. The limit on the expected noise power estimate is applied to avoid the SNR measure tending towards infinity.

3. EVALUATION OF METHODS

3.1. Objective Parameters

The performance of a VAD method is evaluated using the objective parameters. For measuring the amount of clipping and noise detected as speech, the output from the VAD methods are compared to the ideal VAD decisions. The ideal VAD decisions are obtained by manually marking a clean speech recorded in a quiet environment for speech and non-speech periods. VADs are evaluated using four traditional objective parameters [1], [15], [18].

3.1.1.Front End Clipping (FEC)

FEC occurs when speech is misclassified as noise while passing from noise into speech activity. FEC is obtained using the equation (3.1).

$$\% FEC = \frac{N_F}{N_{speec h}} \times 100 \tag{3.1}$$

Where N_F , is the number of samples misclassified as noise when passing from noise to speech activity and N_{speech} is total number of samples of speech from an ideal VAD.

3.1.2.Mid-Speech Clipping (MSC)

Mid-speech clipping occurs when speech is misclassified as noise during an utterance. The MSC measure in percentage is obtained from the equation (3.2)

$$\% MSC = \frac{N_M}{N_{speec h}} \times 100 \tag{3.2}$$

Where N_M , is the number of samples misclassified as noise during an utterance.

3.1.3.Over Hang (OVER)

OVER is the measure of noise interpreted as speech while passing from speech to non-speech or noise period. OVER is measured using the equation (3.3)

$$\% OVER = \frac{N_0}{N_{silence}} \times 100 \tag{3.3}$$

Where N_0 , is the number of samples interpreted as speech while passing from speech to silence period and $N_{silence}$ is the total number of samples from silence period of an ideal VAD.

3.1.4. Noise Detected as Speech (NDS)

This is a measure of noise interpreted as speech within a silence period. NDS is calculated by the equation (3.4)

$$\% NDS = \frac{N_N}{N_{silence}} \times 100 \tag{3.4}$$

Where N_N refers to the number of samples interpreted as speech while in silence period.

The four objective parameters are illustrated in figure (3.1)



Figure 3-1. Objective Parameters [1]

FEC and MSC collectively give the measure of the amount of clipping introduced in the signal. OVER and NDS parameters indicate false alarms. The

clipping errors (FEC and MSC) degrade the speech quality and reduce speech intelligibility. The insertion errors (OVER and NDS) reduce the effectiveness of the VAD. Therefore, it is of vital to reduce clipping errors at all cost for better speech intelligibility.

3.2. NOIZEUS: A Noisy Speech Corpus [19],[20]

The noisy speech corpus (NOIZEUS) database was originally developed to facilitate research groups to compare different speech enhancement algorithms. The database consists of 30 IEEE sentences spoken by 3 male and female speakers corrupted by different real-world noises at different SNRs. The noise in the speech corpus was taken from AURORA database [21] . The different noise environments with SNRs 0dB, 5dB, 10dB and 15dB are train noise, babble noise, car noise, exhibition hall noise, restaurant noise, street noise, airport noise and train station noise. These speech sentences from IEEE database [22] were recorded in sound proof booth and noise was artificially added to the speech signal. The sentences were downsampled to 8 KHz from 25 KHz.

For evaluating various VAD algorithms, NOIZEUS database was used. The test database included speech signals recorded in quiet environment by a male and female speaker. Noise from AURORA database was taken and added artificially to these speech signals. Speech signal with four types of additive noises was used for the analysis and evaluation purpose. The different additive noises are airport noise, babble noise, restaurant noise and train noise. The database for four SNR values 0dB, 5dB, 10dB and 15dB was used.

4. RESULTS AND ANALYSIS

The five VAD algorithms described in chapter 2 were implemented in MATLAB version 7.6.324 (R2008a). The VAD decisions for each method were computed and objective parameters (FEC, MSC, OVER, NDS) were obtained. The VAD methods were tested in four different noise environments with SNR values 0dB, 5dB, 10dB, and 15dB for airport noise, babble noise, restaurant noise and train noise. The results are shown in the following sections for each method for male and female speakers.

