THE HONG KONG POLYUTECHNIC UNIVERSITY

DEPARTMENT OF ELECTRONIC AND INFORMATION ENGINEERING

VOICE ACTIVITY DETECTION FOR NIST SPEAKER RECOGNITION EVALUATIONS

By

YU HON-BILL

A thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Philosophy

July 2011

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

______________________ (Signed)

_____YU Hon-Bill_____ (Name of student)

Abstract

VOICE ACTIVITY DETECTION FOR NIST SPEAKER RECOGNITION EVALUATIONS

Since 2008, interview-style speech has become an important part of the NIST Speaker Recognition Evaluations (SREs). Unlike telephone speech, interview speech has a substantially lower signal-to-noise ratio, which necessitates robust voice activity detectors (VADs). This dissertation highlights the characteristics of interview speech files in NIST SREs and discusses the difficulties in performing speech/non-speech segmentation in these files. To overcome these difficulties, this dissertation proposes using speech enhancement techniques as a pre-processing step for enhancing the reliability of energy-based and statistical-model-based VADs. A decision strategy is also proposed to overcome the undesirable effects caused by impulsive signals and sinusoidal background signals. The proposed VAD is compared with five popular VADs.

1. *Average-Energy (AE)-Based VAD.* This is an energy-based VAD with decisions governed by the linear combination of average magnitude of background noises and signal peaks.

2. *Automatic Speech Recognition (ASR) Transcripts.* In this VAD, speech/non-speech decisions are based on the ASR transcripts provided by NIST.

3. *VAD in the ETSI-AMR Option 2 Coder.* This VAD is part of the Adaptive Multi-Rate (AMR) codec released by the European Telecommunication Standard Institute (ETSI).

4. *Statistical-Model (SM)-Based VAD.* This VAD assumes that the complex frequency components of signals and noises follow a Gaussian distribution

and uses likelihood-ratio tests in the frequency domain for speech/non-speech decisions.

5. *Gaussian-Mixture-Model (GMM)-Based VAD.* This is an extension of the statistical-model-based VAD, which considers the long-term temporal information and harmonic structure in noisy speech.

These five VADs have been evaluated on the NIST 2010 dataset. The comparison of VADs leads to seven findings:

1. Noise reduction is vital for VAD under extremely low SNR;

2. Removal of the sinusoidal background noise is of primary importance as this kind of background signal could lead to many false detection in AE-based VAD;

3. A reliable threshold strategy is required to address the impulsive signals;

4. ASR transcripts provided by NIST do not produce accurate speech and non-speech segmentations;

5. Spectral subtraction contributes to both AE- and SM-based VADs;

6. Spectral subtraction makes better use of background spectra than the likelihood-ratio tests in the SM-based VAD; and

7. The proposed SS+AE-VAD outperforms the SM-based VAD, the GMM-based VAD, the AMR speech coder, and the ASR transcripts provided by NIST SRE Workshop.

ACKNOWLEDGMENTS

I owe my deepest gratitude to various bodies from The Hong Kong Polytechnic University, where I have the opportunity to study with. I am heartily thankful to my supervisor Dr. Man-Wai Mak, Associate Professor of The Hong Kong Polytechnic University, for his encouragement, guidance and support that enable me to develop and understand any research work. I do appreciate his assistance in writing papers and this dissertation. I would also like to thank Prof. Helen Meng and Weiwu Jiang for their technical support.

It is my pleasure to thank those who have taught me during my postgraduate study. I am indebted to all the staff of the department of Electronic and Information Engineering and the clerical staff in the General Office, who have created a creative environment for me to study in.

Besides, I would like to acknowledge the Research Office of The Hong Kong Polytechnic University for its unfailing support over the past two years.

Finally, I offer my regards and blessings to my fiancee and my parents, especially my deceased father for his everlasting encouragement. Last but not least, I would like to honor the omnipresent God for giving me the strength to plod on despite my constitution wanting to give up and throw in the towel, thank you my dear Lord.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

NIST Speaker Recognition Evaluations (SREs)[1] are biannual activities in the speaker recognition community. The evaluations aim at calibrating the capabilities of text-independent speaker recognition technology and measuring the performance of the state-of-the-art approaches to speaker recognition. This dissertation develops voice activity detection (VAD)[2] techniques for extracting speech segments from the interview speech files of recent NIST SREs.

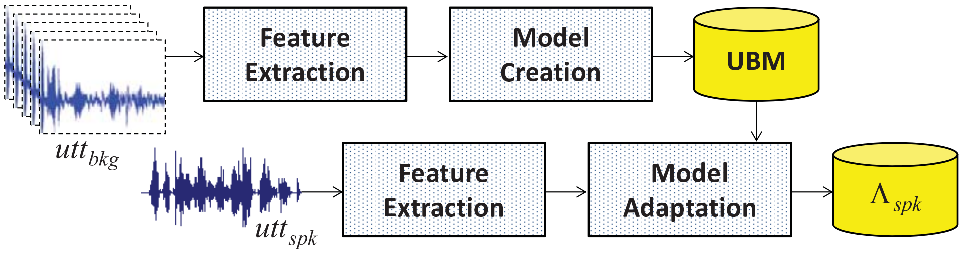


Figure 1.1: The enrollment phase in speaker verification.

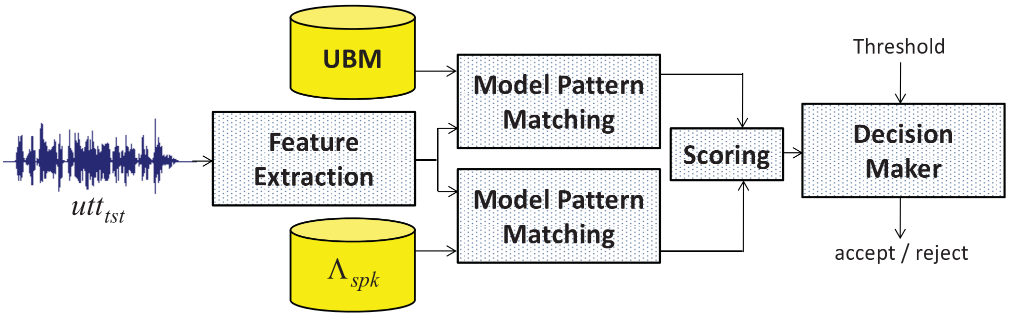


Figure 1.2: The verification phase in speaker verification.

[1] http://www.itl.nist.gov/iad/mig/tests/sre

[2] In this dissertation, the abbreviation VAD stands for voice activity detection or voice activity detector depending on the context

1.1 Speaker Verification

In speaker verification, the identity of a claimant is verified based on his or her own voices [5]. The applications of speaker verification include transaction authentication, access control and information retrieval. For example, Banco Bradesco, a Brazil's private bank, uses Nuance's speaker verification solution to verify its 15 million customers over the phone [6]. In a similar example, ABN AMRO uses VoiceVault's speaker verification system in its telephone banking services [7]. More recently, NAP Personal Banking in Australia and T-mobile of Deutsche Telekom in Netherlands provide voice authentication for their customers [8, 9].

The two main stages in speaker verification are enrollment and verification [10], as illustrated in Fig. 1.1 and Fig. 1.2, respectively. The utterances from background speakers (utt_{bkg}), target speakers (utt_{spk}) and test speakers (utt_{tst}) are transformed into speaker-specific feature vectors, e.g., MFCC [11, 12]. Before enrollment and verification, a universal background model (UBM) – which is typically a Gaussian mixture model with a large number of mixture components – is firstly trained by using the speech of a large number of background speakers to represent the characteristics of the general population. During the enrollment stage, a target-speaker model (Λ_{spk}) is created by adapting the UBM [13] using the feature vectors obtained from the utterances of the target-speaker. During verification, the feature vectors (X_{tst}) extracted from a test speaker (an unknown person) are compared against the background and target-speaker models to give a similarity score, $S(X_{tst}) = \log p(X_{tst}|\Lambda_{spk}) - \log p(X_{tst}|\Lambda_{ubm})$. The score is then compared against a decision threshold to determine whether the test speaker is a genuine speaker or an impostor.

The utterances can be collected from the conversations of speakers in telephone calls, meetings, and interviews. Different transducers and recording environments could affect the performance of speaker recognition systems. In recent NIST SREs, more emphasis has been put on interview speech.

1.2 Importance of VAD in Speaker Verification

NIST SREs have been focusing on text-independent speaker verification over telephone channels since 1996. In recent years, NIST introduces interview-style speech into the evaluations. For example, the speech files in NIST 2008 SRE contain conversation segments of approximately five minutes of telephone speech and three minutes of interview speech, and the speech files in NIST 2010 SRE contain interview recordings with duration ranging from three to fifteen minutes. In each speech file, about half of the conversation contains speech, and the remaining part contains pauses or silence intervals. The inclusion of non-speech intervals in the speech files necessitates voice activity detection (VAD) because these intervals do not contain any speaker information. VAD is a very useful technique for improving the performance of speaker recognition systems working in such scenario. In particular, VAD can be used to identify speech segments prior to the feature extraction process.

The determination of speech segments is critical for speaker verification because misclassifying non-speech segments as speech segments means that incorrect information will be used for speaker modeling and for decision making. Speech/Non-speech detection can be formulated as a statistical hypothesis problem aimed at determining to which class a given speech segment belongs. However, a high level of background noise can cause numerous detection errors. This is because the noise partly or completely masks the speech signal [14]. A robust decision rule that works under noisy conditions is therefore essential. Most of the existing VAD algorithms are effective under clean acoustic environments, but they could fail badly under adverse acoustic conditions [15].

1.3 Contributions and Organization of the Dissertation

This dissertation proposes using spectral subtraction to remove the background noise as much as possible before applying the energy-based VAD and statistical-model-based VAD. The advantage of using spectral subtraction is that it allows us to introduce a nonlinear filtering operation on the noisy signal, which has the effect of emphasizing the speech signal at high SNR regions and suppressing the background noise (to almost zero) in low SNR regions. This nonlinear operation effectively boost the SNR of the whole speech file, which makes the subsequent energy-based VAD uncomplicated. Experimental evaluations suggest that the VAD is particularly suitable for extracting speech segments from the interview-speech files of NIST SREs.

This work has the following contributions to VAD aand speaker verification.

1. Compare the effectiveness of the state-of-the-art VADs – including the statistical-model(SM)-based VAD and the Gaussian-mixture-model(GMM)-based VAD – for speaker verification;

2. Develop a novel technique to extract reliable speech and non-speech segments for training the GMM models in GMM-VAD;

3. Develop a threshold determination strategy that can largely eliminate the influence of impulsive signal; and

4. Discover that spectral subtraction is a simple and good preprocessor for the energy-based VAD and SM-VAD.

Chapter 3 highlights the special characteristics of interview speech in recent NIST SREs and demonstrates how these characteristics cause difficulties in extracting the speech segments accurately. Then, Chapter 4 describes different VADs and argues that spectral subtraction is an essential step in overcoming the difficulties. Further evidences are then reported in Chapter 6 where the proposed VAD shows promising results in NIST 2008 and 2010 SREs.

Chapter 2

LITERATURE REVIEW ON VAD

VAD is an essential part of many speech processing and communication systems. For example, VAD helps enhancing system capacity and reducing power consumption of portable communication devices via discontinuous transmission (DTX) of coded speech [16]. Many VAD methods have been used in the past. Some popular ones use periodicity measure [17], zero-crossing rate [18], pitch [19], energy [20], spectrum analysis [21], higher order statistics in the LPC residual domain [22], or combinations of different features [23].

