2006

A Novel Robust Mel-Energy Based Voice Activity Detector for Nonstationary Noise and Its Application for Speech Waveform Compression

Syed, Q. Waheeduddin
Louisiana State University and Agricultural and Mechanical College

A NOVEL ROBUST MEL-ENERGY BASED VOICE ACTIVITY DETECTOR FOR NONSTATIONARY NOISE AND ITS APPLICATION FOR SPEECH WAVEFORM COMPRESSION

A Thesis
Submitted to the Graduate Faculty of the
Louisiana State University and
Agriculture and Mechanical College
in partial fulfillment of the
requirements of the degree of
Master of Science in Electrical Engineering
in
The Department of Electrical Engineering

by
Waheeduddin Q. Syed
B.S., Jawaharlal Technological University, 2004
December 2006

ACKNOWLEDGEMENTS

I would like to begin with thanking the almighty god for making everything possible. Next, I would like to sincerely acknowledge my major professor Dr. Hsiao-Chun Wu for his guidance. This work would not have been possible without his able guidance and unwavering support for me. I sincerely thank my committee members Dr. Subhash Kak, and Dr. Lu Peng for their valuable comments and suggestions. I would also like to thank Mr. Mike Brookes for providing some of supportive Matlab routines in the form of speech processing tool box (VoiceBox). Finally, I would also like to thank my parents, sisters and fellow researchers for their unflinching support and useful discussions throughout my research.

This research work has been supported by Information Technology Research Award for National Priorities from the National Science Foundation (NSF-ECS 0426644), Research Initiation Grant from the Southeastern Center for Electrical Engineering Education, Research Enhancement Award from the Louisiana-NASA Space Consortium, NSF-Louisiana EPSCOR Pilot Fund, and Faculty Research Grant from the Louisiana State University.

TABLE OF CONTENTS

ABSTRACT

The voice activity detection (VAD) is crucial in all kinds of speech applications. However, almost all existing VAD algorithms suffer from the nonstationarity of both speech and noise. To combat this difficulty, we propose a new voice activity detector, which is based on the Mel-energy features and an adaptive threshold related to the signal-to-noise ratio (SNR) estimates. In this thesis, we first justify the robustness of the Bayes classifier using the Mel-energy features over that using the Fourier spectral features in various noise environments. Then, we design an algorithm using the dynamic Mel-energy estimator and the adaptive threshold which depends on the SNR estimates. In addition, a realignment scheme is incorporated to correct the sparse-and-spurious noise estimates. Numerous simulations are carried out to evaluate the performance of our proposed VAD method and the comparisons are made with a couple existing representative schemes, namely the VAD using the likelihood ratio test with Fourier spectral energy features and that based on the enhanced time-frequency parameters. Three types of noise, namely white noise (stationary), babble noise (nonstationary) and vehicular noise (nonstationary) were artificially added by the computer for our experiments. As a result, our proposed VAD algorithm significantly outperforms other existing methods as illustrated by the corresponding receiver operating curves (ROCs). Finally, we demonstrate one of the major applications, namely speech waveform compression, associated with our new robust VAD scheme and quantify the effectiveness in terms of compression efficiency.

1. INTRODUCTION

1.1. Motivation

Nowadays, speech processing techniques can be applied in a wide variety of devices such as cellular handsets, internet search machines, call-in telephony services, etc. Despite of its constant growth, voice activity detection, one of important speech processing problems, is still intriguing to many researchers [1]. A *voice activity detector* is a pre-processing system for speech recognition systems, isolated word boundary detection systems, cell phones and speech enhancement systems. Voice activity detection (VAD) algorithm is designed to distinguish the speech from the background noise among short-time frames. The importance of the VAD system to the speech processing applications can be easily found in the existing literature. For instance, the VAD system plays a key role in the spectral subtraction techniques [2]. Furthermore, a major source of the errors incurred in the automatic speech recognition systems (ASRs) is the inaccurate detection of the beginning and the ending of utterances according to [3]. Therefore, a robust voice activity detector can definitely improve the overall performance of many speech applications.

A typical voice activity detector can be divided into two parts, namely the feature extraction module and the pattern classification module [4, 5]. In the early developed VAD algorithms, the features were extracted from the short-time energy, zero-crossing rates [6], linear predictive coding coefficients [7] and cepstral coefficients [8]. Recently, Mel-energy features [3], the wavelet transforms [9], the correlation coefficients [10], and the likelihood ratios [11, 12] have been adopted as the underlying features for the VAD techniques.

In the next section a more detail information is provided regarding some of the existing speech detection techniques.

1.2. Review of Existing Speech Detection Techniques

In this section we present a literature survey of the existing wok done in the area of voice activity detection. More details of different speech detection techniques are described in the following subsections

1.2.1. Speech Detection Using Energy and Zero Crossing Rate

A simple but efficient speech detection algorithm has been proposed in [6] which classify the frames based on the differential parameters of logarithmic energy and zero crossing rate. In addition, the recommendation for G.729 Annex B [13] proposed a feature vector consisting of linear prediction coefficients, full-band energy, low-band energy, and zero-crossing rate. This standard was developed in collaboration with France Telecom-CNET (FT-CNET), the University of Sherbrooke, NIT and AT&T Bell labs.

1.2.2. Speech Detection Using Wavelet Transforms

A new VAD method based on the *perceptual wavelet packet transform* (PWPT) and the *Teager energy operator* (TEO) was proposed in [9]. Accordingly, first, the speech is decomposed into the critical sub-bands using PWPT and then the voice activity shape (VAS) parameter is derived from these sub-bands using TEO. Finally, the VAS parameters are used for the speech detection. The main advantage of this new VAD scheme in [9] is that it does not need the preset thresholds or *a priori* knowledge of the signal-to-noise ratio (SNR) as compared to any other conventional VAD method.

1.2.3. Speech Detection Using Correlation Coefficient

In [10], a new VAD algorithm was proposed to improve the word boundary detection for the variable background noise levels. Noise parameters are first estimated from the initial

frames and then these parameters are updated during the silence periods using a first-order autoregressive filter. The robust parameters used in this algorithm are the correlation coefficients for the instantaneous spectrum and an average of background noise spectrum. Subsequently, a statistical approach using a simple binary Markov model is taken for the speech detection.

1.2.4. Speech Detection Using Statistical Model Based VAD

Sohn et al. had proposed a statistical model based VAD (SMVAD) in which the decision rule was derived from the likelihood ratio test (LRT) by estimating the unknown parameters using the maximum likelihood (ML) criterion [11]. Further improvement was achieved by optimizing the decision rule using the Decision-directed (DD) method for the estimation of the unknown parameters [12]. The proposed algorithm further optimized the decision rule by adapting the decision threshold using the measured noise energy in order to achieve the robustness in low SNRs.

1.2.5. Speech Detection Using Cepstral Features and Mel-Energy Features

Haigh et al. showed the robustness to different background noise levels for the successful end-of-speech detection using the thresholds based on the cepstral features [8]. Lin et al. proposed a robust word boundary detection method (ETF VAD) based on the enhanced time frequency (ETF) and the minimum Mel-scale frequency band (MIMSB) parameters extracted from the multi-band spectral analysis using the Mel-scale frequency banks [3].

Among all of the adopted features for VAD, the Mel-spectra have been shown to be very promising in the previous VAD method using the threshold based on the minimum sub-band Mel-energy [3]. According to several experiments in [3], the Mel-spectral features would lead to the most robust VAD performance compared with almost all of

other features. However, the comprehensive studies associated with the Mel-spectral features for VAD cannot be found in the existing literature.

In most of the aforementioned techniques, the feature extraction is followed by the threshold detection. In the realistic environments, there exist nonstationary noises such as babble noise and car noise. Therefore, the static thresholds, which can only depend on the information extracted from the first few frames, would cause numerous classification errors [10]. Hence, dynamic or adaptive thresholds were proposed to combat the problem of nonstationary noises [3]. Nevertheless, how to appropriately adjust the threshold dynamically is still very challenging up to now [3, 10].

In this thesis, we propose a new adaptive threshold, which depends on the signal-to-noise ratio (SNR) estimates and results from the dynamical speech and noise information. Consequently, it can lead to much better VAD performance in the presence of nonstationary noise. Furthermore, we extend our new VAD technique for the application of the speech waveform compression, which can be used in the voice communications and storage [14]. In the next section we discuss some of the applications of the VAD.

1.3. VAD Applications

VAD algorithm is a crucial front-end mechanism of many speech processing applications, such as robust speech recognition, speech compression and speech enhancement. More details of the aforementioned applications are described in the following subsections.

1.3.1. Speech Recognition

Numerous VAD techniques have been proposed which are used in the front-end of the speech recognition systems. A robust VAD algorithm improves the recognition accuracy and simplifies the speech recognition system structure.

1.3.2. Speech Compression

Another VAD application is related to the removal of the unvoiced frames to reduce the voice data necessary for transmission. In addition, the VAD is used in the cellular systems for reducing co-channel interference and power consumption of any subscriber's device [15]. As the outcome from the VAD, the speech compression techniques can be used in multimedia and voicemail applications for the efficient voice data storage [16].

1.3.3. Speech Enhancement

VAD is also used in the front end of the noise suppression algorithms. A robust VAD is crucial for improving the performance of the noise suppression algorithms [17], such as Wiener filtering and spectral subtraction.

In this thesis, we first justify the advantages of the Mel-energy features via the Bayes hypothesis analysis instead of the psychophysical conjectures in the existing literature [3]. Then a new robust VAD algorithm is proposed, which is based on the Mel-energy features and the adaptive threshold detection. Such a dynamical threshold can be derived from the SNR estimates in [18]. The outcomes of our VAD algorithm can be utilized for the silence deletion. Many simulations are performed to compare our proposed VAD algorithm with the existing VAD techniques, namely the VAD method using the likelihood ratio test with Fourier spectral energy features (SM VAD) in [12] and the enhanced time frequency based robust word boundary detection algorithm (ETF VAD) [3].