4.1. VAD Based on Zero Crossing Rate and Energy Measure

This method works on the principle of extracting energy and zero crossing rate features from the input speech signal and comparing them to the threshold to classify the frames into voiced and unvoiced classes. Usually voiced segments have high energy and low zero crossing rate and unvoiced segments have low energy and high zero crossing rate. This is shown in the figure 4.1 for a male speaker.



Figure 4-1.Energy and ZCR Measurements of VAD based on Energy and Zero Crossing Rate

It can be seen from the figure that for voiced segments, the energy measurement is high and zero crossing rate measure is low. The energy measurement is low and zero crossing count is high for unvoiced segments.

The objective parameters for VAD method based on ZCR and energy measure for male and female speaker are presented in table (4.1) below.

Noise Environment		VAD Based on Zero Crossing Rate and Energy									
			Male S	peaker			Female	Speaker			
Noise	SNR(db)	FEC	MSC	OVER	NDS	FEC	MSC	OVER	NDS		
Airport	15	0.685	11.4123	29.8503	7.5795	22.2726	22.6113	0	19.2257		
Airport	10	4.3871	37.7715	5.1154	15.0655	1.022	18.5172	20.5251	28.8783		
Airport	5	8.4682	39.1852	2.7137	5.0218	0.768	31.2167	6.3644	29.939		
Airport	0	1.5523	43.1351	12.8197	22.645	4.3965	45.2225	14.9562	24.5823		
Babble	15	0	20.8279	9.8877	10.0125	5.9627	19.2368	13.9486	31.0263		
Babble	10	2.0624	22.3	12.2583	30.131	14.9129	33.7506	0	15.0093		
Babble	5	5.3491	30.8337	2.8384	13.9426	4.6384	50.9132	12.8348	12.8083		
Babble	0	4.9701	29.398	0.811	45.2277	4.4569	58.7325	13.8425	8.5388		
Restaurant	15	2.5361	24.8652	0	5.0218	0.6531	17.0235	23.1238	22.3813		
Restaurant	10	7.0179	37.7423	2.7137	2.5265	0.768	15.5842	16.07	38.4513		
Restaurant	5	2.8349	24.9016	0	16.3755	8.1398	39.3747	10.183	18.1649		
Restaurant	0	0	16.7978	12.8197	25.1716	25.0605	47.1093	1.1138	14.9828		
Train	15	0.5976	9.1167	0	27.6669	4.8803	17.5254	9.5996	14.9828		
Train	10	2.1353	31.3001	2.7137	25.0468	2.165	19.255	6.3644	19.2522		
Train	5	11.9662	32.4807	0	10.0437	1.0099	27.298	3.2352	20.3394		
Train	0	4.2851	45.6056	0	7.5172	5.479	40.3846	0	30.9732		

Table 4.1 Objective parameters for VAD based on Zero Crossing Rate and Energy Measure

FEC and MSC collectively give the amount of clipping and are called as clipping errors (FEC+MSC). OVER and NDS measures give the false alarm percentages in the detected voiced and unvoiced segments. These are called as insertion errors together (OVER+NDS). Although the FEC measure is low, the amount of MSC measured in the detection is very high which results in degradation of speech quality and reduction of speech intelligibility. The VAD works well in restaurant noise environments for both male and female speakers compared to other environments. The VAD performance in babble noise environment is very poor. The method performs well under noise conditions below 10dB SNR value. The insertion error for this method is very less compared to the clipping errors. FEC measure for the male speaker increases with decrease in SNR value with low values and for female speaker, this measure varies between high and low. MSC measure is high for both male and female speakers under all noise environments. NDS measure is high for female speaker and less for male speaker. The overall performance of the VAD based on zero crossing rate and energy measure is very poor as it introduces a lot of clipping errors which reduces the speech quality.

4.2. LED: Linear Energy-Based VAD

LED method extracts energy feature from the input signal and compares with a threshold computed during initial period of the input signal. If the energy is higher than the threshold, the incoming frame is classified into voiced frame and if the energy is lower than the threshold, the frame is classified into unvoiced frame. The energy for voiced frame is high and for unvoiced frame, it is low. Figure 4.2 show the energy measurement for the LED method for a male speaker.



Figure 4-2. Energy Measurement for LED method

The objective parameters for Linear Energy-Based VAD (LED) are tabulated in the table 4.2.