During the last decade different VAD methods have been applied to real-time speech transmission on the Internet [24], mobile communication services [16] and noise reduction for digital hearing aid devices [25]. These methods can be generally divided into three categories and they will be briefly explained in this chapter.

2.1 Conventional Time- or Frequency-Based Methods

Early VADs extract parameters such as LPC distance [26], energy levels, and zero crossing rates [18, 20, 27] from speech signals and compare these parameters with a set of thresholds for detecting the speech regions of an utterance. In 1993, an energy-based speech detector was proposed [28] and its algorithm can be described by the following equations:

$$E = \sum_{t=0}^{T-1} s^2(t) \quad \text{or} \quad E = \sum_{k=0}^{K-1} S_k^2, \tag{2.1}$$

$$E_p = 1.5E_d, \tag{2.2}$$

$$E_d^{new} = (1-p)E_d^{old} + pE, \tag{2.3}$$

where E is the frame energy computed either from signal samples $s(t)$ or from DFT coefficients S_k of a frame. This energy is compared with the threshold E_p derived from the background noise energy E_d, i.e., speech is detected if E is greater than E_p; otherwise, the background noise energy E_d is modified according to Eq. 2.3.

The zero-crossing rate of a signal is defined as the number of times the signal crosses the 'zero' line [29]. The number of zero crossings for noise is random and unpredictable; in contrast, the number of zero crossings for speech signals lies in a fixed range. Specifically, the number of zero crossings for a 10-ms frame lies between 5 and 15 [24]. This characteristic has been used to formulate decision rules [24] that are independent of energy and hence is able to detect the low-energy phonemes of a word or utterance. Typically, the following decision rule is used:

$$\text{Frame} = \begin{cases} \text{`Speech'} & \text{if } Z_l \leq N_{ZCR}(m) \leq Z_u, \\ \text{`Non-speech'} & \text{otherwise,} \end{cases} \tag{2.4}$$

where $N_{ZCR}(m)$ is the number of zero crossings detected at frame m, and Z_l and Z_u are respectively the lower and upper limits of the zero-crossings in typical speech frames.

The detection accuracy of these earlier methods, however, could degrade dramatically under adverse acoustic conditions.

2.2 VAD Methods in Speech Coders

Advanced speech coders typically use more sophisticated methods in their VAD than the conventional methods mentioned earlier. For example, the European Telecommunication Standard Institute (ETSI) has released two VAD options [30] for the Adaptive Multi-Rate (AMR) codec.

In Option 1 of the ETSI-AMR coder (AMR1) [1], speech is firstly passed through a filterbank and the signal level in each band is calculated. The SNR of these bands together with the output of a pitch detector, a tone detector and a complex-signal analysis module are used to make VAD decisions. Thus, the classification of voice activity depends on a mixture of acoustic information including pitch, tone, complex-signal correlation and the energy levels of 9 frequency bands.

In Option 2 of the coder (AMR2), the decision logic is based on the energy of 16 channels (frequency bands), background noise, channel SNR, frame SNR, and long-term SNR [31]. As an enhanced version of the original VAD, AMR2 takes advantage of speech encoder parameters and is more robust against environmental noise than AMR1 and G.729 [32].

One advantage of this coder is that the VAD decision threshold is adapted dynamically according to the acoustic environment, allowing on-line speech/non-speech detection under non-stationary acoustic environments.

2.3 Statistical-Model-Based Methods

More recently, research has focused on statistical-model-based VAD where individual frequency bins of speech are assumed to follow a parametric density function [4]. In this approach, VAD decisions are based on a Likelihood Ratio Test (LRT) where the geometric mean of the log-likelihood ratios of individual frequency bins are estimated from observed speech signals. The statistical model can be Gaussian [4]. However, it has been recently found that Laplacian and Gamma models are more appropriate for handling a wide variety of noise conditions [33]. Using an online version of the Kolmogorov-Smirnov test, the type of models can be selected adaptively for different noise types and SNRs [33]. Furthermore, Gaussian mixture models trained with clean speech and noise have been used to provide an appropriate decision rule for speech/non-speech detection [2, 32].

To improve the robustness of VAD under adverse acoustic environment, contextual information derived from multiple observations has been incorporated into the LRT (MO-LRT) [34].

2.4 Applications of VAD

VAD has been applied to many areas of speech processing, including coding, enhancement and recognition [14]. For speech coding, taking ETSI-AMR as an example, the coder works at eight different bit rates ranging from 12.2 kb/s to 4.75 kb/s [35]. The coder is equipped with a voice activity detector [36–38] that enables silence compression, facilitates channel interference reduction, and extends the battery life time for mobile communications.

VAD plays an important role in speech enhancement. The purpose of speech enhancement is to improve the speech quality under noisy environments. There are two main difficulties for designing a speech enhancement system. One is the lack of explicit statistical models for the speech and noise signals. Another one is non-stationary property of speech and possibly also the noise signals. In practice, the noise source is assumed to be additive and contains no correlation with the clean speech signals. Spectral subtraction has been proposed in 1979 for lowering the level of background noise [39]. This method is popular because of its simplicity and ease of implementation.

The quality of speech signal has a strong influence on the performance of speech recognition systems. To improve performance, VAD can be used to remove the non-speech frames from the speech signals, a technique commonly known as frame dropping [40].

2.5 VAD for NIST SRE

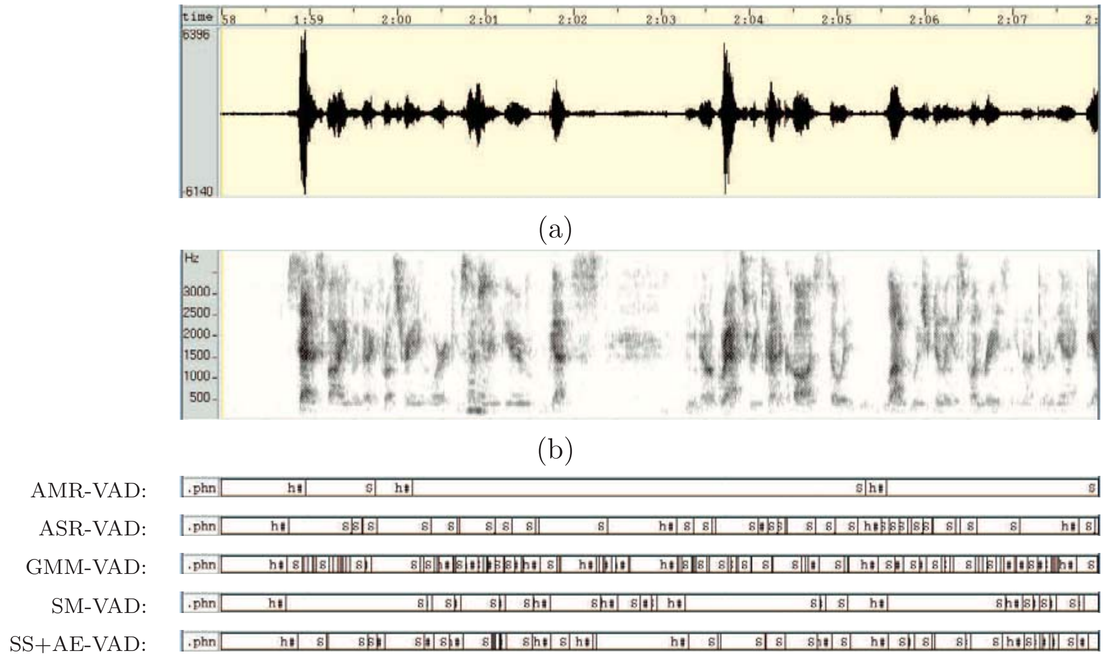
In recent NIST SREs, several sites provided the details of their VAD in the system descriptions. Typically, these systems use energy-based methods that estimate a file-dependent decision threshold according to the maximum energy level of the file [41]. Some sites used the periodicity of speech frames or the power of noise-removed speech frames to make speech/non-speech decisions [42, 43]. An alternative approach is to use the ASR transcripts supplied by NIST to remove the non-speech segments [44].

Chapter 3

CHARACTERISTICS OF INTERVIEW SPEECH IN NIST SRES

In early NIST SREs, researchers seldom pay attention to VAD. This is because the telephone speech files in early SREs have high signal-to-noise ratios (SNRs), making VAD a trivial task. The high SNR in telephone speech is resulted from the close proximity between speaker's mouth and the handset. In interview speech, however, different microphone types were used for recording. For example, twelve microphones were used in NIST 2008 SRE,[1] and in NIST 2010 SRE, the interviewees used different types of far-field microphones, such as lavaliere microphones, camcorders, and hanging microphones [45]. These microphones lead to four types of speech files. This chapter highlights the characteristics of these files and explains why these characteristics cause difficulty to VAD.

[1] Some of these microphones are of the same models, but they were placed at different positions with respect to the speakers.



(a)

(b)

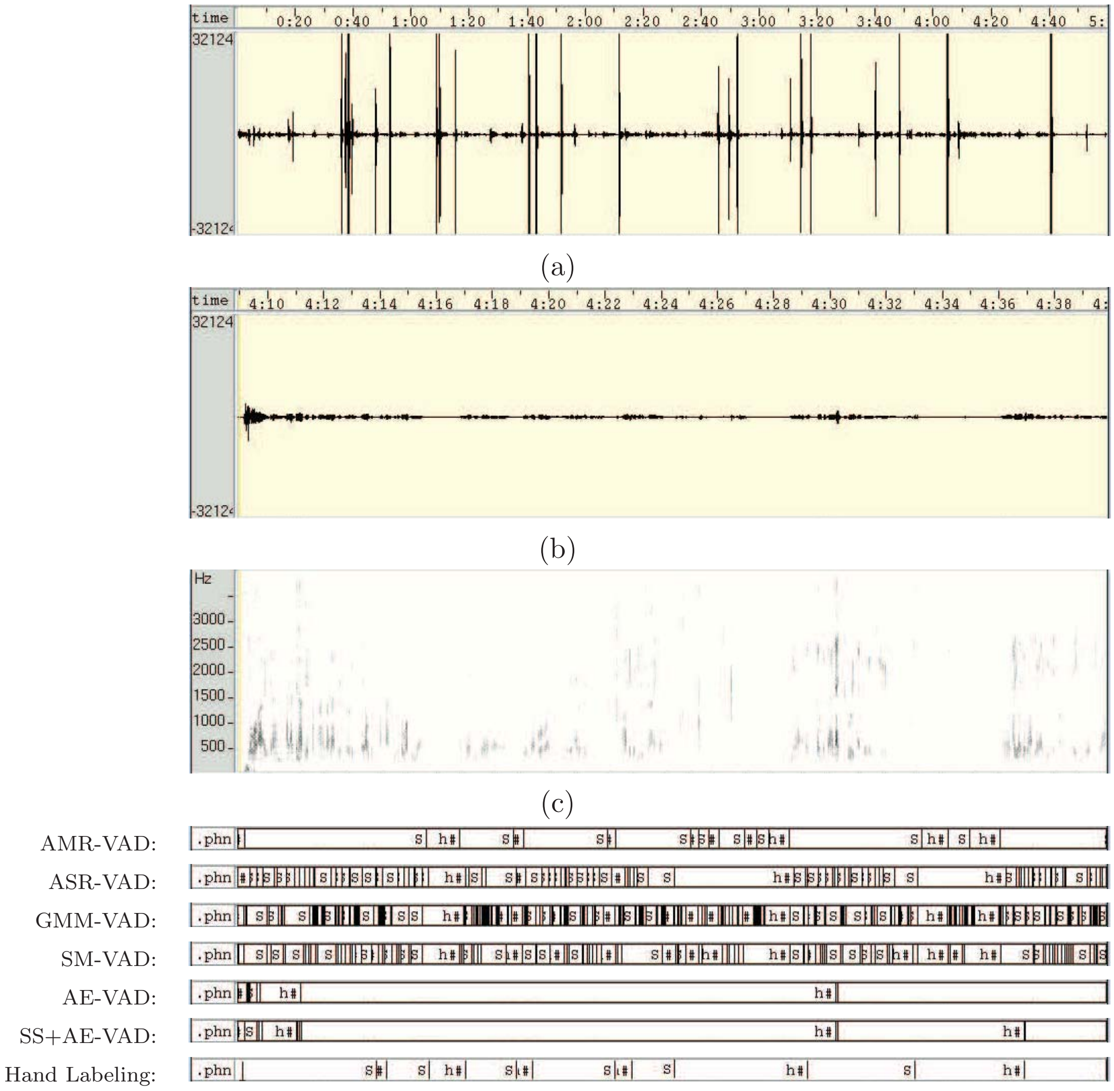
(c)