The rest of this thesis is organized as follows. The time-frequency features of the speech signals are studied and analyzed in Chapter 2. In Chapter 3, we introduce the new robust VAD scheme, using the Mel-energy features and the SNR-based adaptive

threshold. The speech compression application is presented in Chapter 4. The simulation results to demonstrate the effectiveness of our proposed VAD algorithm will be presented and the concluding remarks will be finally drawn in Chapter 5.

2. TIME-FREQUENCY REPRESENTATION AND ANALYSIS

The features play a key role in all voice activity detectors. The ambiguity due to the unreliable features, especially in the conditions of low SNRs and/or nonstationary background noises, will cause the misdetection very often [1]. According to a few simulations in [3], robust speech detection can be achieved using the Mel-spectral features compared to the Fourier spectral features. The existing literature provides the explanation simply based on the auditory psychophysics that the human ears perceive acoustic waves along the nonlinear scale in the frequency domain, which forms the Mel filter bank [3]. In addition, the dimensionality reduction but not in the tradeoff of performance can be achieved for speech detection and recognition using the Mel-spectral features. In the subsequent sections, we formulate the Fourier spectral and the Mel-spectral features for VAD. Then, we compare the VAD performances using these two features via the Bayes hypothesis testing analysis accordingly to show the effectiveness of the Mel-spectral features.

2.1. Fourier Spectral Features

Fourier spectral features are obtained from the short-time Fourier transform. A primary Fourier spectral feature, *short-time Fourier energy* $\left|x_{freq}[n,k]\right|^2$, is defined as [19]:

$$\left|x_{freq}[n,k]\right|^2 \equiv \left|\sum_{m=0}^{N-1} x(n-m)w(m)\exp\left(-j\frac{2\pi km}{N}\right)\right|^2, \quad (1)$$

where $x(n)$ is the discrete-time speech signal, $w(m)$ is the window sequence and N is the window size. According to Eq. (1), it is noted that $\left|x_{freq}[n,k]\right|^2$ is a double-indexed function with time index n and frequency index k. Usually, the *short-time framed Fourier*

energy $E_{FT}[n',k]$ will be collected for $n = 0, \Delta n, 2\Delta n, \cdots, n'\Delta n, \cdots$, and $n' \in \mathsf{Z}^+ \cup \{0\}$ such that

$$E_{FT}[n',k] \equiv \left| x_{freq}[n'\Delta n, k] \right|^2, \text{ for } n' \in \mathsf{Z}^+ \cup \{0\}, \tag{2}$$

where $\Delta n > 0$ is the *frame advance step size*.

2.2. Mel Spectral Features

The Mel-spectral features can be acquired through the weighted Fourier spectral features via the Mel filter bank which is a uniformly spaced filter bank on a nonlinearly wrapped frequency scale, known as the Mel-scale, as illustrated in Figure 1.

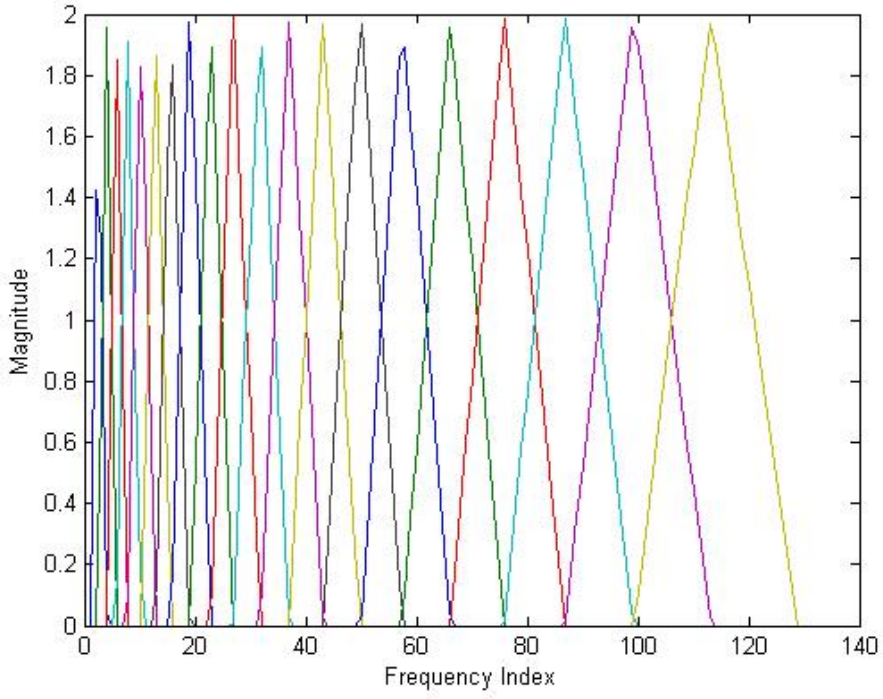


Figure 1. A Mel filter bank composed of the triangular band-pass filters each with a bandwidth and the spacing in accordance with the Mel-scale in the frequency domain.

The relationship between the Mel-scale frequency f_{mel} and the conventional frequency f_{con} (in Hz) is given by [3]:

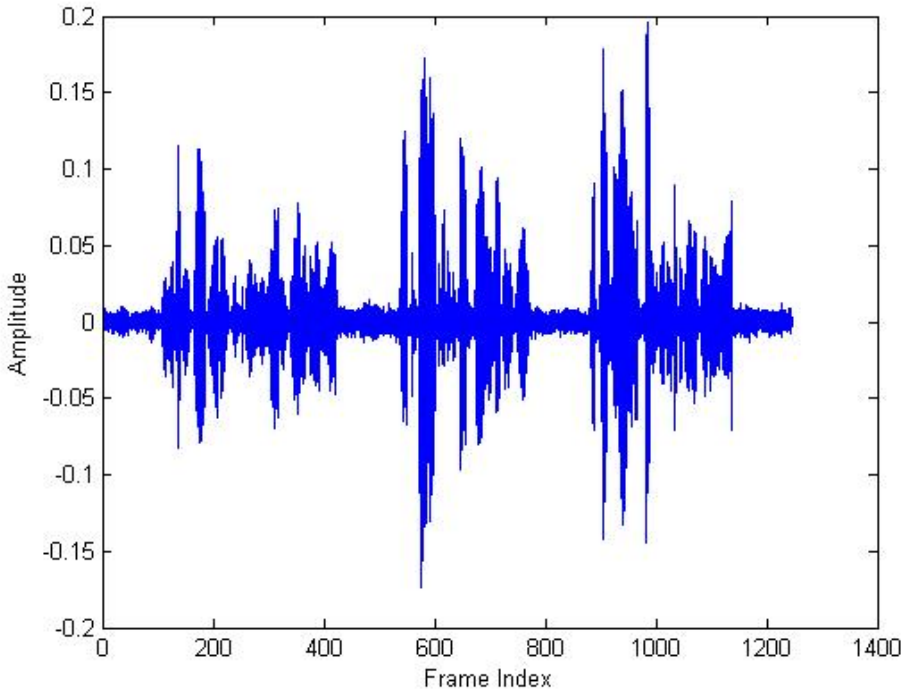
$$f_{mel} = 2595 \log\left(1 + \frac{f_{con}}{700}\right). \tag{3}$$

Without loss of generality, we choose a set of 20-band Mel filters throughout this thesis as illustrated in Figure 1. As depicted in Figure 1, the squared magnitude response of the i^{th} Mel filter, $\left|H_{mel}(i,k)\right|^2,\ 0 \le k \le N-1, 1 \le i \le 20,$ specifies the individual weighting factor for the k^{th} frequency component of the Fourier spectra [3]. According to Eqs. (2) and (3), the *short-time framed Mel-energy* $E_{mel}(n',i)$ is given by

$$E_{mel}(n',i) = \sum_{k=0}^{N-1} E_{FT}[n',k]\left|H_{mel}(i,k)\right|^2, \quad 1 \le i \le 20, \tag{4}$$

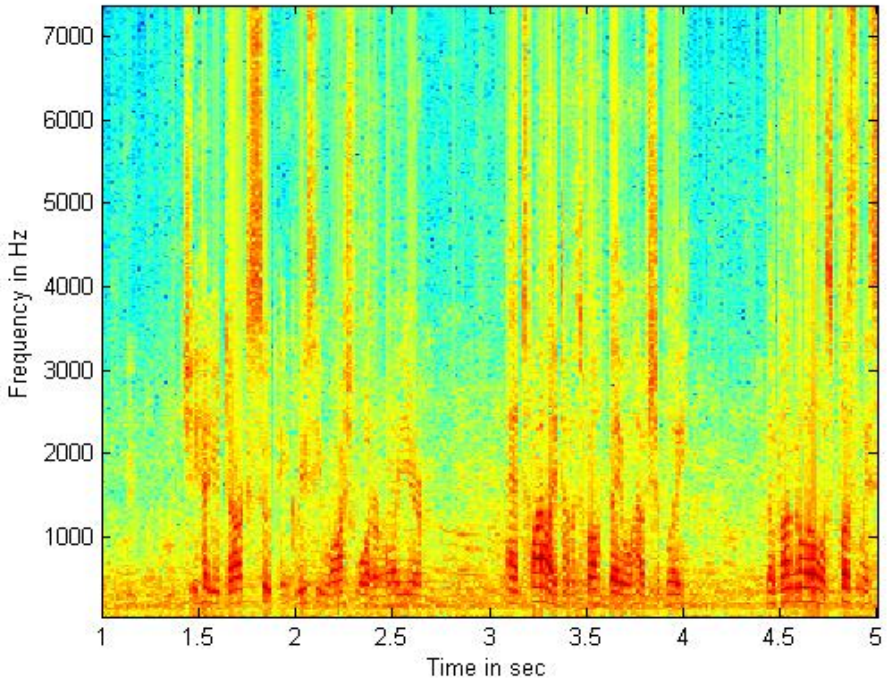
where, i is the Mel filter index. It is noted that the frequency dimensionality of $E_{mel}(n',i)$ is reduced from N of $E_{FT}[n',k]$ to 20.