The results from table 4.2 indicate that the VAD has less percentage of average clipping error. However, the average insertion error is quite high in this method which would reduce the effectiveness of the method. Under babble noise environment, the VAD has least percentage of clipping errors and highest percentage of insertion errors. For babble noise, restaurant noise and train environments, the clipping errors are less and insertion errors are high for female speaker compared to the male speaker. The VAD performs well under train noise environment. FEC measure for both speakers is lesser than the MSC after the detection of voiced and unvoiced segments. The false alarms in the detected speech are high for this method. The VAD maintains the speech intelligibility but reduces the effectiveness of the VAD.

Noise Environment		LED: Linear Energy-Based VAD									
			Male	Speaker		Female Speaker					
Noise	SNR(db)	FEC	MSC	OVER	NDS	FEC	MSC	OVER	NDS		
Airport	15	0.102	4.1175	42.4517	7.5172	5.479	14.5743	2.1215	17.1307		
Airport	10	0.6049	20.4708	7.0493	22.5515	0.1693	20.3919	15.8048	17.1573		
Airport	5	3.2211	26.8838	0.2183	10.1061	2.1045	11.6715	31.7157	12.7818		
Airport	0	2.5361	28.6984	0	22.6762	12.1372	23.3551	26.6773	8.5919		
Babble	15	0.2842	0.5903	27.4797	37.617	2.4734	0	66.5341	2.148		
Babble	10	0.2842	11.1208	17.4984	22.6762	1.022	14.5682	10.7133	23.4155		
Babble	5	0.3862	15.2237	15.4398	10.1061	8.7506	17.0114	0.5834	36.1973		
Babble	0	3.2211	24.6028	4.554	27.6669	0.6531	8.2729	45.4787	23.389		
Restaurant	15	0.2842	14.6261	19.9938	20.0561	1.5058	8.2487	30.655	10.6603		
Restaurant	10	0.102	12.2941	0.2183	20.0873	0.0544	12.6149	49.7746	21.2941		
Restaurant	5	1.0858	18.7582	17.0306	5.053	1.2518	7.7528	52.9303	2.148		
Restaurant	0	3.1992	10.5451	59.7006	15.0343	0.2238	28.6829	30.0716	15.0093		
Train	15	1.9531	7.6301	0	2.5265	1.9896	4.3783	20.8168	2.148		
Train	10	0.3862	16.9946	0	7.5483	0	23.2886	11.9862	19.1726		
Train	5	3.8041	7.0034	24.5165	17.592	0.0544	4.8803	18.987	14.9828		
Train	0	4.2851	24.5445	9.5446	17.5608	9.7303	4.3965	48.4222	6.4439		

Table 4.2 Objective parameters	s for LED: Linear Energy-Based VA	AD
--------------------------------	-----------------------------------	----

4.3. ALED: Adaptive Linear Energy-Based Detector

ALED method is an improvement of the LED method. In this method, the threshold is updated adaptively. The value of p is varied adaptively according to the table 4.1. The energy measurement for a female speaker is shown in figure 4.3. This method works on the same principle of LED.



Figure 4-3. Energy Measurement of ALED Method

The objective parameters for Adaptive Linear Energy-Based Detector are presented in table below.

The VAD works reasonable well for SNR values ranging from 5dB to 15dB. The results from table 4.3 indicate that the FEC and MSC measures increase with decrease in the SNR value. The VAD has lower FEC and MSC measures and higher OVER and NDS measures. The VAD shows good performance in train noise environment which has both less percentage of clipping and insertion errors. This method is the improvement of the previous method. In this method, the amount is insertion errors are reduced with adaptive threshold. However, the amount of clipping errors is slightly higher in babble noise and train noise environment compared to the previous method. The method improves on the effectiveness of VAD.