Figure 3.1: (a) A short segment of low-energy interview speech in NIST 2010 SRE with high-energy spikes. (b) The spectrogram of the same short segment. (c) Speech/non-speech decisions (`S` for speech and `h#` for silence) made by five different VADs, which are abbreviated in Table 6.1.

3.1 Impulsive Signals

Some files contain a number of spikes caused by plosive sounds or the speaker speaking too close to the microphone, as illustrated in Fig. 3.1. The presence of the impulsive signals causes problems in determining the VAD decision threshold, because the spikes affect the maximum SNR in the file. If the decision threshold is based on the background amplitude and the maximum amplitude, the presence of these spikes will lead to overestimation of the decision threshold, causing low-energy speech segments to be mistakenly detected as non-speech.

Some of the files in NIST 2010 SRE contain a large number of spikes that seriously mask the amplitude of speech segments, as illustrated in Fig. 3.2.



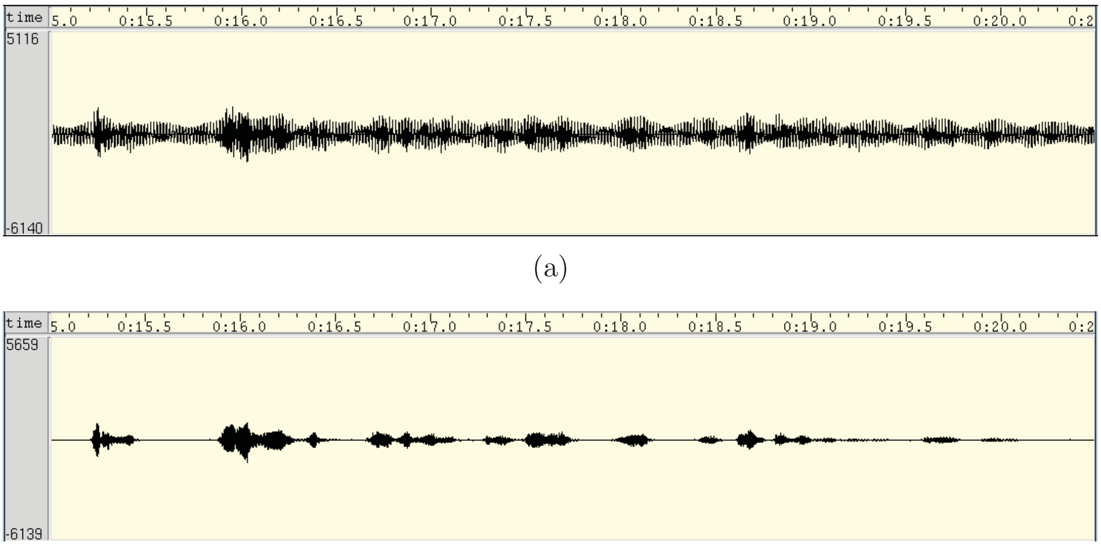
(a)

(b)

(c)

(d)

Figure 3.2: (a) Waveform of low-energy microphone speech in NIST 2010 SRE with numerous high-energy spikes. (b) The short segment of the same utterance. (c) The corresponding spectrogram of the same short segment. (d) Speech/non-speech decisions (`S` for speech and `h#` for silence) made by six VADs abbreviated in Table 6.1 and listening tests (Hand Labeling).



(a)

(b)

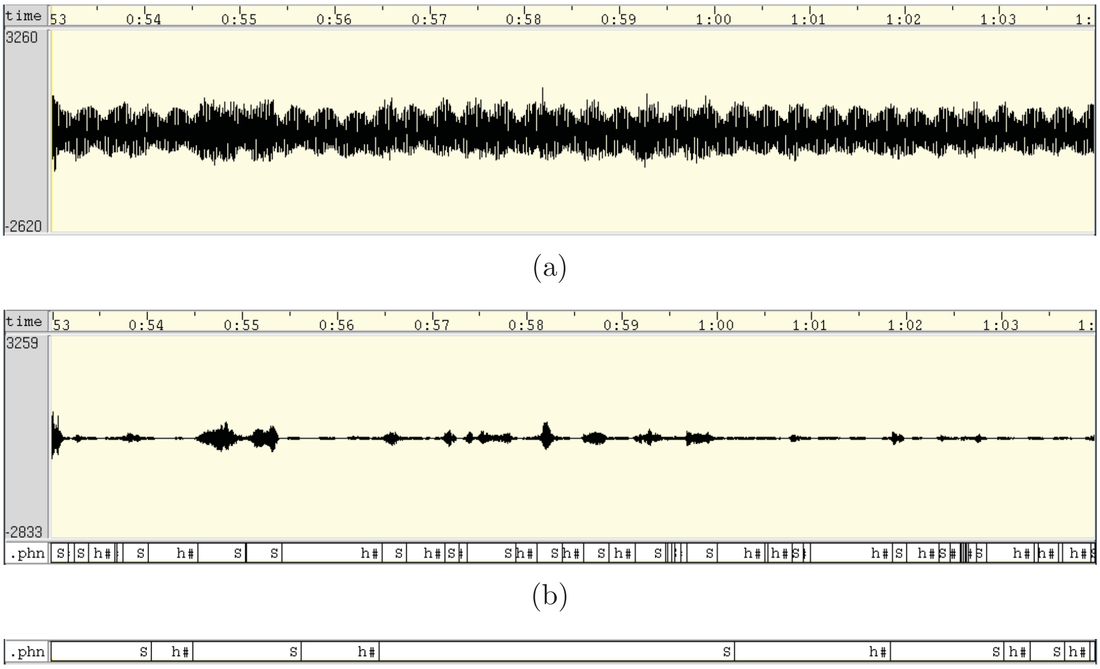
Figure 3.3: (a) A short segment of a speech file in NIST 2008 SRE. The segment contains a high-level of periodic background noise. (b) The same segment after performing spectral subtraction.

3.2 Low-energy Speech Superimposed on Periodic Background Signals

Some files contain low-energy speech superimposed on periodic background noise, as exemplified in Fig. 3.3 and Fig. 3.4.

3.3 Low Signal-to-Noise Ratio

Depending on the microphone types, some of the interview speech segments have extremely low SNR, causing problems in conventional VAD. Fig. 3.5(a) shows the waveform of an interview speech file (ftvhv.sph) in NIST 2008 SRE, and Fig. 3.5(c) highlights a short segment of the same file. Evidently, the SNR is very low. This low SNR will cause numerous errors in energy-based VAD, as evident in the lower panel (labeled with .phn) of Fig. 3.5(c).

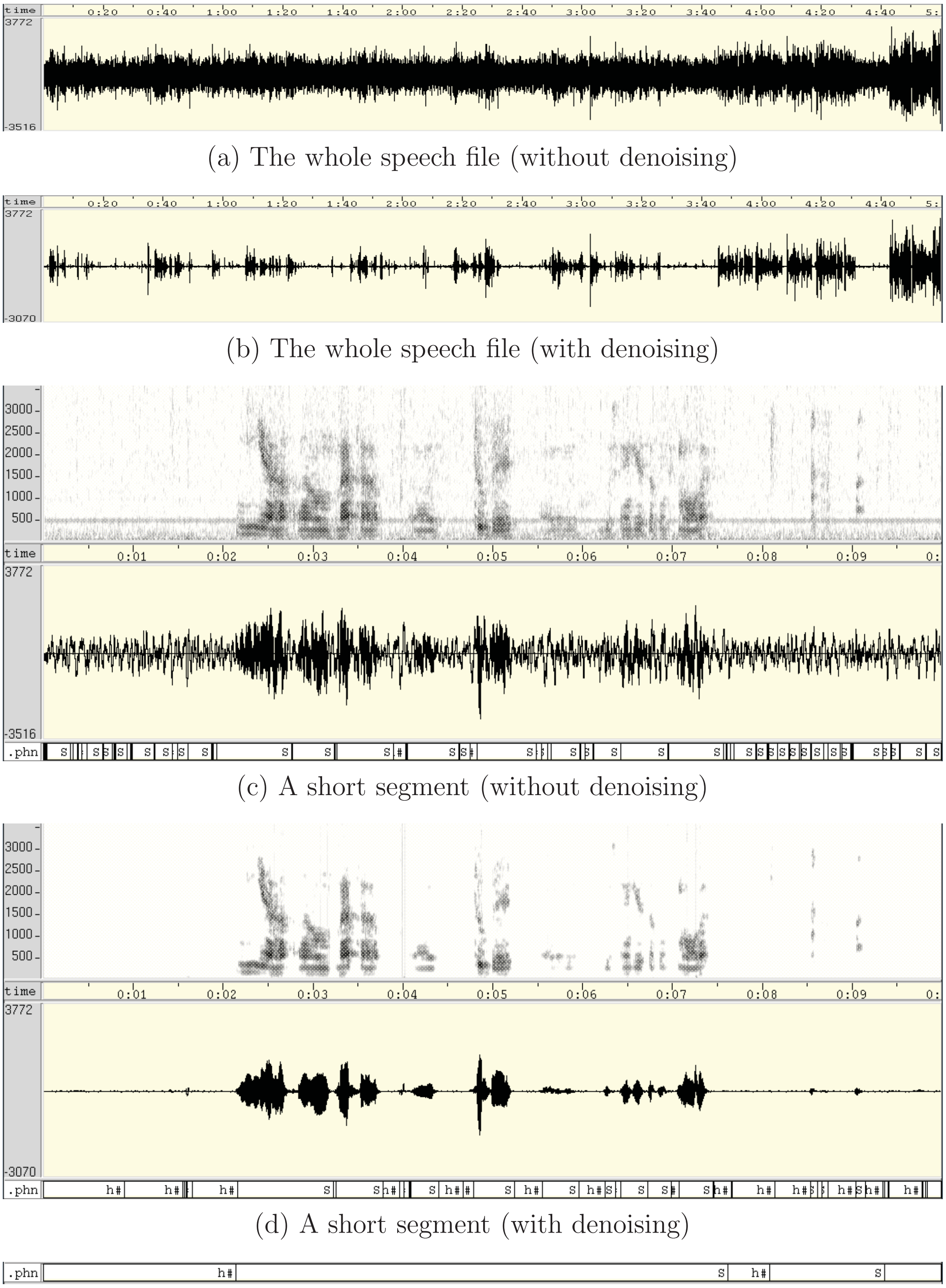


(a)

(b)

(c)

Figure 3.4: (a) A short segment of low-energy interview speech in NIST 2008 SRE superimposed on periodic background noise. (b) The same segment after spectral subtraction. The VAD decisions (`S` for speech and `h#` for silence) are shown in the bottom panel. (c) VAD decisions made by an ETSI-AMR coder.