Finally, Figure 2 illustrates the spectrograms for the Fourier spectral features and Mel-spectral features respectively. The spectrogram is a three-dimensional plot of the energy

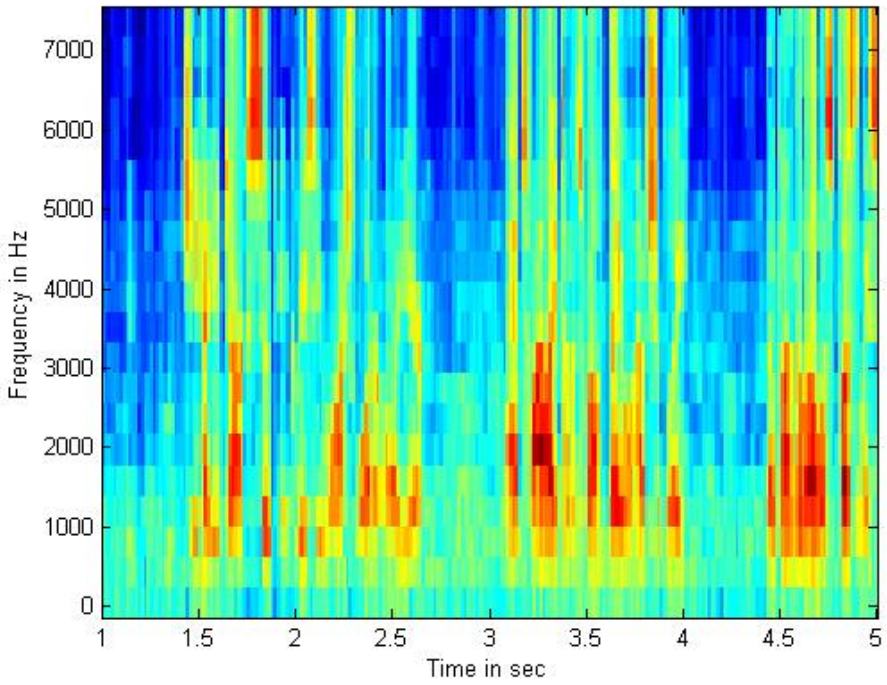


(a)

Figure 2. (a) Input speech signal, (b) Spectrogram for Fourier spectral features and (c) Spectrogram for Mel-spectral features (fig. cont'd.).


7000
6000
5000
4000
3000
2000
1000
Frequency in Hz
1
1.5
2
2.5
3
3.5
4
4.5
5
Time in sec

(b)


7000
6000
5000
4000
3000
2000
1000
0
Frequency in Hz
1
1.5
2
2.5
3
3.5
4
4.5
5
Time in sec

(c)

for the temporal frequency contents of a signal as it changes over time. The vertical axis represents the frequencies range from 0 to 8000 Hz, the horizontal axis indicates the incremental time towards the right, and the colors specify the most important acoustic peaks for a given time frame. In a decreasing order, red, orange, yellow, green, cyan, blue, magenta, gray and white represent the energies from the highest to the lowest, accordingly.

According to Figure 2, it can be observed that the distinction between speech and noise intervals using Mel-spectral features is much clearer than those using Fourier spectral features. Comparative analysis and simulations will be presented in the next section to further illustrate the superiority of the Mel-spectral features.

2.3. Comparative Studies between Fourier and Mel Spectral Features

In this section, we will provide the simulation results for the speech/noise classification to justify the advantage of the Mel-energy features over the Fourier energy features extracted on frame-by-frame basis. We establish an optimal Bayes classifier in [20] to evaluate the effectiveness of the extracted features under the assumption that the entire feature vectors are drawn from the multi-dimensional Gaussian processes. The general framework of the Bayes classifiers is shown in Figure 3. Provided speech data, the framed Fourier and Mel-energies are acquired to establish the corresponding Bayes classifiers and the ground truth (the true speech/noise frame labels) are applied to determine the optimal threshold. Then the outcomes of each classifier will be compared with the ground truth. The Bayes classifiers can be constructed as follows. Each feature vector is of dimension d ($d=N$ for Fourier spectral features, $d=20$ for Mel-spectral features). For a feature vector in an arbitrary frame, let us denote it as $\vec{X} \in \mathsf{R}^d$. Thus, the speech/noise frame classification becomes a binary Bayes hypothesis test, where

$H_s : \vec{X} \in \omega_s$ ($\vec{X}$ is extracted from a frame in the presence of both speech and noise) and $H_n : \vec{X} \in \omega_n$ ($\vec{X}$ is extracted from a frame in the presence of noise only) are the two corresponding values.

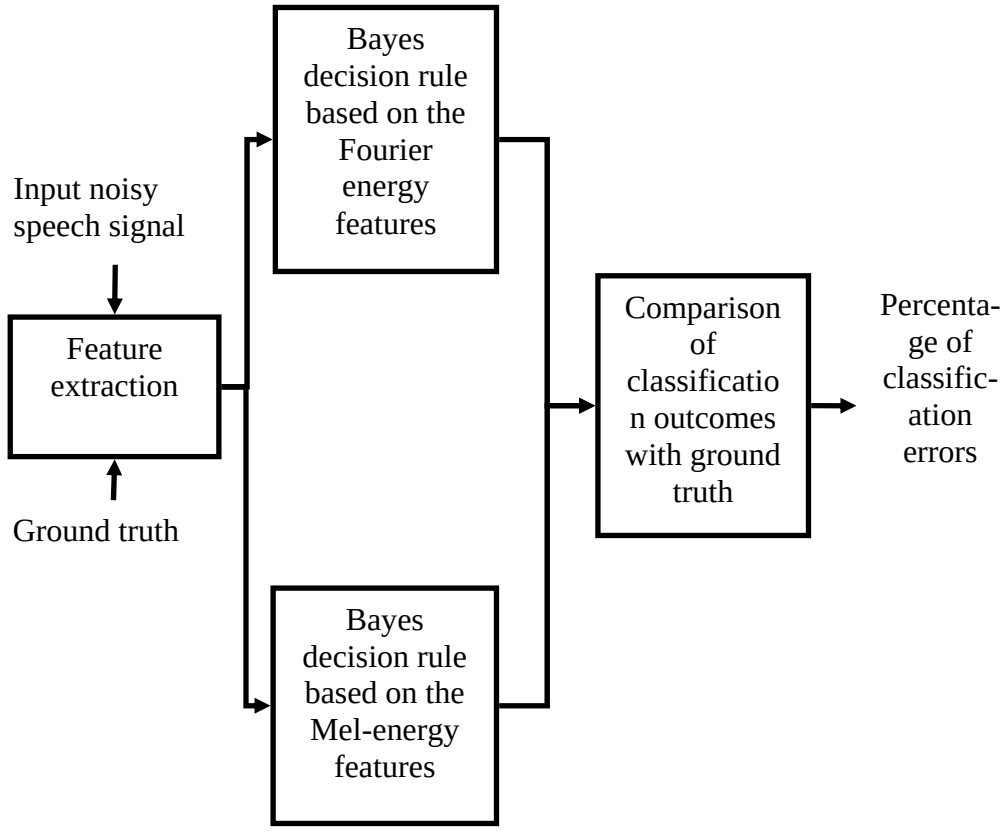


Figure 3. Comparison of different features using Bayes classifiers.

Let the two *a priori* probabilities be $P(\omega_s)$ and $P(\omega_n)$ respectively. Then, the *a posteriori* probabilities are given by [20]:

$$P(\omega_s \mid \vec{X}) = \frac{P(\vec{X} \mid \omega_s)P(\omega_s)}{P(\vec{X})}$$
$$P(\omega_n \mid \vec{X}) = \frac{P(\vec{X} \mid \omega_n)P(\omega_n)}{P(\vec{X})} \quad , \tag{5}$$

where $P(\vec{X}) = P(\vec{X} \mid \omega_s)P(\omega_s) + P(\vec{X} \mid \omega_n)P(\omega_n)$ is a common factor associated with the nonparametric probability.

The conditional probabilities are given by

$$P(\vec{X} \mid \omega_s) = \frac{1}{\sqrt{(2\pi)^d \left|\tilde{\tilde{\Sigma}}_s\right|}} \exp\left[-\frac{1}{2}\left(\vec{X}_s - \vec{\mu}_s\right)^T \tilde{\tilde{\Sigma}}_s^{-1}\left(\vec{X}_s - \vec{\mu}_s\right)\right]$$
$$P(\vec{X} \mid \omega_n) = \frac{1}{\sqrt{(2\pi)^d \left|\tilde{\tilde{\Sigma}}_n\right|}} \exp\left[-\frac{1}{2}\left(\vec{X}_n - \vec{\mu}_n\right)^T \tilde{\tilde{\Sigma}}_n^{-1}\left(\vec{X}_n - \vec{\mu}_n\right)\right], \quad (6)$$

where $\vec{\mu}_s$, $\vec{\mu}_n$ are the mean vectors and $\tilde{\tilde{\Sigma}}_s$, $\tilde{\tilde{\Sigma}}_n$ are the covariance matrices associated with the extracted feature vectors $\vec{X}$ for Hypotheses H_s and H_n respectively. According to Eq. (5), the Bayes classifier depends on $P(\vec{X} \mid \omega_s), P(\vec{X} \mid \omega_n)$, $P(\omega_s)$ and $P(\omega_n)$ only. The ground truth can be utilized to determine $P(\omega_s)$ and $P(\omega_n)$. In addition, the ground truth can also be utilized to determine the logarithms of the conditional probabilities as well, such that

$$\log\left[P(\vec{X} \mid \omega_s)\right] = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\left(\left|\tilde{\tilde{\Sigma}}_s\right|\right) - \frac{1}{2}(\vec{X} - \vec{\mu}_s)^T \tilde{\tilde{\Sigma}}_s^{-1}(\vec{X} - \vec{\mu}_s); \quad (7)$$

$$\log\left[P(\vec{X} \mid \omega_n)\right] = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\left(\left|\tilde{\tilde{\Sigma}}_n\right|\right) - \frac{1}{2}(\vec{X} - \vec{\mu}_n)^T \tilde{\tilde{\Sigma}}_n^{-1}(\vec{X} - \vec{\mu}_n). \quad (8)$$