Noise Environment		ALED: Adaptive Linear Energy-Based Detector									
			Male S	Speaker			Female Speaker				
Noise	SNR(db)	FEC	MSC	OVER	NDS	FEC	MSC	OVER	NDS		
Airport	15	0.102	4.1175	42.4517	7.5172	1.6207	15.5358	15.9639	14.9828		
Airport	10	0.6049	21.6295	7.0493	22.5515	0.1693	17.9669	15.8048	12.8878		
Airport	5	4.3871	16.9582	0.2183	7.5795	3.5559	11.1756	31.7157	12.7818		
Airport	0	2.5361	28.1154	0	20.1809	12.1372	22.3754	28.7987	8.5919		
Babble	15	0.2842	1.1806	27.4797	35.1528	2.4734	0.4898	66.5341	2.148		
Babble	10	0.2842	11.1208	17.4984	22.6762	1.022	20.8817	3.7656	27.685		
Babble	5	0.3862	16.9655	15.4398	10.1061	5.0617	22.3452	0.5834	23.4421		
Babble	0	3.2211	24.6028	2.0586	20.1497	0.6531	8.2729	41.2357	25.537		
Restaurant	15	0.2842	14.6261	19.9938	17.5296	1.5058	8.2487	30.655	10.6603		
Restaurant	10	0.102	12.2941	0.2183	17.5608	0.0544	13.1047	49.7746	21.2941		
Restaurant	5	1.0858	21.0902	17.0306	7.5795	2.2194	8.2426	33.0416	8.5388		
Restaurant	0	3.1992	12.3087	52.2146	20.0561	1.1913	28.1991	30.0716	10.7133		
Train	15	1.9531	7.6301	0	5.053	1.9896	4.3783	20.8168	2.148		
Train	10	0.3862	16.98	0	12.5702	0.5987	23.7724	11.9862	17.0247		
Train	5	3.8041	9.3572	24.5165	17.592	0.0544	1.9533	50.0663	6.4174		
Train	0	4.2851	28.0426	9.5446	17.5608	9.7303	4.3965	48.4222	6.4439		

Table 4.3 Objective parameters for ALED: Adaptive Linear Energy-Based Detector

4.4. A Pattern Recognition Approach to Voiced-Unvoiced Classification

In this method, parameters such as zero crossing count, N_z log-energy, E_s normalized autocorrelation coefficient, C_1 first linear predictor coefficient and normalized prediction error are computed and using the minimum weighted Euclidean distance, the frames are classified into voiced and unvoiced classes.

For voiced segments, the zero crossing count is low, log-energy measurement is high, normalized autocorrelation coefficient is close to unity and first predictor coefficient value is around -5 and for the unvoiced segments, the zero crossing count is high, log-energy measurement is low, normalized autocorrelation coefficient is close to zero and first predictor coefficient value is around 1. These measurements are shown for a female speaker in figure 4.4.



Figure 4-4.Extracted Features from Speech Signal using VAD based on Pattern Recognition Approach

The objective parameters for the VAD based on pattern recognition approach are presented in table 4.4.

The results obtained from table 4.4 indicate that the method performs very well to maintain good speech intelligibility. The FEC+MSC measure is very low in this method. The amount of clipping errors is low which makes this method a good VAD. This VAD method performs well in airport noise and babble noise environments. In airport noise environment, the amount of insertion error is slightly higher than in babble noise environment. The method provides good results for SNR values ranging from 5dB and higher.

Noise Environment		A Pattern Recognition Approach to Voiced-Unvoiced Classification									
			Male S	Speaker		Female Speaker					
Noise	SNR(db)	FEC	MSC	OVER	NDS	FEC	MSC	OVER	NDS		
Airport	15	1.5523	2.3612	14.8784	7.5795	1.3062	4.8863	13.3652	36.3034		
Airport	10	0.102	6.4641	12.0399	37.7105	0	10.2141	18.4566	23.4421		
Airport	5	2.7183	18.7946	14.8784	30.2246	0	8.8111	29.5943	31.9809		
Airport	0	0	14.7136	12.6949	37.617	0	29.7412	21.9305	25.6165		
Babble	15	0	4.7078	14.8784	25.1716	0.2238	13.1592	26.0408	23.4951		
Babble	10	0	6.4786	7.174	20.0873	0	6.3437	16.5739	36.3034		
Babble	5	0	8.8398	4.8971	37.5858	0	15.6205	3.2352	40.5198		
Babble	0	0.8672	25.8199	7.7043	22.5827	2.9451	22.8592	12.2514	17.1042		
Restaurant	15	0.102	6.4568	12.1647	40.1435	0.8406	7.8133	4.349	36.2768		
Restaurant	10	0.685	7.0762	2.277	30.1934	0.1693	12.1976	7.5842	29.8329		
Restaurant	5	0.5028	11.1937	0	30.131	2.7576	11.7017	14.4259	29.8595		
Restaurant	0	2.1353	18.2262	4.7723	30.131	0.7075	27.812	14.4259	34.1554		
Train	15	2.5361	13.4601	5.3337	5.053	1.5058	10.716	13.9486	8.5919		
Train	10	0.102	8.8107	5.3337	25.1716	4.5114	14.5985	0	25.6696		
Train	5	1.5523	20.5582	14.5352	15.0655	2.2738	14.1388	5.3567	19.2257		
Train	0	0.5028	28.2247	5.3337	30.131	5.4245	29.2271	1.6441	36.2768		