(a) The whole speech file (without denoising)

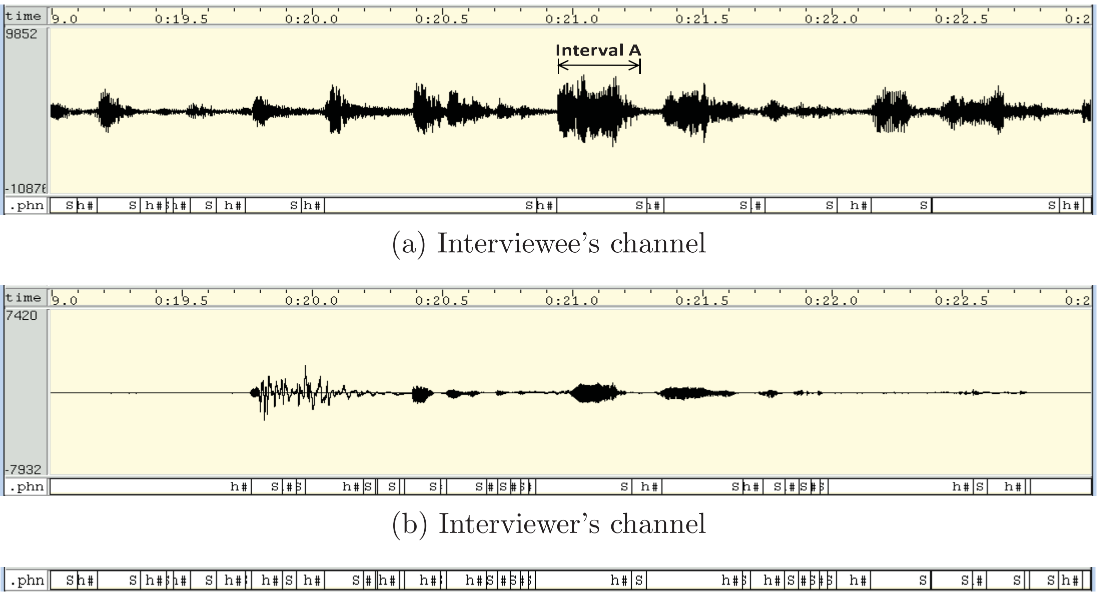
(b) The whole speech file (with denoising)

(c) A short segment (without denoising)

(d) A short segment (with denoising)

(e) A short segment

Figure 3.5: Waveform, spectrogram, and speech/non-speech decision of an interview-speech file in NIST 2008 SRE without [(a) and (c)] and with [(b) and (d)] denoising. (e) VAD decisions of the ETSI-AMR coder, Option 2 [1]. For (c)–(e), the results of VAD are shown in the panels labelled with `.phn`, with `S` and `h#` representing speech and non-speech intervals, respectively.



(a) Interviewee's channel

(b) Interviewer's channel

(c) Crosstalk removed segmentation in Interviewee's channel

Figure 3.6: (a) and (b) show the waveform of a short speech segment from an interviewee and interviewer respectively. The corresponding VAD results (`S` for speech and `h#` for silence) are displayed under their waveform. (c) VAD results of the interviewee's speech after performing crosstalk removal.

3.4 Crosstalks

Each interview speech file in NIST 2010 SRE contains two channels, one recording the speech of an interviewee and the other the speech of an interviewer. As far-field microphones were used for recording interviewee's speech, a low-energy crosstalk signal appears in the interviewee's channel when the interviewer is talking, causing the VAD mistakenly considers the crosstalk as belonging to the interviewee. This situation is exemplified in Fig. 3.6(a) in which the microphone of the interviewee's channel picks up the speech of the interviewer in Interval A. This problem can be solved by using the signal in the interviewer's channel as follows. First, the noise in the interviewer's channel is removed by spectral subtraction, which is followed by identifying the speech segments in the interviewer's channel. Then, the intervals for which the VAD detects speech in both channels are reverted to non-speech. As can be seen from Fig. 3.6(c), Interval A has been successfully reverted to non-speech by using the above strategy.

Chapter 4

VOICE ACTIVITY DETECTION

After investigating the characteristics of interview speech in NIST SREs, we implement different VAD methods to detect the speech segments in these files. This chapter describes statistical-model(SM)-based VAD [4] incorporated with fixed decision threshold. By extending the SM-based VAD, a Gaussian-mixture-model-based VAD [46] that considers the long-term spectro-temporal and static harmonic features [2] is explained. An energy-based VAD that uses spectral subtraction [39, 47, 48] as a preprocessor is then proposed.

4.1 Statistical-Model-Based VAD

4.1.1 Formulation

Recent state-of-the-art VADs are based on likelihood ratio tests where the distributions of the frequency components of speech and noise are approximated by a statistical model (SM). In SM-based VAD, noisy speech $y(t)$ is assumed to be a combination of clean speech $x(t)$ and uncorrelated additive noise $b(t)$, resulting in $y(t) = x(t) + b(t)$, where t represents the sample index. Their corresponding K-dimensional DFT are denoted as $Y(m)$, $X(m)$, and $B(m)$, respectively, where m denotes the frame index. Two hypotheses are made for each frame:

$$\begin{aligned} &H_0 : \text{speech absent} : Y(m) = B(m), \\ &H_1 : \text{speech present} : Y(m) = X(m) + B(m). \end{aligned} \tag{4.1}$$

The distributions of the noisy DFT coefficients conditioned on the above hypotheses are given by [49]

$$p(Y|H_0) = \prod_{k=0}^{K-1} \frac{1}{\pi\lambda_{B,k}} \exp\left\{-\frac{|Y_k|^2}{\lambda_{B,k}}\right\} \tag{4.2}$$

$$p(Y|H_1) = \prod_{k=0}^{K-1} \frac{1}{\pi[\lambda_{B,k} + \lambda_{X,k}]} \exp\left\{-\frac{|Y_k|^2}{\lambda_{B,k} + \lambda_{X,k}}\right\}, \tag{4.3}$$

where $\lambda_{B,k}$ and $\lambda_{X,k}$ denote the variances of additive noise and clean speech for frequency bin k, respectively.

The likelihood ratio for the k-th frequency bin at frame m is

$$\Lambda_k(m) \triangleq \frac{p(Y_k(m)|H_1)}{p(Y_k(m)|H_0)} = \frac{1}{1+\xi_k(m)} \exp\left\{\frac{\gamma_k(m)\xi_k(m)}{1+\xi_k(m)}\right\}, \tag{4.4}$$

where

$$\gamma_k(m) = \frac{|Y_k(m)|^2}{\lambda_{B,k}}, \tag{4.5}$$

$$\xi_k(m) = \frac{\lambda_{X,k}(m)}{\lambda_{B,k}}, \tag{4.6}$$

are respectively defined as the *a posteriori* signal-to-noise ratio (SNR) and the *a priori* SNR. In practice, $\lambda_{B,k}$'s are estimated from non-speech regions; therefore, $\gamma_k(m)$ can be computed easily. To compute $\xi_k(m)$, Sohn et al. [4] suggest to apply the decision direct formulation:

$$\hat{\xi}_k(m) = \alpha \frac{\hat{X}_k^2(m-1)}{\lambda_{B,k}(m-1)} + (1-\alpha)\max\{\gamma_k(m) - 1, 0\} \tag{4.7}$$

where $\hat{X}_k(m-1)$ is the estimated spectrum of the previous frame obtained using the MMSE estimator [49].

To account for the correlation in consecutive speech frames, an HMM-based hangover scheme is adopted. In this scheme, the sequence of frame states (H_0 or H_1) is modeled as a first-order Markov process. Given a set of observations up to frame m, $\mathcal{Y}(m) = \{Y(m), Y(m-1), \ldots, Y(1)\}$, the forward variable is defined as $\alpha_i(m) \triangleq p(q(m) = H_i, \mathcal{Y}(m))$ where $q(m)$ denotes the state of the m-th frame and is either H_0 or H_1. By using the forward procedure [50], $\alpha_i(m)$ can be written as follows:

$$\alpha_i(m) = \begin{cases} P(H_i) \cdot p(Y(1)|q(1) = H_i), & \text{if } m = 1 \\ (\alpha_0(m-1)a_{0j} + \alpha_1(m-1)a_{1j}) \cdot p(Y(m)|q(m) = H_i), & \text{if } m \geq 2. \end{cases} \tag{4.8}$$

where $a_{ij} \triangleq P(q(m) = H_j | q(m-1) = H_i)$.

After some mathematical manipulations, the decision rule can be written as [4]

$$\Gamma(m) = \frac{\alpha_1(m)P(H_0)}{\alpha_0(m)P(H_1)} = \frac{a_{01} + a_{11}\Gamma(m-1)}{a_{00} + a_{10}\Gamma(m-1)} \frac{P(H_0)}{P(H_1)} \Lambda(m) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.9}$$

where η is a decision threshold and $\Lambda(m) = \left[\prod_{k=0}^{K-1} \Lambda_k(m)\right]^{\frac{1}{K}}$ is the geometric mean of the likelihood ratios for the individual frequency bins.

To improve the robustness of VAD under adverse acoustic environment, contextual information derived from multiple observations has been incorporated into the LRT (MO-LRT) [34]. Given the observation vectors $\{X(l-n), \ldots, X(l), \ldots, X(l+n)\}$, the MO-LRT is defined as follows:

$$\varphi(l,n) = \sum_{m=l-n}^{l+n} \ln \left[\frac{p(X(m)|H_1)}{p(X(m)|H_0)}\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.10}$$

where η is a fixed threshold, and l denotes the frame being classified as speech (H_1) or non-speech (H_0). Thus, the decision is based on a sliding window consisting of observation vectors around the current frame. It was found that this decision rule achieves significant improvements in speech/non-speech discrimination when compared with those that rely on a single observation only [51].

4.1.2 Threshold Determination and VAD Decision Logic

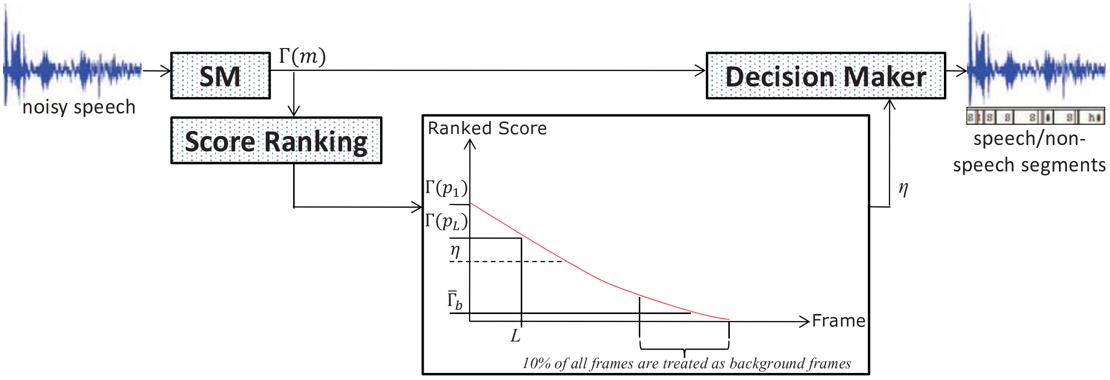


Figure 4.1: The structure of SM-VAD incorporated with a fixed threshold.