According to Eqs. (5)- (8), the Bayes decision rule is given by

$$\Gamma\left(\omega_s \mid \vec{X}\right) \underset{\vec{X} \in \omega_n}{\overset{\vec{X} \in \omega_s}{>}} \Gamma\left(\omega_n \mid \vec{X}\right), \quad (9)$$

where the discriminant functions are defined as

$$\Gamma\left(\omega_s \mid \vec{X}\right) \equiv \log\left[P(\omega_s)\right] - \frac{1}{2}\log\left(\left|\tilde{\tilde{\Sigma}}_s\right|\right) - \frac{1}{2}(\vec{X} - \vec{\mu}_s)^T \tilde{\tilde{\Sigma}}_s^{-1}(\vec{X} - \vec{\mu}_s)$$
$$\Gamma\left(\omega_n \mid \vec{X}\right) \equiv \log\left[P(\omega_n)\right] - \frac{1}{2}\log\left(\left|\tilde{\tilde{\Sigma}}_n\right|\right) - \frac{1}{2}(\vec{X} - \vec{\mu}_n)^T \tilde{\tilde{\Sigma}}_n^{-1}(\vec{X} - \vec{\mu}_n).$$

For quantificational convenience, we artificially add the clean speech with noise (SNR = 5 dB) and carry out the Bayes classifiers as given by Eq. (9). The speech data are

randomly picked from the TIMIT database (three male and three female speakers) [21] while the white and the babble noises are taken from the NOISEX-92 database [22]. The ground truth comes from the speech frame labels specified in the TIMIT database. The outcomes of the Bayes classifiers in terms of the percentages of classification errors for both Mel-spectral features and Fourier spectral features are presented in the Table I. According to Table I, the Mel-spectral features lead to much better speech/noise frame detection performance than the Fourier spectral features, in both white and babble noises.

Table I. Percentages of classification errors using the Bayes classifiers for white noise and babble noise.

Types of Noise	Male speakers			Female speakers		
	Percentage of classification errors			Percentage of classification errors		
	Samples	Mel-spectral features	Fourier spectral features	Samples	Mel-spectral features	Fourier spectral features
White noise (SNR=5dB)	Speaker1	14.13%	31.43%	Speaker1	17.40%	61.86%
	Speaker2	23.85%	31.98%	Speaker2	24.92%	33.98%
	Speaker3	12.18%	42.26%	Speaker3	31.35%	34.52%
	Average	16.72%	35.23%	Average	24.55%	43.46%
Babble noise (SNR=5dB)	Speaker1	13.66%	44.00%	Speaker1	17.26%	11.12%
	Speaker2	19.85%	31.98%	Speaker2	23.68%	65.92%
	Speaker3	15.90%	42.26%	Speaker3	32.62%	45.57%
	Average	16.47%	39.42%	Average	24.52%	40.87%

3. SPEECH DETECTION USING MEL-SPECTRA

In this chapter, we describe the feature extraction as well as the adaptive threshold estimation and propose the new Mel-energy based adaptive threshold voice activity detector (ME VAD). Furthermore, we discuss a realignment mechanism to correct the classification errors generally encountered after the detection.

First, feature extraction is presented in Section 3.1. Then a threshold adaptation procedure is described in Section 3.2. Finally, the realignment mechanism is presented in Section 3.3.

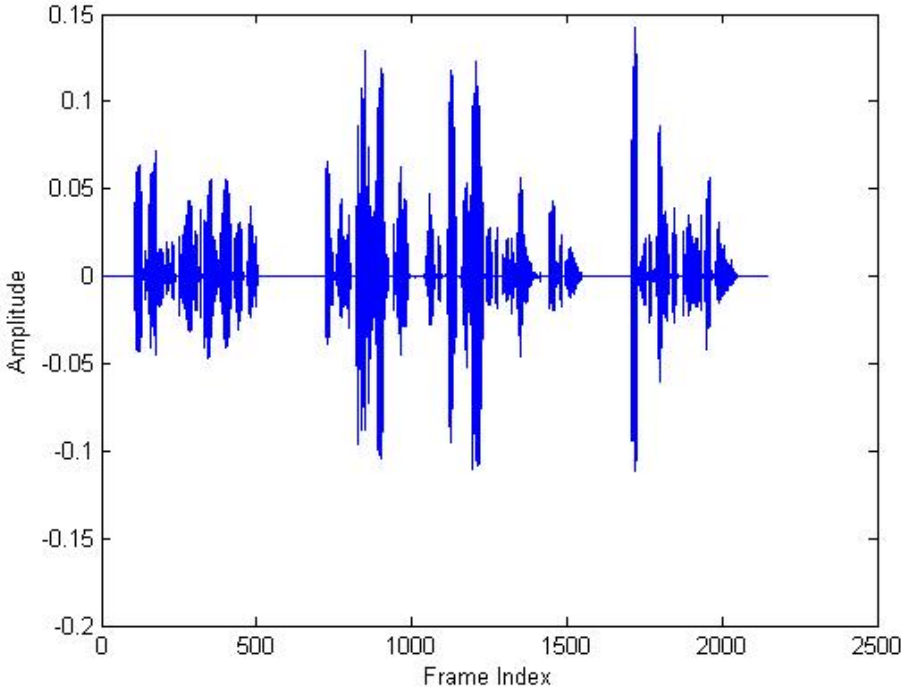
3.1. Mel-Energy Feature Extraction

Feature extraction is the crucial part of the voice activity detection and its methodology is described in this section. Consider a time-domain noisy speech signal x, which is divided into the overlapping frames of size 256 each. Furthermore, each frame is weighted by a 256-point Hamming window. Initially, the *short-time framed Fourier energy* $E_{FT}[n',k]$ of this signal is determined as described in Eq. (2) of Section 2.1. Next, the *short-time framed Mel-energy* $E_{mel}(n',i)$ for the i^{th} frequency band of the n'^{th} frame are calculated as describe in Eq. (4) of Section 2.2.

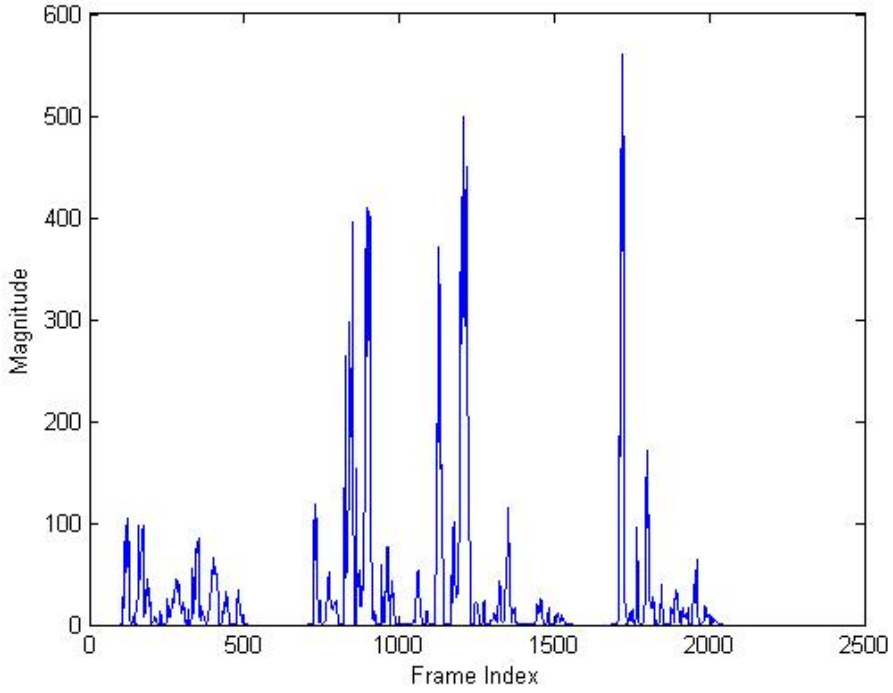
Finally, $E_{mel}(n',i)$, for i=1, …, 20, is summed over the 20 frequency bands to obtain the *robust energy indicator* $I(n')$:

$$I(n') = \sum_{i=1}^{20} E_{mel}(n',i) \ . \tag{10}$$

For example, the parameters $I(n')$ corresponding to a clean speech signal are depicted in Figure 4 and the parameters $I(n')$ for a speech signal corrupted by the babble noise with SNR= 5 dB are shown in Figure 5 respectively.