Table 4.4 Objective parameters for Pattern Recognition Approach to Voiced-Unvoiced Classification

4.4.1.VAD based on Statistical Measures

This method works on the principal of calculating the Signal-to-Noise Ratio (SNR) measure and comparing it with an optimal threshold value using estimated noise statistics. If the SNR measure is higher than the threshold, then the frame is classified into voiced class. If it is lower than the threshold, then it is classified into unvoiced class. The SNR measure is high for voiced segments and low for unvoiced segments. This is shown in figure 4.5 for a male speaker.



Figure 4-5.SNR Measure of VAD based on Statistical Measures

The objective parameters for VAD based on statistical measures are presented in table 4.5 for four different noise environments.

The results from the table indicate that the VAD based on statistical measures performs well in restaurant noise environment when compared to other environments. The method performs better for SNR value above 10dB. The FEC measure is high in this method and MSC measure is lower. Most of the insertion error is constituted by OVER parameter. The amount of clipping error introduced in this method is high as the SNR value decreases below 10dB. The VAD has lower insertion error compared with the other methods. The VAD exhibited poor performance in train environment with high clipping errors which leads to reduction in speech quality. However, the insertion error was found to be lower than in other environments.

		VAD Based on Statistical Measures									
Noise Env	ironment		Male S	peaker	Female Speaker						
Noise	SNR(db)	FEC	MSC	OVER	NDS	FEC	MSC	OVER	NDS		
Airport	15	8.7669	16.3460	12.6949	0	10.0085	0	15.9905	0		
Airport	10	8.7669	18.6780	32.1273	0	10.9761	17.4226	15.9905	0		
Airport	5	0	9.3354	32.6887	0	24.0445	0	46.5659	0		
Airport	0	0	10.5014	57.2052	0	8.4482	2.9088	50.3050	0		
Babble	15	1.8510	12.8407	12.9445	0	7.5411	5.8116	46.5659	0		
Babble	10	25.8563	3.5053	51.8715	0	3.8582	0	47.3880	0		
Babble	5	0.6850	19.8440	37.4610	0	12.3186	7.7528	30.6285	0		
Babble	0	3.0171	15.1727	22.4891	0	19.5090	10.6555	23.4421	0		
Restaurant	15	0	14.0140	57.2052	0	22.6415	15.4874	19.2257	21.2411		
Restaurant	10	2.1353	9.3354	37.4610	0	6.7610	0	35.6669	0		
Restaurant	5	0	23.3348	42.2333	0	10.0085	0	15.9905	0		
Restaurant	0	17.7744	3.5053	57.2052	0	1.6207	0	50.3050	0		
Train	15	27.7073	3.5053	0	0	36.3026	0	3.2352	0		
Train	10	28.8733	7.0106	7.6107	0	38.2378	0	12.2514	0		
Train	5	27.7073	3.5053	0	0	31.4647	0	0	0		
Train	0	31.2054	4.6713	0	0	42.1081	0	0	0		

Table 4.5 Objective parameters for VAD based on Statistical Measures

4.5. Summary

The results of VADs have been presented in section 4.1 through 4.5. The results indicate that the performance of VADs improves as SNR value increases. Among all the methods analyzed, VAD based on energy measurement and zero crossing rate has poor performance as it introduces great amount of clipping error and insertion error. The total percentage of correct classifications is very low. On the other hand, VAD based on pattern recognition approach exhibits very good performance and it maintains good signal quality. LED and ALED methods have high insertion errors, thereby, making the methods less effective. VAD based on statistical measure works best for SNR values above 10dB.