The VAD accuracy is highly dependent on the decision threshold η, we advocated a method to determine the threshold that is fixed across the whole utterance. For this method, the SM scores $\Gamma(m)$ of the entire utterance are ranked in descending order as shown in Fig. 4.1. Then, a fixed percentage of scores in the lower and upper ends of the ranked list are selected and assumed to be the background frames and peak frames, respectively. The VAD's fixed decision threshold is a linear combination of the score mean of the lower end ($\bar{\Gamma}_b$) and the minimum score in the upper end as follows:

$$\eta = \nu\bar{\Gamma}_b + (1 - \nu)\min\{\Gamma(p_1), \ldots, \Gamma(p_L)\}, \tag{4.11}$$

where $0 \ll \nu < 1$ is a weighting factor and $\{\Gamma(p_1), \ldots, \Gamma(p_L)\}$ are top-L scores. Note that L cannot be too large; otherwise the rank list may include the peaks of some high-energy speech frames, which will lead to under-estimation of η. However, when L is too small, some medium-amplitude spikes will be missed. It was found that the influence of spikes can be largely eliminated by using the minimum amplitude in this ranked list, as evidenced by the VAD result in the Fig. 3.1(c).

The above procedure raises the issue of determining an appropriate percentage for the lower and upper ends of the ranked score list. These percentages can be founded by inspecting several interview speech files in NIST 2005–2008 SREs. By examining some of these files, we found that it is fairly safe to consider 10% of a speech file contain background frames and 5% of the file contain signal peaks.

4.2 Gaussian-Mixture-Model-Based VAD

This section outlines the procedures of extracting the long-term temporal cepstra and harmonic features [2] from noisy signals and explains how to use these features to discriminate between speech frames and non-speech frames by using Gaussian Mixture Models (GMMs) [46].

4.2.1 Static Harmonic Features

Mel-frequency cepstral coefficients (MFCCs) are known to be inadequate for discriminating speech and non-speech frames, primarily because of the similarity between the static MFCC vectors of speech and background noise. On the other hand, the harmonic structures of speech and background noise are more distinguishable and more noise robust [52]. Based on this argument, Fukauda et al. [2] extracted the harmonic-structure-based features from the middle range of the cepstral coefficients obtained from the discrete cosine transform (DCT) of the power spectral coefficients. Fig. 4.2 shows the procedure of extracting the harmonic-structure-based features.

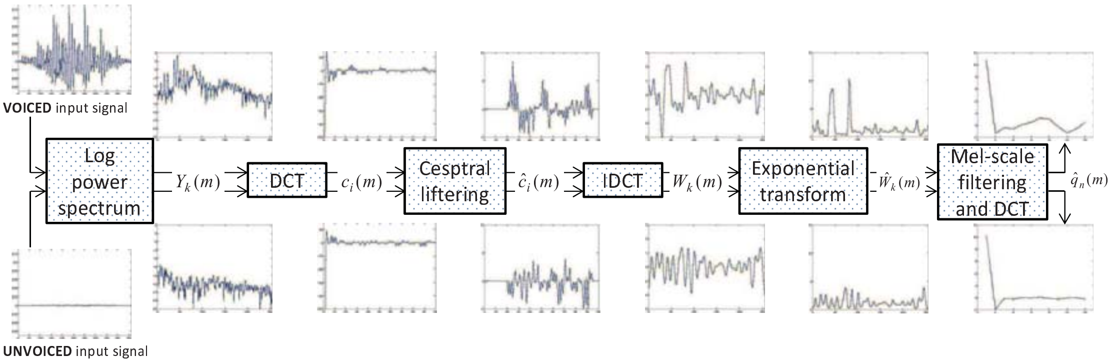


Figure 4.2: The procedure Eq. 4.12 to 4.15 of extracting the harmonic-structure-based features (after [2]).

The power spectrum is first obtained from the observed speech, which is followed by taking logarithm to produce a log power spectrum $Y_k(m)$, where m and k are the frame index and frequency bin index, respectively. Then, a cepstrum

$c_i(m)$ is obtained by applying DCT to the log-power spectrum:

$$c_i(m) = \omega_i \sum_{k=1}^{K} Y_k(m) \cos \frac{\pi(2k-1)(i-1)}{2K}, \; i = 1, \ldots, I \tag{4.12}$$

where $\omega_i = 1/\sqrt{K}$ if $i = 1$, and $\omega_i = \sqrt{2/K}$ otherwise; K is the length of $\boldsymbol{Y}(m)$ and i is the cepstral index. The cepstral coefficients $c_i(m)$ with small and large indexes i are liftered out because they include long and short oscillations. On the other hand, the coefficients in the middle part of the cepstrum capture the harmonic structure information in the human voice. Therefore, the following liftering process is applied to the cepstrum $c_i(m)$:

$$\hat{c}_i(m) = \begin{cases} \lambda c_i(m), & \text{if } (i < D_L) \text{ and } (i > D_H) \\ c_i(m), & \text{otherwise} \end{cases} \tag{4.13}$$

where $\lambda \, (= 10^{-3})$ is a small constant, D_L and D_H are the lower- and upper-limit of cepstral indexes corresponding to the range of pitch frequencies in human voice. For example, for F_0 ranges between 100 and 400 Hz, $D_L = 20$ and $D_H = 80$.[1] The liftered cepstrum $\hat{c}_i(m)$ is converted back to the log power spectrum by Inverse-DCT:

$$W_k(m) = \sum_{i=1}^{I} \omega_i \hat{c}_i(m) \cos \frac{\pi(2k-1)(i-1)}{2I}, \; k = 1, \ldots, K \tag{4.14}$$

followed by the exponential transform to obtain the linear power spectrum

$$\hat{W}_k(m) = \exp\left(W_k(m)\right). \tag{4.15}$$

The coefficients $\hat{W}_k(m)$ are finally converted to mel-cepstrum $\hat{q}_n(m)$ by applying a mel-scale filter bank and DCT, where n is the bin number of the harmonic structure-based mel cepstral coefficients. This feature captures the envelope information of the local peaks in the frequency spectrum corresponding to the harmonic information in the speech signals.

[1] All utterances in NIST SREs were sampled at 8kHz.

4.2.2 Long-Term Dynamic Features

Dynamic (spectro-temporal) features capture the variation of the spectral envelopes along the time axis. They are typically obtained by estimating the derivative of 5 to 9 consecutive acoustic vectors. The first-order derivative of a sequence of cepstral vectors is called delta cepstrum, and the second-order derivative is called delta-delta cepstrum.

Denote the cepstral sequence as $\boldsymbol{C} = [\boldsymbol{c}(1), \boldsymbol{c}(2), \ldots, \boldsymbol{c}(m), \ldots, \boldsymbol{c}(M)]$, where $\boldsymbol{c}(m) = [c_1(m), c_2(m), \ldots, c_I(m)]^T$ is an I-dimensional cepstral vector. The I-dimensional delta cepstrum is given by [53]

$$\Delta c_i(m) = \frac{\sum_{n=-N}^{N} n c_i(m+n)}{\sum_{n=-N}^{N} n^2}, \ i = 1, \ldots, I \tag{4.16}$$

where $c_i(m)$ is the i-th coefficient of $\boldsymbol{c}(m)$ and $2N+1$ cepstral vectors are used for estimating the delta cepstrum. The value of N is set to eight to extract long-term temporal information, leading to long-term dynamic features.

4.2.3 Threshold Determination and VAD Decision Logic

GMM-based VAD is a kind of statistical-model-based VAD in which the K-dimensional feature vectors $\boldsymbol{y}(m), m = 1, \ldots, M$, are assumed to follow a mixture of Gaussian distributions. The probability density functions (PDF) of speech $(i = 1)$ and non-speech $(i = 0)$ vectors in the j-th Gaussian are given by

$$p_j(\boldsymbol{y}(m)|H_i) = \frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}_{ij}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}(m) - \boldsymbol{\mu}_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1}(\boldsymbol{y}(m) - \boldsymbol{\mu}_{ij})\right] \tag{4.17}$$

where $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\Sigma}_{ij}^{-1}$ are the mean vector and covariance matrix for either speech $(i = 1)$ or non-speech $(i = 0)$ model. The Gaussian density functions are then linear combined to give the mixture distributions:

$$p(\boldsymbol{y}(m)|H_i) = \sum_{j=1}^{J} \beta_{ij} p_j(\boldsymbol{y}(m)|H_i) \tag{4.18}$$

where β_{ij}'s are the mixture coefficients, H_0 and H_1 represent non-speech and speech hypothesis (model), respectively.

The decision rule is obtained by comparing the log-likelihood ratio

$$\mathcal{L}(m) = \log p(\boldsymbol{y}(m)|H_1) - \log p(\boldsymbol{y}(m)|H_0) \tag{4.19}$$

with decision threshold η:

$$\mathcal{L}(m) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.20}$$

The decision threshold η is determined by a strategy similar to that of SM-VAD described in Section 4.1.2. Specifically, 20% and 5% of a speech file are assumed to contain background frames and signal peaks, respectively.

Architecture of GMM-Based VAD

Unlike the SM-based VAD, the GMM-based VAD require the training of two GMMs – one representing speech and another one representing non-speech. This means that some speech files with speech and non-speech segmentations are required. In theory, the segmentations had better to be the ground-truths, i.e., they need to be done by listening tests and human inspections of spectrograms. This is not a problem if clean speech files are available and the VAD is tested on the same files but with noise added to them, e.g., the experiments in [2]. However, in NIST SREs, the requirement of ground-truth segmentations will cause difficulty because no clean speech files are available for the listening tests or spectrogram inspections. Even if we can find some interview-style speech files with high enough SNR for the listening tests, they may be too clean and therefore cannot represent the realistic situations in other noisy speech files. Furthermore, hand labeling of large amount of speech files is simply too laborious and time-consuming. To overcome this difficulty, this dissertation proposes method that can determine the segmentations that are close enough to the ground-truths for training the GMMs without human intervention.