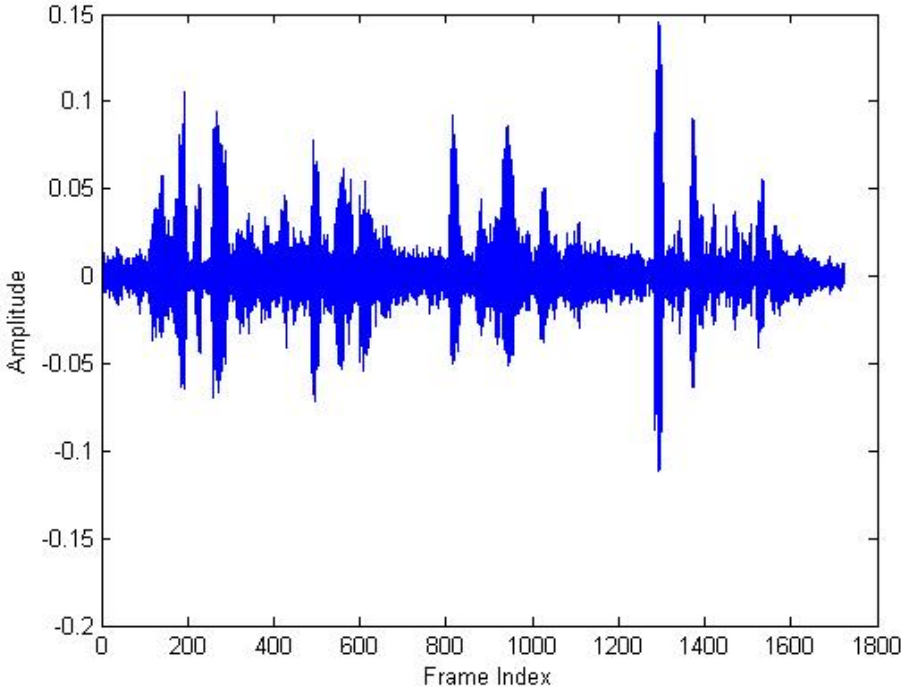


(a)

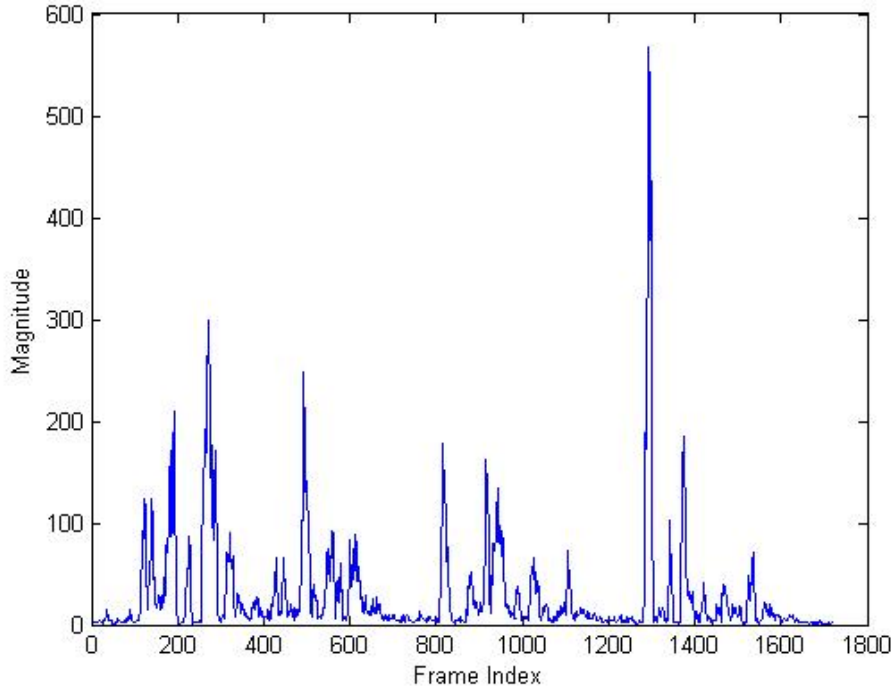


(b)

Figure 4. Robust energy indicator $I(n')$ and its corresponding time-domain speech signal: (a) clean speech waveform, (b) $I(n')$ for clean speech.



(a)



(b)

Figure 5. Robust energy indicator $I(n')$ and its corresponding time-domain speech signal: (a) waveform corrupted by babble noise (SNR=5dB), (b) $I(n')$ for corrupted speech.

3.2. Adaptive Threshold Detection

We propose a new "two-threshold" scheme for a better speech/noise classification. Since the first N_{noise} frames consist of noise only, we can determine the first threshold or the *a priori* threshold $\eta_{apr}(n')$, for the frame n', as

$$\eta_{apr}(n') \equiv \min\left[1.2E_n, \frac{E_{\max}(n') + E_n}{2}\right], \tag{11}$$

where $E_n \equiv \frac{1}{N_{noise}} \sum_{m=1}^{N_{noise}} I(m)$ denotes the average energy of first few noise frames; $E_{\max}(n') = \max_{1 \le m \le n'} \{ I(m) \}$. Accordingly, the *a priori* decision rule is given by

$$I(n') \underset{noise}{\overset{speech}{\gtrless}} \eta_{apr}(n'). \tag{12}$$

After the first N_{noise} noise frames, the classified noise energy indicator $N(n') = I(n')$, if the n'th frame is the noise frame index according to Eq. (12), it is stored in a noise energy buffer while $S(n') = I(n')$ is stored in a speech energy buffer if such a frame is the speech frame instead. Once the sufficient collection of both $N(n')$ and $S(n')$ is available, the SNR estimation can be achieved using [18].

Next, we like to introduce the procedure for estimating the *temporal SNR* and the *a posteriori threshold* $\eta_{aps}(n')$ as follows. The temporal speech energy estimate $\hat{S}(n')$ can be obtained as

$$\hat{S}(n') \equiv \sum_{m=n'-B_s+1}^{n'} S(m) - \sum_{m=n'-B_n+1}^{n'} N(m), \tag{13}$$

where B_S and B_n specify the speech buffer and noise buffer sizes, respectively. Thereby, the temporal SNR can be calculated as

$$SNR(n') \equiv 10\log_{10}\left[\frac{\hat{S}(n')}{\sum_{m=n'-B_n+1}^{n'} N(m)}\right]. \tag{14}$$

According to Eq. (14), we can estimate the *temporal noise energy estimate* $\hat{N}(n')$ as

$$\hat{N}(n') \equiv \frac{E_{\max}(N_{noise})}{1+\gamma\, SNR(n')}, \tag{15}$$

where γ is the control parameter. Thus, the *a posteriori threshold* $\eta_{aps}(n')$ can be determined as

$$\eta_{aps}(n') \equiv \min\left[1.2\hat{N}(n'), \frac{E_{\max}(n')+\hat{N}(n')}{2}\right]. \tag{16}$$

The *a posteriori* speech/noise frame classification can be achieved using

$$I(n') \underset{noise}{\overset{speech}{\gtrless}} \eta_{aps}(n'). \tag{17}$$

According to Eqs. (11) and (16), the two thresholds are dynamically adapted and therefore they can track the nonstationarity. Finally, these speech/noise frame labels are sent through a realignment mechanism to remove the sparse occurrence of speech/noise as described in the next section.

3.3. Realignment Mechanism

Maleh et al. concluded that the hangover schemes are not effective in correcting isolated VAD errors (i.e. a speech frame among a sequence of noise frames or vice versa) and hence they proposed an *isolated error correction mechanism* (IECM) [5]. Our realignment mechanism is simply based on the similar approach using the majority voting to forcefully re-assign the labels resulting from Eq. (17) consistent with the majority

every 5 to 7 successive frames. Finally the proposed algorithm can be summarized using the flowchart in Figure 6.

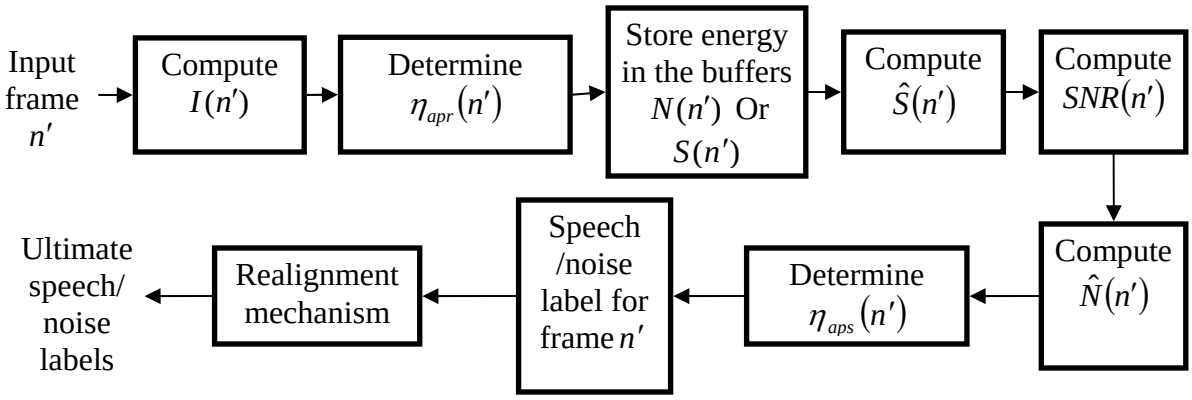


Figure 6. Flow chart of our proposed ME VAD algorithm.

4. SPEECH WAVEFORM COMPRESSION

The recently increasing demand in the voice communications, storage and voicemail applications is the driving force behind the improved cost-effective speech compression algorithms. Researchers are still interested in developing the efficient speech compression algorithms nowadays. Section 4.1 is focused on the application of the VAD outcomes as described in Chapter 3, to perform the speech waveform. In Section 4.2, the corresponding reconstruction procedure is also addressed.

4.1. Speech Compression Using Silence Deletion

This thesis focuses on the waveform compression using the silence deletion. A considerable portion, up to 60% of a two-way conversation, of normal speech belongs to silence. In practical applications, such as Global System for Mobile Communications (GSM), the silence detection and the comfort noise injection are applied for a higher coding efficiency [13]. In principle, the input signal is classified into the segments of active voice (speech) and inactive voice (silence or background noise) and the compression is performed by the deleting those inactive voice intervals. The silence deletion algorithm proposed by Loo et al. was capable of discarding up to 50% of the speech portion if the original speech signal is in a noise-free environment [14]. The speech waveform compression for the noisy speech signals is carried out with the help of the classified labels from the ME VAD scheme. In digital communications, this silence deletion leads to the dual-mode speech coding. The full-rate speech coder operates in the active speech mode, but a low-rate transmission is employed for the silence mode involving much fewer bits. Such a two-mode codec will lead to very efficient data communications.

Furthermore, for voice storage applications, the markings about the beginnings, ends and durations of the inactive intervals (silence) are stored as the additional information to the compressed speech waveforms. As illustrated in Figure 7, the inactive or silence intervals will be removed except that those markings will be recorded for the later speech reconstruction. The compressed signal consists of the segmented active speech waveforms together with the markings about the inactive (silence) intervals.