Some of the errors in the measurements are due to manual mismarking of the test signals. The limiting factors on $\eta_k'(f_l)$ and $\hat{P}_{vv}(f_l)$ contributes to the total error in VAD based on statistical measures.

5. CONCLUSION

This thesis has been a survey of existing VAD methods. The various VAD methods were investigated for this purpose and five VAD methods were selected. These methods were studied and implemented in MATLAB. The implemented methods were analyzed based on the objective parameters calculated under four different noise environments for a male and a female speaker. The five methods selected were based on threshold calculation. The results from VAD based on energy and zero crossing measurement displayed poor performance and VAD based on pattern recognition approach exhibited very good performance. The VAD methods work well for SNR values above 10dB. The Clipping errors (FEC + MSC), were minimum in VAD based on pattern recognition approach and maximum in VAD based on energy and zero crossing measurement. The insertion errors (NDS + OVER) was found to be low in VAD based on statistical measures and high in LED method.

REFERENCES

- F. Beritelli, S. Casale, A. Cavallaero, "A robust voice activity detector for wireless communications using soft computing," *Selected Areas in Communications, IEEE Journal on*, vol.16, no.9, pp.1818-1829, Dec 1998
- [2] K. Li, M. N. S. Swamy and M. O. Ahmad, "An Improved Voice Activity Detection Using Higher Order Statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, pp. 965-974, 2005.
- [3] R. Le Bouquin-Jeannès and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Commun.*, vol. 16, pp. 245-254, 4, 1995.
- [4] J. Ramírez, J. M. Górriz, J. C. Segura, "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness". Robust Speech Recognition and Understanding. pp. 1-22, Ed.: M. Grimm and K. Kroschel, I-TECH Education and Publishing, 2007. ISBN: 978-3-902613-08-0
- [5] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoicedsilence classification with applications to speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, pp. 201-212, 1976.
- [6] A. M. Noll, "Cepstrum pitch determination," J. Acoust. Soc. Amer., vol. 41, 293-309, Feb. 1967.
- [7] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *TENCON '93. Proceedings. Computer, Communication, Control and Power Engineering.1993 IEEE Region 10 Conference on*, 1993, pp. 321-324 vol.3.
- [8] J. Stegmann and G. Schroder, "Robust voice-activity detection based on the wavelet transform," in *Speech Coding for Telecommunications Proceeding*, 1997, 1997 IEEE Workshop on, 1997, pp. 99-100.
- [9] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 139, pp. 377-80, 08, 1992.
- [10] Jongseo Sohn, Nam Soo Kim and Wonyong Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, pp. 1-3, 1999.
- [11] R. G. Bachu, S. Kopparthi, B. Adapa and B. D. Barkana, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy," 2010.
- [12] S. Kirill, V. Ekaterina and S. and Boris, "Dynamical Energy-Based Speech/Silence Detector for Speech Enhancement Applications," *Proceedings of the World Congress on Engineering 2009*, vol. Vol I, pp. pp801-806, 2009.
- [13] R. Venkatesha Prasad, A. Sangwan, H. S. Jamadagni, M. C. Chiranth, R. Sah and V. Gaurav, "Comparison of voice activity detection algorithms for VoIP," in

Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on, 2002, pp. 530-535.

- [14] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
- [15] A. Davis, S. Nordholm and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 412-424, 2006.
- [16] A. Davis and S. Nordholm, "A low complexity statistical voice activity detector with performance comparisons to ITU-T/ETSI voice activity detectors," in Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on, 2003, pp. 119-123 Vol.1.
- [17] S. Haykin, *Communication Systems*, 3rd ed. New York: Wiley, 1994.
- [18] D. K. Freeman, G. Cosier, C. B. Southcott and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on,* 1989, pp. 369-372 vol.1.
- [19] Hu, Y. and Loizou, P. (2007). "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Communication*, 49, 588-601.
- [20] "NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithm" [online]. Available: www.utdallas.edu/~loizou/speech/noizeus
- [21] H. Hirsch, and D. Pearce (2000). "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions." *ISCA ITRW ASR2000*, Paris, France, September 18-20.
- [22] IEEE Subcommittee (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Trans. Audio and Electroacoustics*, AU-17(3), 225-246.