Fig. 4.3 shows the architecture of GMM-based VAD for NIST SREs. Unlike

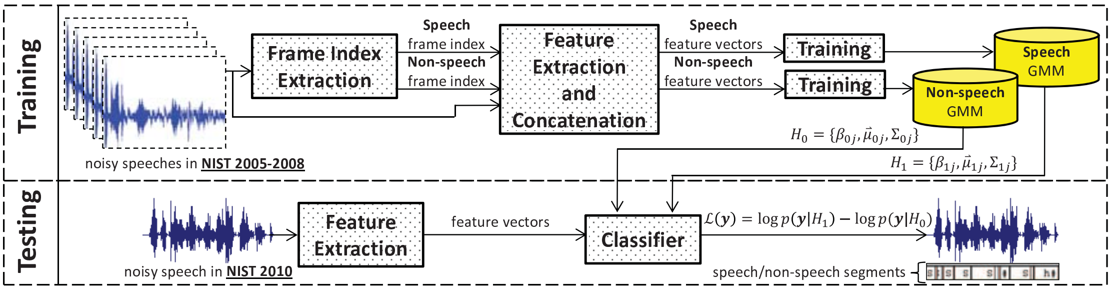

Training
noisy speeches in NIST 2005-2008
Frame Index Extraction
Speech frame index
Non-speech frame index
Feature Extraction and Concatenation
Speech feature vectors
Non-speech feature vectors
Training
Training
Speech GMM
Non-speech GMM
$H_0 = \{\beta_{0j}, \vec{\mu}_{0j}, \Sigma_{0j}\}$
$H_1 = \{\beta_{1j}, \vec{\mu}_{1j}, \Sigma_{1j}\}$
Testing
noisy speech in NIST 2010
Feature Extraction
feature vectors
Classifier
$\mathcal{L}(\boldsymbol{y}) = \log p(\boldsymbol{y}|H_1) - \log p(\boldsymbol{y}|H_0)$
speech/non-speech segments

Figure 4.3: Overview of the GMM-based VAD. See Fig. 4.4 for the algorithm of frame index extraction.

the VAD in [2], our GMM-based VAD contains an extra processing block (Frame Index Extraction) that finds the frame indexes of speech and non-speech segments with very high confidence of being correct. This seems to create a chicken-and-egg problem because if a reliable VAD exists, we do not need to build a new one in the first place. However, having some reliable speech and non-speech segments does not mean that we need a reliable VAD to detect both at the same time. The idea is that we can always make a simple energy-based VAD "very" reliable in detecting speech but extremely unreliable in detecting non-speech by adjusting the decision threshold such that it can achieve a very low false alarm (consider non-speech as speech) but having a very high missing rate (consider speech as non-speech). A similar argument applies to the reliable detection of non-speech. Note that as this simple VAD can only maintain either the false alarm or missing rate low but not both, it can only be used as a pre-processing step in more sophisticated VADs such as the one illustrated in Fig. 4.3.

The idea is to leverage the large number of speech files in NIST SREs available for training the GMMs. Specifically, for each interview-style speech files in the training set (e.g., past NIST SREs), a simple energy-based VAD is used to determine the energy of all frames. Then, the frames are ranked in ascending order of energy as illustrated in Fig. 4.4. The top 5% of the ranked list are discarded because the high energy is most likely caused by spiky signals instead of speech. Because of the simplicity of the energy-based VAD, there will be many false alarms and misses in the detections. Therefore, only a small percentage in the upper- and lower-part of the ranked list are considered as speech and non-speech, respectively. In other words, about 99% of the frames in the middle of the ranked list will be ignored, and only the frames with a very high confidence of having a correct segmentation are retained for training the GMMs.

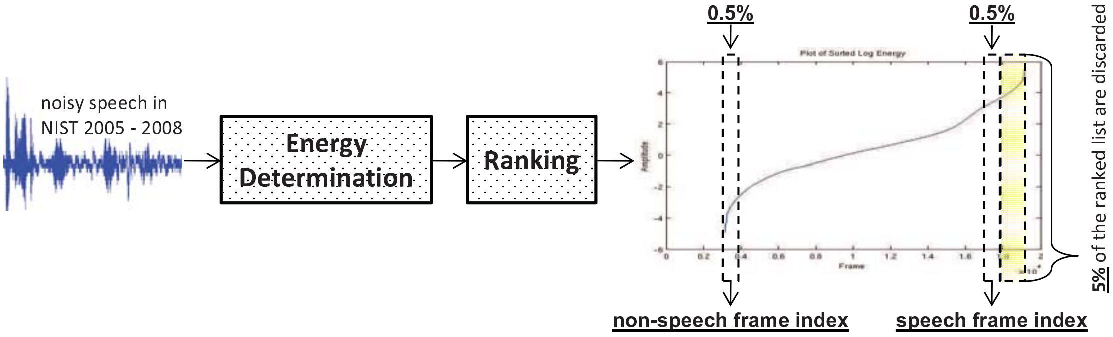


Figure 4.4: The procedure of extracting the frame indexes representing the speech and non-speech segments in the processing block “Frame Index Extraction” in Fig. 4.3.

Given the frame indexes of speech and non-speech segments, static harmonic features and long-term dynamic features are extracted and concatenated (i.e., $\boldsymbol{y} = [\hat{\boldsymbol{q}}^T \; \Delta\boldsymbol{c}^T]^T$), forming two streams of feature vectors as shown in Fig. 4.3. These concatenated features vectors are then used to train the GMMs.

In this work, 3569 interview-style utterances from NIST SRE 2005–2008 were used for training the GMMs. This amount to 280,010 training vectors per GMM. The number of mixtures J was set to 32, and all Gaussians have a full covariance matrix. Similar to the SM-based VAD described in Section 4.1.2, the decision threshold η was determined by considering 20% and 5% of a speech file contain background frames and signal peaks, respectively.

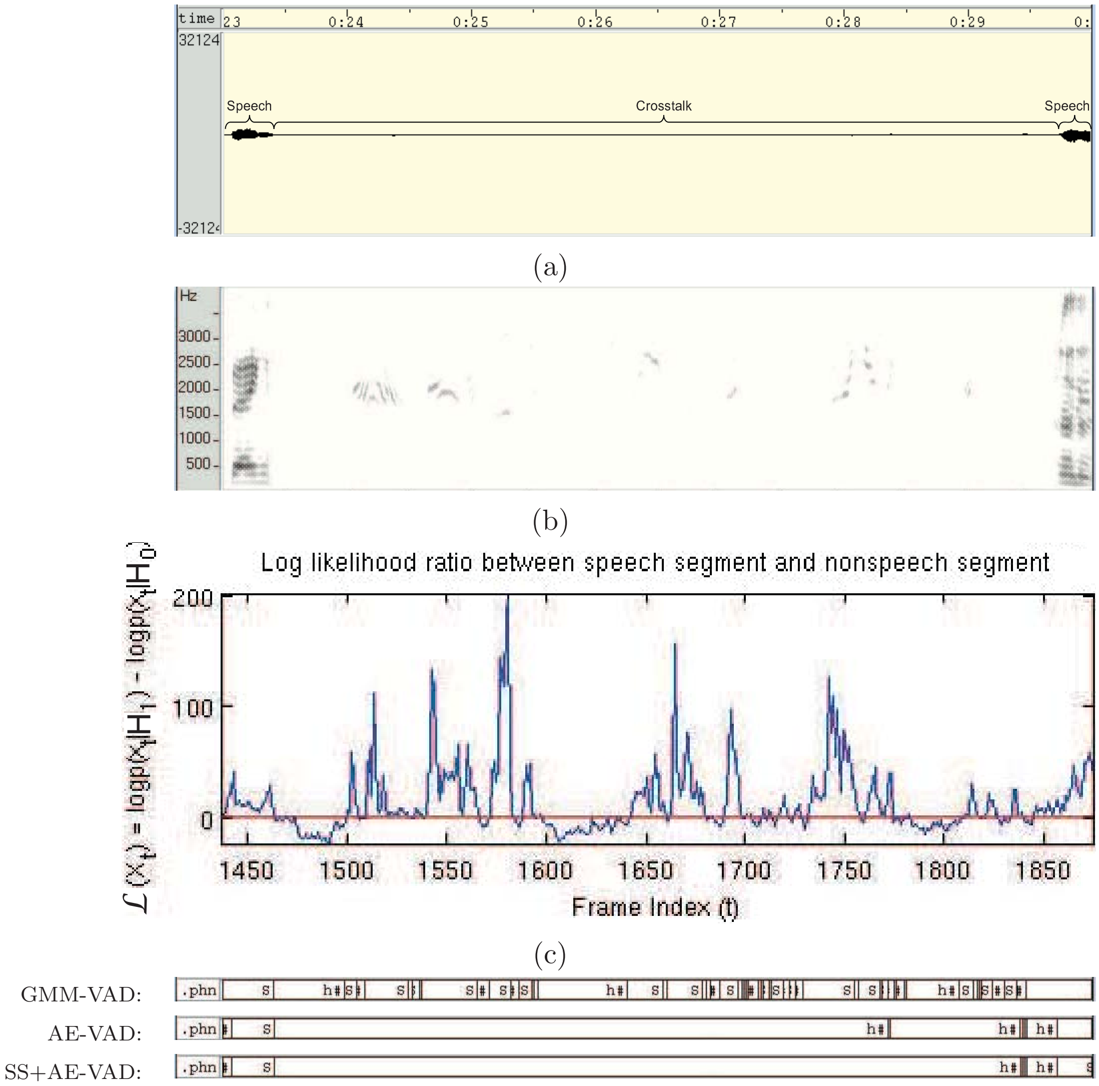
4.2.4 Characteristics of GMM-Based VAD

Unlike the SM-based VAD, the GMM-based VAD uses GMMs to model the distribution of multi-dimension acoustic features. This approach makes the VAD less susceptible to spiky signals because these signals have low-level of harmonic contents and their temporal property is also different from that of speech signals.

This ability is evident in Fig. 3.2 where the segmentations of GMM-based VAD are closest to those obtained by hand labeling.

However, GMM-based VAD also has its own limitations. In particular, because the GMM-based VAD does not rely on SNR, it could falsely detect some weak cross-talks from other speakers as speech segments as long as the cross-talks contain speech-like characteristics. This phenomenon is especially apparent in the "Phonecall-Microphone" speech files[2] in NIST 2010 SRE, as exemplified in Fig. 4.5. Evidently, this drawback can be alleviated by using spectral subtraction as a pre-processor because the weak cross-talks will be considered as background signals so that they can be largely eliminated in the spectral subtraction process. Further discussions on the use of spectral subtraction as a pre-processor can be found in Section 4.3.

[2] Speech passed through a telephone channel but recorded by microphones.



(a)

(b)

(c)

(d)

Figure 4.5: (a) A short segment of phonecall-microphone speech in NIST 2010 SRE with not-so-apparent crosstalk. (b) The spectrogram of the short segment in (a). (c) The log likelihood ratio (blue curve) of the segment obtained by a GMM-based VAD with the fixed decision threshold (red line). (d) Speech/non-speech decisions (`S` for speech and `h#` for silence) made by GMM-based VAD (GMM-VAD), the energy-based VAD without (AE-VAD) and with spectral subtraction (SS+AE-VAD).

4.3 Energy-Based VAD with Spectral Subtraction

Noise removal is a vital step for pre-processing the interview speech files in NIST SREs because many of them have very low SNR. This dissertation proposes to apply spectral subtraction (SS) with a large over-subtraction factor to discard the background noise as much as possible before passing the enhanced speech to an energy-based VAD. Advanced speech enhancement techniques (e.g. MMSE [49] and LSA-MMSE [54]) have not been used because audio quality of reconstructed speech is not the main concern. Instead, it is more important to increase the SNR in speech regions and to minimize the background noise in non-speech regions. Spectral subtraction can well meet this requirement without unnecessarily complicating the whole system.