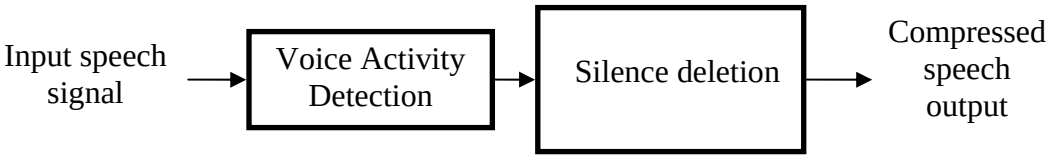


Figure 7. Speech waveform compression using silence deletion.

4.2. Speech Reconstruction

Speech reconstruction is undertaken when we insert the silence intervals back at the appropriate locations specified by the markings. Once the silence intervals are retrieved, the aforementioned dual-mode voice decoder in [14] or the simple zero-patching reconstruction scheme in [16] can be applied. In this thesis, the latter speech reconstruction approach is adopted as depicted in Figure 8.

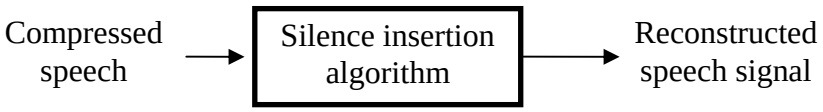
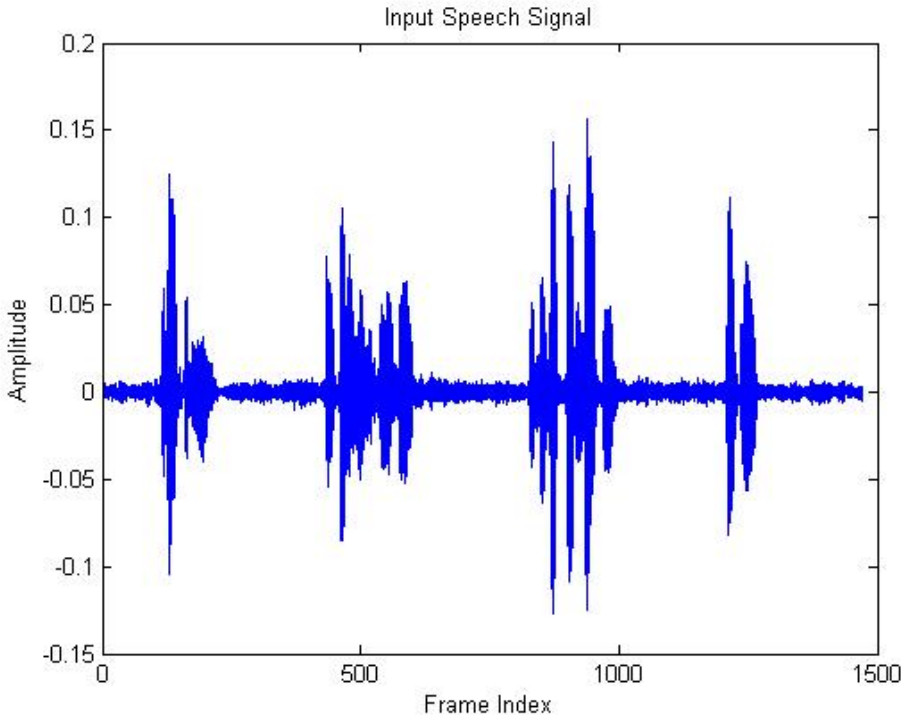


Figure 8. Speech reconstruction by silence insertion.

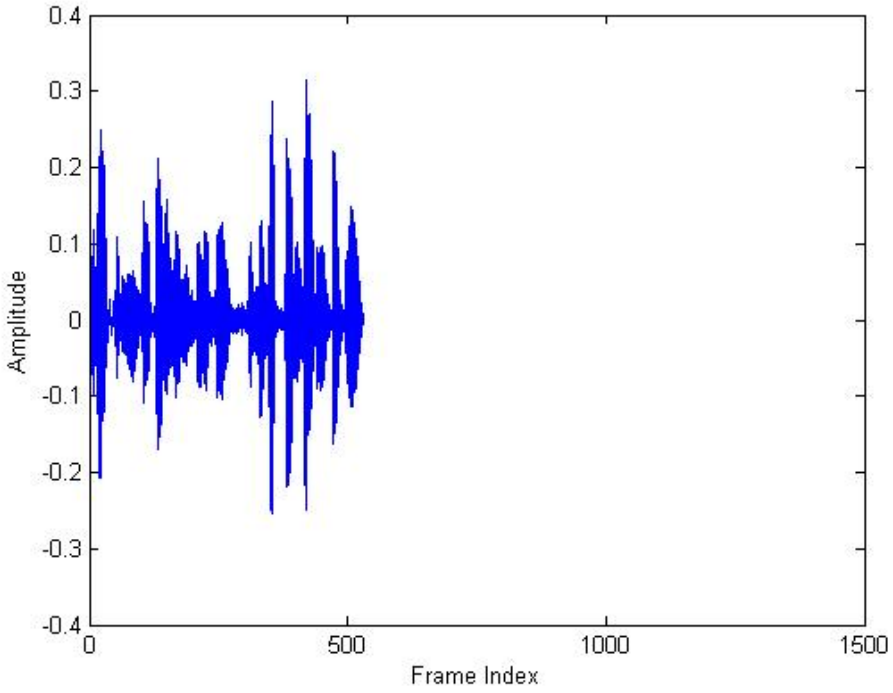
Finally, the procedure of the speech waveform compression and reconstruction is illustrated by the example in Figure 9. Figure 9 (a) depicts the original noisy speech signal waveform. Our ME VAD algorithm is employed to generate the speech/noise classification labels for this signal. Then the speech waveform is compressed using the

silence deletion based on the attained labels, which is shown in Figure 9 (b). Finally, Figure 9 (c) depicts the reconstructed speech signal from Figure 9 (b) after the silence insertion.

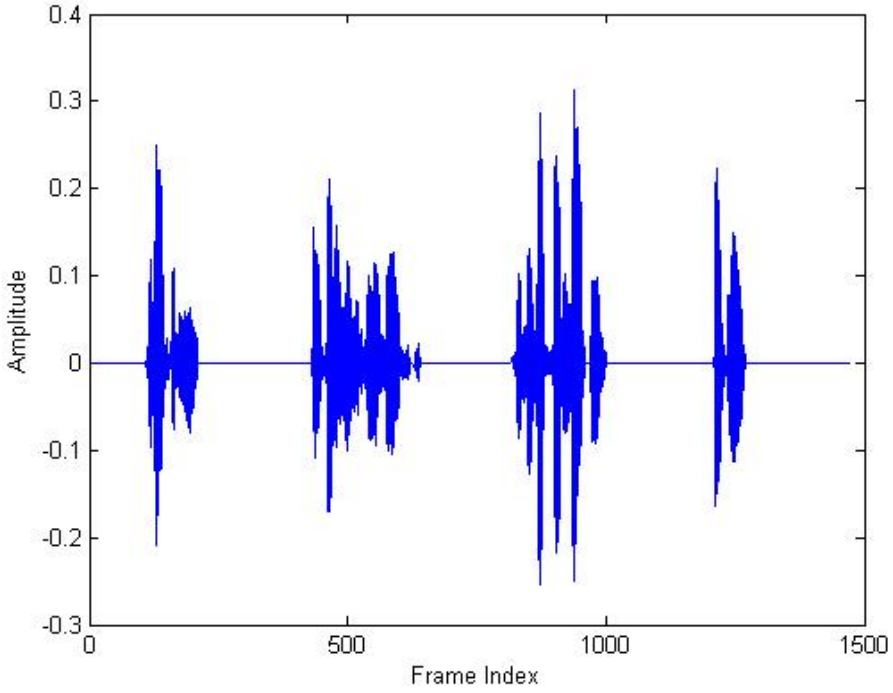


(a)

Figure 9. (a) The speech signal corrupted by babble noise (SNR=15dB), (b) compressed speech signal, and (c) reconstructed signal (fig. cont'd.).


0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
Amplitude
0
500
1000
1500
Frame Index

(b)


0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Amplitude
0
500
1000
1500
Frame Index

(c)

5. SIMULATION AND CONCLUSION

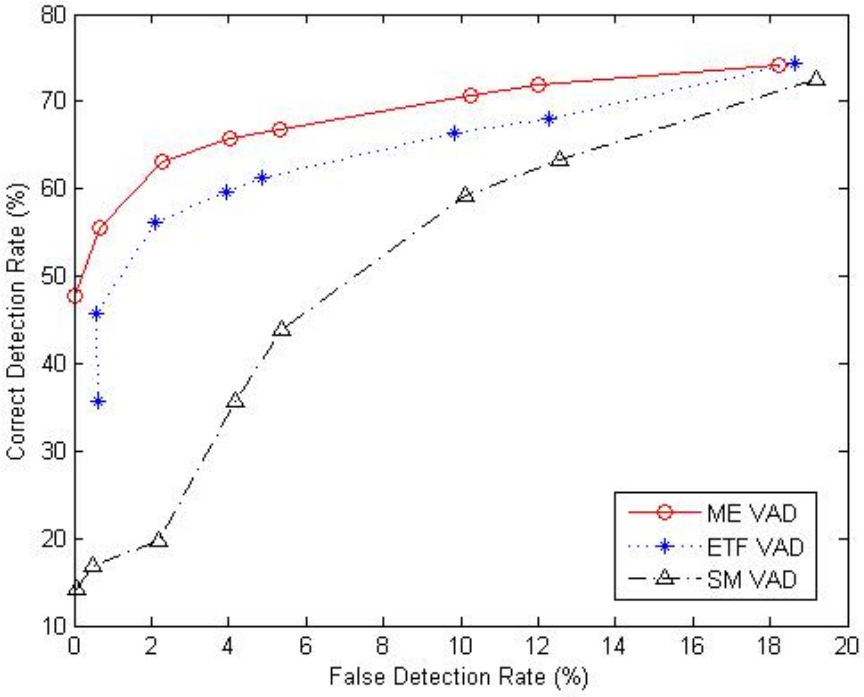
We have implemented the proposed *Mel-Energy Based Adaptive Threshold Voice Activity Detection Algorithm* (ME VAD) and compared the performance for different noise environments at various noise levels with the two major existing algorithms namely, *A Statistical Model Based Voice Activity Detection Algorithm* (SM VAD) in [11] and *An Enhanced Time Frequency Based Robust Word Boundary Detection Algorithm* (ETF VAD) in [3]. The SM VAD and ETF VAD algorithms were carried out and the corresponding parameters complied with [3, 11]. As stated in Section 5.1, the TIMIT database was chosen under testing and evaluation. As introduced in Section 5.2, the Receiver Operating Characteristics (ROC) curves for all algorithms were achieved to make the performance comparison. Finally, the different impacts of the VAD methods on the speech compression results are presented in Section 5.3.