4.3.1 Noise Reduction via Spectral Subtraction

To obtain the enhanced speech $\hat{x}(t)$ from the noisy speech $y(t)$ at frame m, we implemented the spectral subtraction [39, 47, 48] of the form

$$\hat{X}_k(m) = \begin{cases} [|Y_k(m)| - \alpha(m)|\hat{B}_k|]e^{j\varphi_k(m)} & \text{if } |Y_k(m)| > (\alpha(m) + \beta(m))|\hat{B}_k| \\ \beta(m)|\hat{B}_k|e^{j\varphi_k(m)} & \text{otherwise,} \end{cases} \tag{4.21}$$

where k is the frequency bin index, $\varphi_k(m)$ is the phase of $Y_k(m)$, $\hat{B}_k$ is the average spectrum of some non-speech regions, $\alpha(m)$ is an over-subtraction factor for removing background noise, and $0 < \beta(m) \ll 1$ is a spectral floor factor ensuring that the recovered spectra never fall below a preset minimum. When the SNR is low, the spectral floor factor ensures that a low-level of noise is present in the enhanced signal. This noise helps to reduce the musical noise that may otherwise be introduced if the recovered spectrum $\hat{X}_k(m)$ is set to zero.

The value of $\alpha(m)$ and $\beta(m)$ can be computed as

$$\begin{aligned} \alpha(m) &= -\frac{1}{2}\gamma(m) + c \qquad (\alpha_{\min} \leq \alpha(m) \leq \alpha_{\max}) \\ \beta(m) &= \begin{cases} \beta_{\min} & \text{if } \gamma(m) < 1 \\ \beta_{\max} & \text{otherwise} \end{cases} \end{aligned} \tag{4.22}$$

where $\gamma(m) = \frac{\sum_k |Y_k(m)|}{\sum_k |\hat{B}_k|}$ is the *a posteriori* SNR, c is a constant ($= 4.5$ in this work), $\alpha_{\min}$, $\alpha_{\max}$, $\beta_{\min}$, and $\beta_{\max}$ constrain the allowable range of the over-subtraction factor and the noise floor. These limits are set according to the amount of tolerable musical noise in the denoised speech. Note that musical noise is not a main concern in our application as speakers' features were extracted from the original files instead of the enhanced files. We thus set these values such that the speech spectra are over-subtracted when the SNR is low. In this work, we set $\alpha_{\max} = 4$, $\alpha_{\min} = 0.5$, $\beta_{\max} = 0.05$, and $\beta_{\min} = 0.01$. These values were determined by observing the reconstructed waveform of several files.

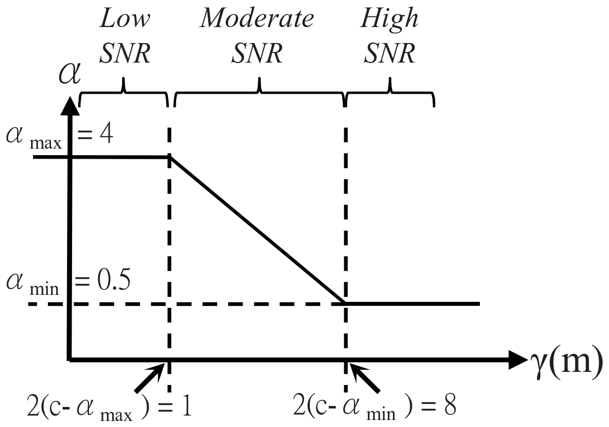


Figure 4.6: Plot of α against γ in the Eq. 4.22.

Eq. 4.22 ensures that when SNR is high (see Fig. 4.6) $\alpha(m)$ will be small, and therefore spectral subtraction (upper-part of Eq. 4.21) will occur, but the amount of subtraction is small. This ensures that the original speech signal will not be significantly distorted by the subtraction process. For moderate SNR, either over-subtraction or noise flooring may be applied to the noisy spectra. For those frequency components that are subject to over-subtraction, the amount of subtraction is larger. This ensures that more noise will be removed. At regions with very low SNR, noise flooring is more likely to occur. For the frequency components that meet the condition, $|Y_k| - \alpha|\hat{B}_k| > \beta|\hat{B}_k|$, majority of noise will be largely removed because α becomes very large.

Note that α should not be too large for two reasons. First, if α is very large, most of the frequency components cannot meet the condition $|Y_k| - \alpha|\hat{B}_k| > \beta|\hat{B}_k|$, which results in losing speech contents in the denoised speech. Second, for those components that can meet the above condition, the degree of over-subtraction will be too significant, resulting in severely distorted speech.

For the value of β, at region of low SNR, β is set to its minimum value. This ensures that the non-speech region of the denoised speech has a low-level of noise. Note that at low SNR, almost all frequency component cannot meet the condition $|Y_k| - \alpha|\hat{B}_k| > \beta|\hat{B}_k|$. As a result, noise flooring will be applied to almost all frequency components. Therefore, keeping β small can help reducing the background noise in the denoised speech. On the other hand, at the region of high SNR, β is set to its maximum value. This strategy helps to avoid the abrupt change in the frequency spectra of the denoised speech at high SNR region. The reason is that although at high SNR, majority of the frequency components can meet the condition $|Y_k| - \alpha|\hat{B}_k| > \beta|\hat{B}_k|$, some of the frequency components may not. For those components that cannot meet this condition, noise flooring will be applied. If the value of β is too small, the corresponding frequency components in the denoised speech will be significantly smaller than the other components, causing significant spectral distortion in the denoised speech.

4.3.2 Applying Spectral Subtraction to the Energy-Based VAD

Fig. 4.7 shows the structure of the proposed energy-based VAD, which we refer to as SS+AE-VAD. Figs. 3.5(b) and (d) show the same speech file and segment as in Figs. 3.5(a) and (c) but after spectral subtraction. Evidently, with the background noise largely removed, speech and non-speech intervals can be correctly detected by

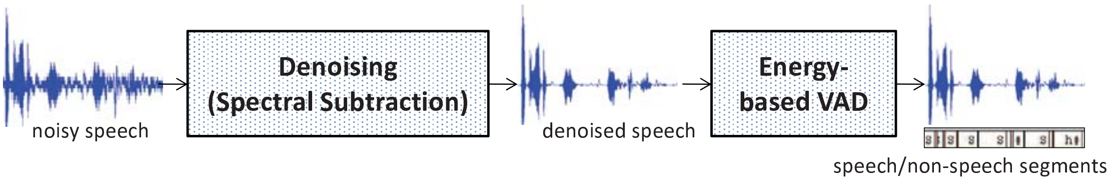


Figure 4.7: The structure of the proposed VAD for NIST SREs.

an energy-based VAD. To highlight the advantage of spectral subtraction, Fig. 3.5 compares the segmentation results of SS+AE-VAD and that of the ETSI-AMR coder (Option 2). The figure suggests that this coder over-estimates the length of speech segments, whereas the SS+AE-VAD correctly detects the speech segments.

To collect more evidences on the advantage of noise removal, we applied energy-based VAD without SS, ETSI AMR, and energy-based VAD with SS to extract the speech segments of 6249 files in NIST 2005–2008. For each file, we used the three detectors to extract the speech segments and computed the ratio between speech-segment length and total-signal length. The distributions of speech-segment-length to total-signal-length ratio are shown in Fig. 4.8. The figure shows that without noise removal, the detector mistakenly determines many non-speech segments as speech segments in a large number of speech files, as evident by the high frequency of occurrences at ratio over 0.9–1.0. On the other hand, with noise removal, the detector considers half of the total signals contain speech in many speech files. The ETSI AMR lies in between VAD with noise removal and VAD without noise removal.

4.3.3 Threshold Determination and VAD Decision Logic

The presence of spikes in some files affects the maximum SNR in these files, which needs to be taken care of when determining the VAD decision threshold. In particular, these spikes lead to overestimation of the decision threshold if it is based on the background amplitude and the maximum amplitude. Consequently, low-energy speech segments could be mistakenly detected as non-speech. To address this problem, we have developed a similar strategy as the one in Section 4.1.2, but

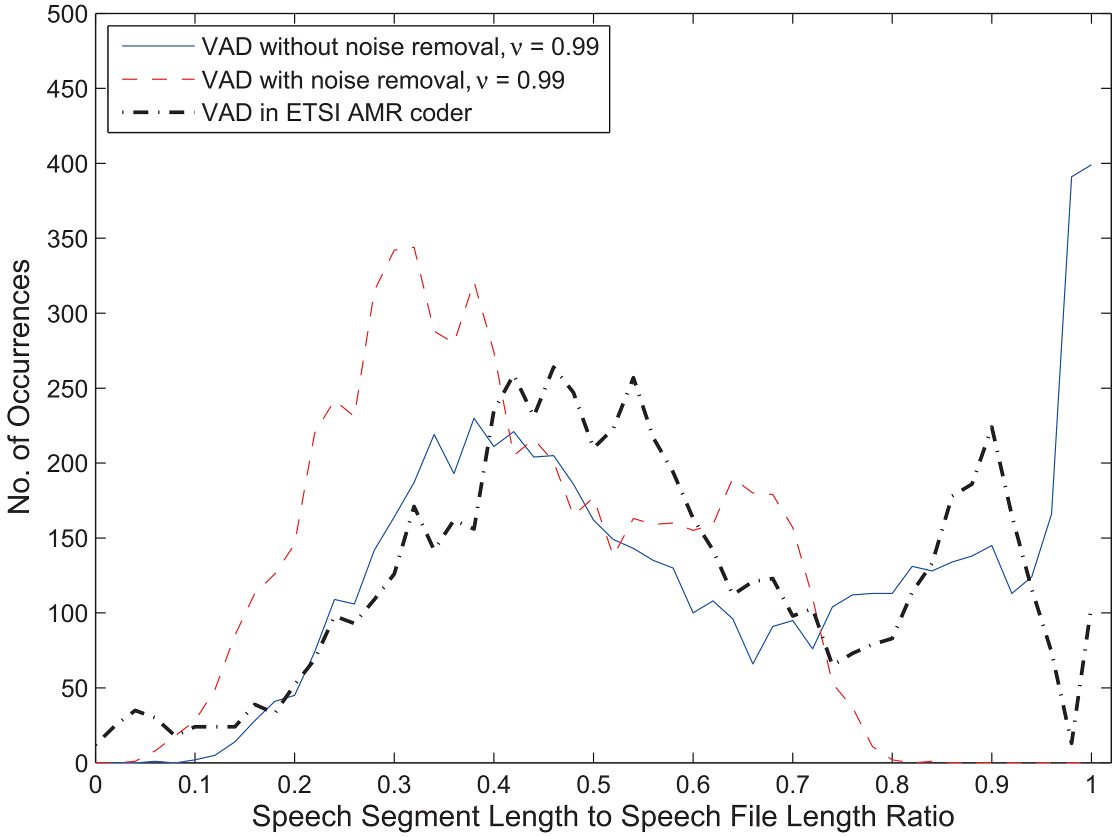


Figure 4.8: Distribution of speech-segment-length to total-signal-length ratio determined by three VAD detectors: energy-based VAD without noise removal (blue), energy-based VAD with noise removal (red dashed), and VAD (Option2) in ETSI-AMR coder (black dashed-dot).

considering signal amplitude rather than statistical scores. The decision threshold is a linear combination of the mean of background amplitude ($\bar{a}_b$) and the minimum of the signal peaks:

$$\eta = \nu \bar{a}_b + (1-\nu)\min\{a(p_1), \ldots, a(p_L)\}, \tag{4.23}$$

where $\{a(p_1), \ldots, a(p_L)\}$ are the amplitudes of L largest-amplitude frames. In this work, L was set to 1% of the total number of frames in the speech file. By comparing the amplitude of each frame in the file with the threshold, those frames with amplitude larger than the threshold are considered as speech frames.