5.1. Database

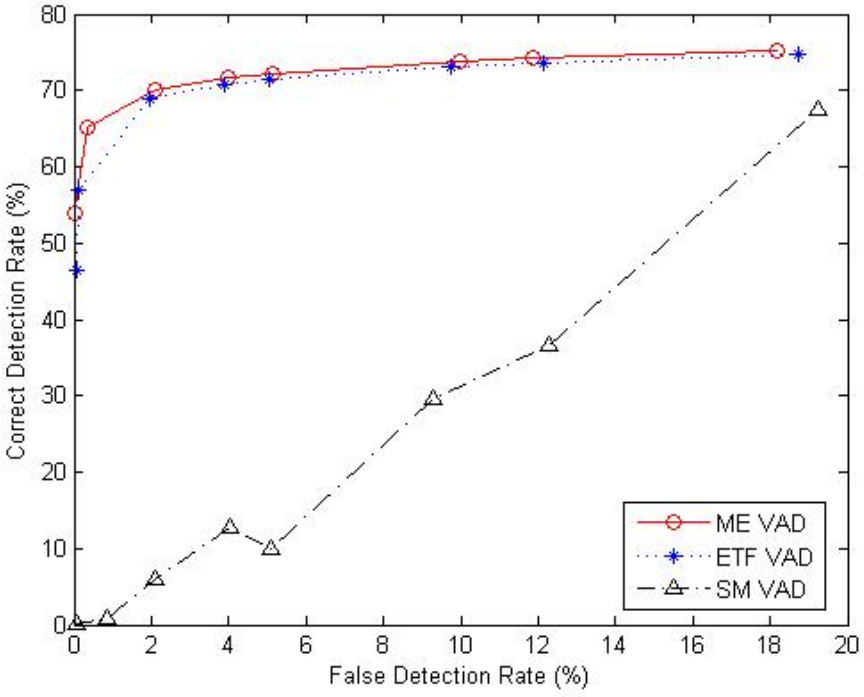
The signals to be experimented were drawn from the TIMIT databases [21], which were corrupted by the nonstationary babble and vehicular noises drawn from the NOISEX-92 database [22]. The sampling frequency was set as 16 kHz. Three male and three female speakers were selected from each of the three different US regions (New England, Southern, and Western) [21]. Thirty-six files were generated by the computer; each data file under test consisted of three different speech utterances from the same speaker concatenated together with the pauses in between; each speech data was added with the aforementioned noise samples for SNR=5 and SNR=15 dB respectively.

5.2. Receiver Operating Characteristics (ROC) of VAD

The Receiver Operating Characteristics (ROC) of a classifier specifies the performance as a trade-off between the selectivity and the sensitivity.



(a)



(b)

Figure 10. Receiver operating characteristics (ROC) for ME VAD, ETF VAD and SM VAD at SNR=5dB, for (a) babble noise (b) vehicular noise.

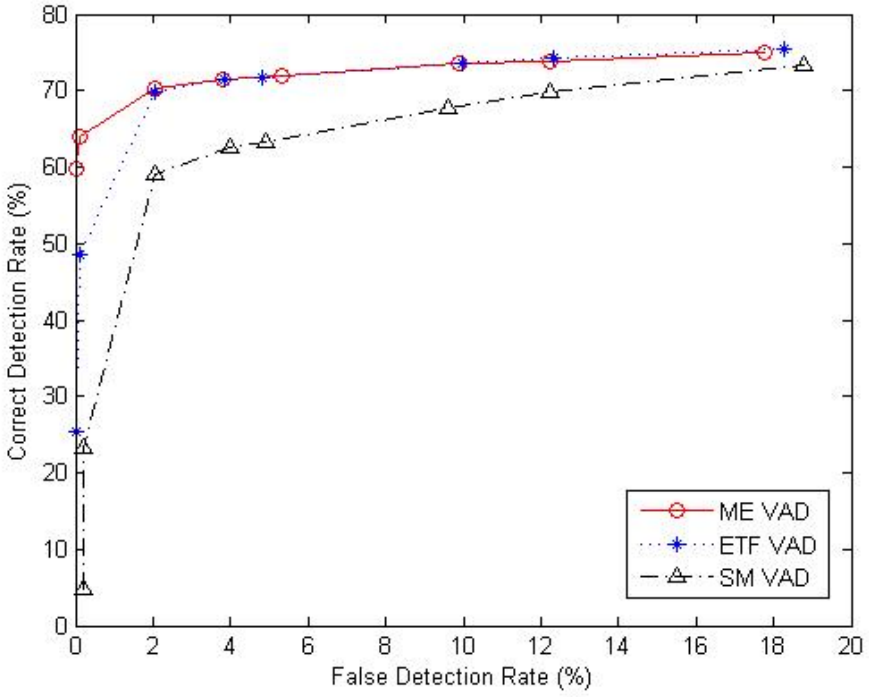
Table II. Correct detection rates for false detection rates=2%, 4% (SNR=5 dB)

SNR 5dB				
Types of Noise	Babble Noise		Vehicular Noise	
VAD algorithms	2%	4%	2%	4%
ME VAD	63.05%	65.68%	70.00%	71.60%
ETF VAD	56.23%	59.61%	68.90%	70.88%
SM VAD	19.72%	35.72%	5.94%	12.80%

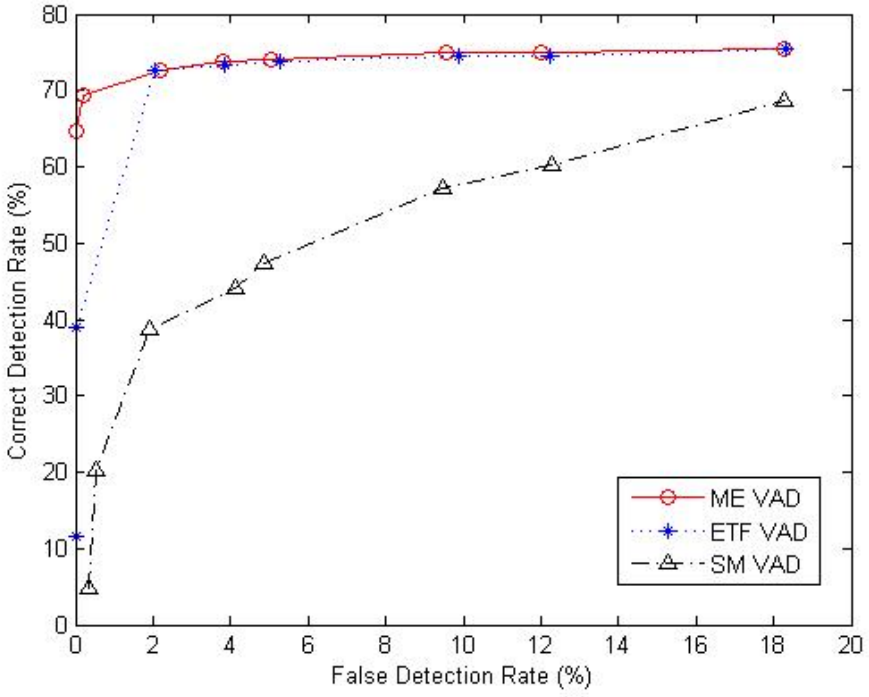
Typically a ROC curve measures the relationship between the percentage of the correctly classified frames (correct detection rate) versus the percentage of the incorrectly classified frames (false detection rate) [23]. The threshold parameter is varied and the corresponding correct detection rate and false alarm rate are observed and plotted accordingly. The higher the exponential increase in the curve the better is the classification accuracy of the VAD, with goal of the VAD being the minimize false detection rate for the highest correct detection rate possible.

The performance of the three algorithms for comparison was evaluated among different types of noise at SNR=5dB, 15dB and the ROC curves are shown in Figures 10 and 11 respectively. Note that the ME VAD performance is always better than the ETF VAD and SM VAD methods for the non-stationary babble and vehicular noises.

In addition, for a better comparison of algorithms we focus on small detection rate. For two normal conditions, at the small false detection rates of 2% and 4%, the correct detection rates were provided in Tables II, III. According to these tables, our ME VAD algorithm is the best for low false detection rates.



(a)



(b)

Figure 11. Receiver operating characteristics (ROC) for ME VAD, ETF VAD and SM VAD at SNR=15dB, for (a) babble noise (b) vehicular noise

Table III. Correct detection rates for false detection rates=2%, 4% (SNR=15 dB)

SNR 15dB				
Types of noise	Babble Noise		Vehicular Noise	
VAD algorithms	2%	4%	2%	4%
ME VAD	70.31%	71.49%	72.58%	73.79%
ETF VAD	69.76%	71.40%	72.72%	73.39%
SM VAD	59.03%	62.61%	38.69%	44.01%

5.3. Speech Compression

In addition, we developed a speech waveform compressor using the ME VAD outcomes. This compression algorithm identifies the beginnings and the ends of the pauses in a speech signal and then deletes the detected silence frames for the waveform compression. The speech reconstruction (decompression) can be performed by re-inserting the pauses at a desired noise level [18]. The performance of any compression scheme is generally measured in terms of *compression efficiency* and *play-back quality* [19]. However, the compression efficiency measure might be misleading since it dramatically varies among data and strictly depends on the durations of the silence periods. Therefore, in this thesis, we define a new compression efficiency measure C_d as

$$C_d \equiv \frac{\text{The number of detected noise frames}}{\text{The number of total frames}} \ (\%). \tag{18}$$

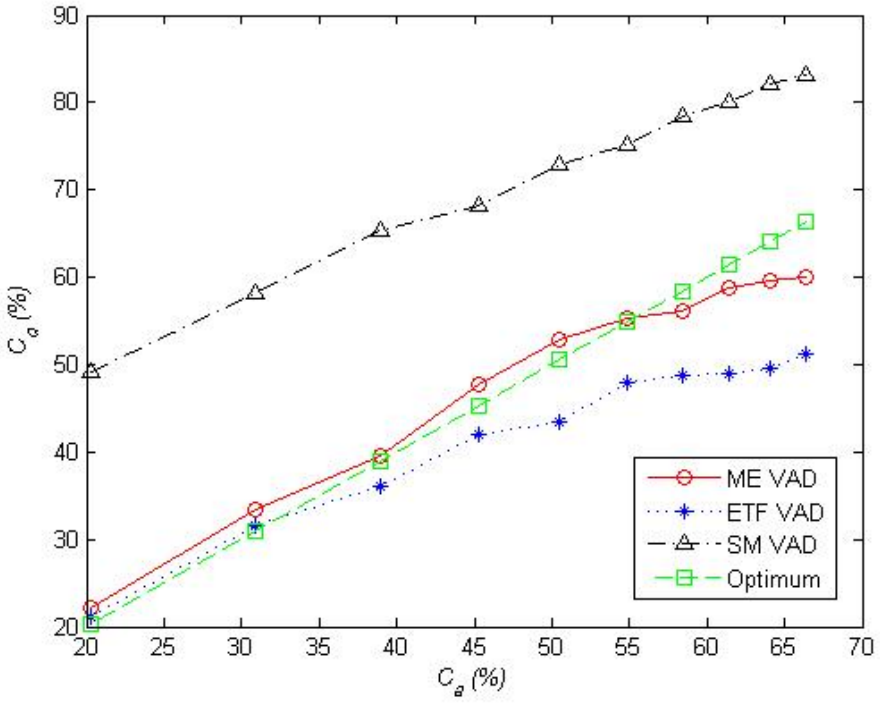
On the other hand, the nature of the original speech data can be characterized as the actual noise percentage measure C_a as

$$C_a \equiv \frac{\text{The number of actual noise frames}}{\text{The number of total frames}} \ (\%). \tag{19}$$

The compression efficiency curves for different speech data files (added with babble and vehicular noises at SNR=5 dB) as depicted in Fig. 12 illustrate the relationship between C_d and C_a from the ground truth. It is obvious that the optimal compression is achieved only when $C_d = C_a$. Thus, the closer the $C_d - C_a$ curve for each scheme to the straight line $C_a - C_a$ as illustrated as "Optimum" in Figure 12, the better the compression performance. According to Figure 12, our ME VAD based speech compression method significantly outperforms others.

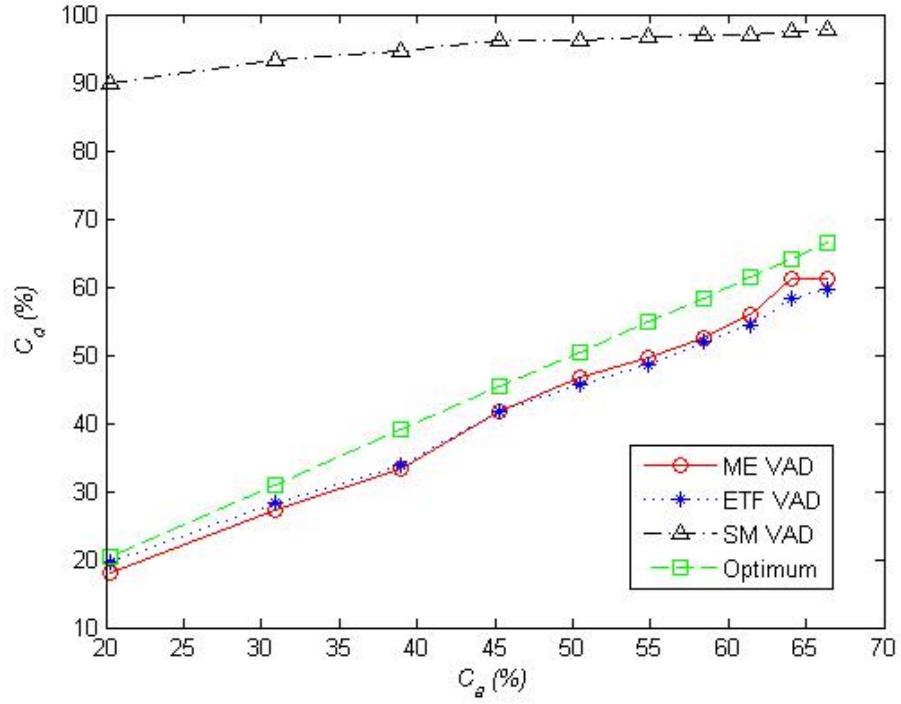
5.4. Conclusion

In this thesis, we investigate the advantage of the speech/noise detection using the Mel-spectral features over the Fourier spectral features via the Bayes hypothesis analysis. Then, we design a robust voice activity detection algorithm using the adaptive *a priori*



(a)

Figure 12. Compression efficiency curves of ME VAD, SM VAD and ETF VAD at SNR=5dB for (a) Babble noise and (b) Vehicular noise (fig. cont'd.).



(b)

and *a posteriori* thresholds incorporated with a realignment scheme. Moreover, we can extend this new voice activity detection method to establish a new speech waveform compressor for voice communications. Simulation results show that our new voice activity detection algorithm and speech compression scheme greatly outperform other existing methods, especially in the nonstationary noisy environments.

REFERENCES

[1] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109-118, February 2002.

[2] A. Fischer and V. Stahl, "On improvement measures for spectral subtraction applied to robust automatic speech recognition in car environments," *Proceedings of Workshop Robust Method Speech Recognition Adverse Conditions*, Tampere, Finland, pp.75-78, May 1999.

[3] C. T. Lin, J. Y. Lin and G.D. Wu, "A robust word boundary detection algorithm for variable noise-level environment in cars," *IEEE Transaction on Intelligent Transportation systems*, vol. 3, no. 1, pp. 89-101, March 2002.

[4] Stadermann, V. Stahl and G. Rose, "Voice activity detection in noisy environments," *Proceedings of European Conference on Speech Communication and Technology, Eurospeech*, pp. 1851-1854, 2001.

[5] K. El-Maleh and P. Kabal, "Comparison of voice activity detection for wireless personal communication systems," *Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering*, vol. 2, pp. 470-473, 1997.

[6] M. Hahn and C .K. Park, "An improved speech detection algorithm for isolated Korean utterances," Proceedings *of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 525-528, March 1992.

[7] L. R. Rabiner and M. R. Sambur, "Voice-unvoiced-silence detection using the itakura LPC distance measure," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 323-326, May 1977.

[8] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," Proceedings of the IEEE Region 10 Conference on Computer, Communication, Control, and Power Engineering, vol. 3, pp. 321-324, October 1993.

[9] S. H. Chen and J. F. Wang, "A wavelet-based voice activity detection algorithm in noisy environments," Proceedings of IEEE International Conference on Electronics, Circuits, and Systems, vol. 3, pp. 995-998, September 2002.

[10] A. Craciun and M. Gabrea, "Correlation coefficient-based voice activity detector algorithm," Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering, vol. 3, pp. 1789-1792, May 2004.

[11] J. Sohn, N. S. Kim and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1-3, January 1999.

[12] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 365-368, May 1998.

[13] A. Benyassine, E. Shlomot and H. Su, "ITU-T recommendation G.729 annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64-73, September 1997.

[14] M.C. Loo and R.W. Donaldson, "An Adaptive silence deletion algorithm for compression of telephone speech," *Proceedings of IEEE Pacific Rim Conference on Communications, Computer, and Signal Processing*, vol. 2, pp.701-705, 1997.

[15] D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd," The voice activity detector for the pan European digital cellular mobile telephone service," *Proceedings of IEEE International Conference of Acoustics, Speech, and Signal Processing*, vol. 1, pp. 369-372, May 1989.

[16] C. K. Gan and R.W. Donaldson, "Adaptive silence deletion for speech storage and voice mail applications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 6, pp.924-927, June 1988.

[17] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208-211, 1979.

[18] V. Saeed and B. Milner, "Noise compensation methods for hidden markov model speech recognition in adverse environments," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 11-21, January 1997.

[19] S. K. Mitra, *Digital Signal Processing: A Computer Based Approach*, 2nd Edition, Mc-Graw Hill, 2001.

[20] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd Edition, Wiley, 2002.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM, NIST, 1986.

[22] A. Varga, H. J .M. Steeneken, M. Tomlinson and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *Technical report*, Defense Evaluation and Research Agency (DERA), Speech Research Unit, Malvern, United Kingdom, 1992.

[23] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, vol. 1, Wiley, New York, 1968.

VITA

Waheeduddin Q. Syed was born on December 29, 1982, in Hyderabad City, India. He received his Bachelor of Technology in Electronics and Communications Engineering degree from Jawaharlal Nehru Technological University, Hyderabad, India, in May 2004. He is enrolled in the Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, Louisiana, since August 2004 to attend graduate school. His research interests are in communications and signal processing.