However, some speech files contain segments with a large DC offset after spectral subtraction, and these segments should be considered as non-speech. Therefore, another decision logic is applied: Frame with extremely low zero-crossing rate (smaller than 10% of background zero-crossing rate) are considered as non-speech. The pseudo-code of the proposed SS+AE-VAD can be found in the appendix A.

Chapter 5

ANALYTICAL COMPARISON OF SM-VAD AND SS+AE-VAD

This chapter compares and contrasts analytically the decision rules of the statistical-model-based VAD (SM-VAD) and the spectral-subtraction-based VAD (SS+AE-VAD) described in Chapter 4.

5.1 Decision Rules of Energy-Based VAD

5.1.1 Without Spectral Subtraction

The decision rule of an energy-based VAD without spectral subtraction is given by

$$\begin{aligned} &\frac{1}{K}\sum_{k=0}^{K-1}|Y_k|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{K}\sum_{k=0}^{K-1}|\hat{B}_k|^2+\eta \\ \Rightarrow &\frac{1}{K}\sum_{k=0}^{K-1}\left[|Y_k|^2-|\hat{B}_k|^2\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta \end{aligned} \tag{5.1}$$

where $k = 0, \ldots, K-1$ is the frequency bin index, K is the frame size, and η is a decision threshold.

5.1.2 With Generalized Spectral Subtraction

With the generalized spectral subtraction, the k-th frequency bin of the enhanced power spectrum $|\hat{X}_k|^2$ is given by $|Y_k|^2 - \alpha|\hat{B}_k|^2$. Then the *a priori* and *a posteriori* SNR are respectively given by $\xi_k = |\hat{X}_k|^2/|\hat{B}_k|^2 = \lambda_{X,k}/\lambda_{B,k}$ and $\gamma_k = |Y_k|^2/|\hat{B}_k|^2 = \lambda_{Y,k}/\lambda_{B,k}$. Consider the generalized spectral subtraction in Eq. 4.21

as a filtering process shown in Fig. 5.1. Then, $|H_k|^2$ can be expressed as

$$\frac{|\hat{X}_k|^2}{|Y_k|^2} = \frac{|Y_k|^2 - \alpha|\hat{B}_k|^2}{|Y_k|^2} = \frac{\frac{|Y_k|^2}{|\hat{B}_k|^2} - \alpha}{\frac{|Y_k|^2}{|\hat{B}_k|^2}} = \frac{\gamma_k - \alpha}{\gamma_k}. \tag{5.2}$$

Therefore, we have $|H_k| = \sqrt{\frac{\gamma_k - \alpha}{\gamma_k}}$, and for $|H_k|$ to be real, $\gamma_k \geq \alpha \ \forall k$ and thus $|Y_k|^2 \geq \alpha|\hat{B}_k|^2$. To make VAD decisions, we compute the energy of $|\hat{X}_k|$ for each frame and compare it with the background energy, i.e.,

$$\frac{1}{K}\sum_{k=0}^{K-1}|\hat{X}_k|^2 = \frac{1}{K}\sum_{k=0}^{K-1}\left[\frac{\gamma_k - \alpha}{\gamma_k}\right]|Y_k|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{K}\sum_{k=0}^{K-1}|\hat{B}_k|^2 + \eta \tag{5.3}$$

According to Eq. 4.21, if $|Y_k|^2 < |\hat{B}_k|^2$, then $|\hat{X}_k|^2 = \beta|\hat{B}_k|^2$. Therefore, when SNR is small, the VAD decision rule is given by

$$\frac{1}{K}\sum_{k=0}^{K-1}\beta|\hat{B}_k|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{K}\sum_{k=0}^{K-1}|\hat{B}_k|^2 + \eta. \tag{5.4}$$

To make correct decisions, β needs to be less than 1, which agrees with the setting of $\beta_{\max}$ and $\beta_{\min}$ in Section 4.3.1.

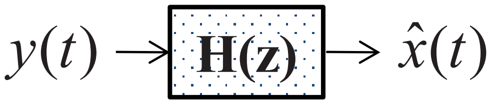


Figure 5.1: Relationship between noisy speech (input) and denoised speech (output).

5.1.3 With Simple Spectral Subtraction

When $\alpha(m) = 1$ and $\beta(m) = 0$ in Eq. 4.21, we have the most basic form of spectral subtraction. The enhanced power spectrum is given by

$$|\hat{X}_k|^2 = \begin{cases} |Y_k|^2 - |\hat{B}_k|^2 & \text{if } |Y_k|^2 > |\hat{B}_k|^2 \\ 0 & \text{otherwise,} \end{cases} \tag{5.5}$$

and using Eq. 5.2 and Eq. 5.3, the VAD decision rule is given by

$$\begin{aligned} &\frac{1}{K}\sum_{k=0}^{K-1}\left[\frac{\gamma_k - 1}{\gamma_k}\right]|Y_k|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{K}\sum_{k=0}^{K-1}|\hat{B}_k|^2 + \eta \\ \Rightarrow &\frac{1}{K}\sum_{k=0}^{K-1}\left[|Y_k|^2 - \frac{|Y_k|^2}{\gamma_k}\right] \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{K}\sum_{k=0}^{K-1}|\hat{B}_k|^2 + \eta \\ \Rightarrow &\frac{1}{K}\sum_{k=0}^{K-1}\left[|Y_k|^2 - |\hat{B}_k|^2\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta + \frac{1}{K}\sum_{k=0}^{K-1}\lambda_{B,k} \\ \Rightarrow &\frac{1}{K}\sum_{k=0}^{K-1}\left[|Y_k|^2 - |\hat{B}_k|^2\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta' \end{aligned} \tag{5.6}$$

where $\eta' = \eta + \frac{1}{K}\sum_{k=0}^{K-1}\lambda_{B,k}$, and Eq. 4.5 has been used in the derivation. Eq. 5.1 and Eq. 5.6 suggest that VAD with simple spectral subtraction can be reduced to energy-based VAD as in Eq. 5.1. Therefore, in order to make spectral subtraction useful for VAD application, we should consider the generalized form in Eq. 5.3.

5.1.4 Benefit of Spectral Subtraction

Without spectral subtraction, the ratio between the energy of speech and non-speech can be determined by

$$\text{SNR}_{(withoutSS)} = \gamma = \frac{\sum_k |Y_k|^2}{\sum_k |\hat{B}_k|^2}. \tag{5.7}$$

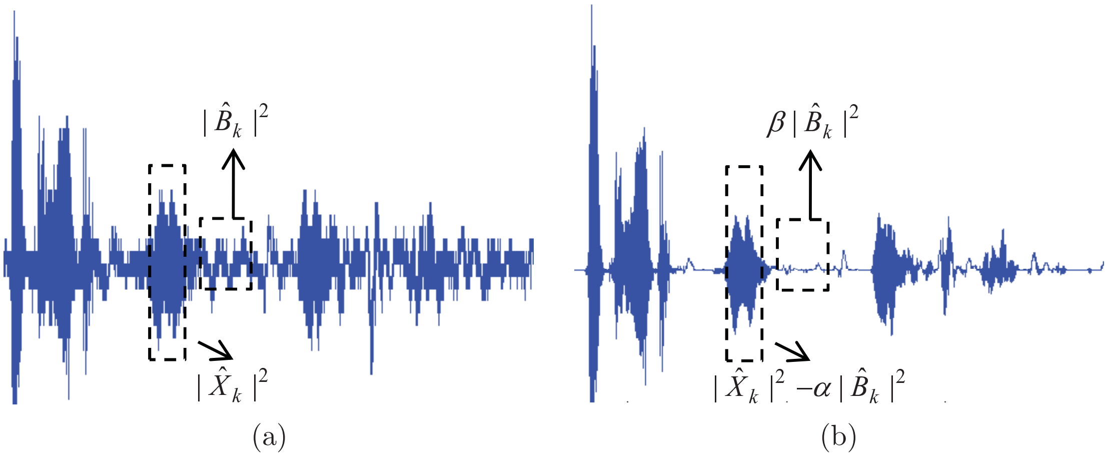


Figure 5.2: Waveform of a short segment (a) before and (b) after performing spectral subtraction.

With spectral subtraction, the signal-to-noise ratio becomes

$$\mathrm{SNR}_{(withSS)} = \gamma' = \frac{\sum_k \left[|Y_k|^2 - \alpha|\hat{B}_k|^2\right]}{\beta \sum_k |\hat{B}_k|^2} = \frac{\gamma - \alpha}{\beta}, \tag{5.8}$$

where γ is the SNR without SS in Eq. 5.7. The minimum value of γ for which spectral subtraction is beneficial can be found by the following inequality:

$$\begin{aligned} &\mathrm{SNR}_{(withSS)} > \mathrm{SNR}_{(withoutSS)} \\ \Rightarrow &\frac{\gamma - \alpha}{\beta} > \gamma \\ \Rightarrow &\gamma(1-\beta) > \alpha \\ \Rightarrow &\gamma > \frac{\alpha}{1-\beta}. \end{aligned} \tag{5.9}$$

This inequality provides a guideline for setting the limits of α and β. Let us use the setting in Section 4.3.1 as a numerical example. Using $\alpha_{\max} = 4$, $\alpha_{\min} = 0.5$, $\beta_{\max} = 0.05$ and $\beta_{\min} = 0.01$, we have

$$0.53 = \frac{0.5}{1 - 0.05} = \frac{\alpha_{\min}}{1 - \beta_{\max}} < \frac{\alpha}{1 - \beta} < \frac{\alpha_{\max}}{1 - \beta_{\min}} = \frac{4}{1 - 0.01} = 4.04.$$

Therefore, as long as the SNR γ is greater than 0.53, the SNR after spectral subtraction will be higher than that before spectral subtraction, thus bringing benefit to the energy-based VAD. Fig. 5.2 further exemplifies this situation.

5.2 Decision Rules of SM-Based VAD

5.2.1 Without Spectral Subtraction

To simplify analysis, let us consider the Itakura-Saito distortion (ISD) based decision rule (Eq. 6 of [4]) in SM-based VAD:

$$\log \hat{\Lambda} = \frac{1}{K} \sum_{k=0}^{K-1} \{\gamma_k - \log \gamma_k - 1\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{5.10}$$

where $\gamma_k = |Y_k|^2 / |B_k|^2$.

5.2.2 With Spectral Subtraction

With spectral subtraction, we replace γ_k in Eq. 5.10 with $\gamma'_k = \frac{\gamma_k - \alpha}{\beta}$, leading to the following decision rule:

$$\begin{aligned}
&\frac{1}{K} \sum_k [\gamma'_k - \log \gamma'_k - 1] \underset{H_0}{\overset{H_1}{\gtrless}} \eta' \\
\Rightarrow &\frac{1}{K} \sum_k \left[\frac{\gamma_k - \alpha}{\beta} - \log\left(\frac{\gamma_k - \alpha}{\beta}\right) - 1\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta' \\
\Rightarrow &\frac{1}{K} \sum_k [\gamma_k - \alpha - \beta \log(\gamma_k - \alpha) + \beta \log \beta - \beta] \underset{H_0}{\overset{H_1}{\gtrless}} \eta' \beta \\
\Rightarrow &\frac{1}{K} \sum_k [\gamma_k - \beta \log(\gamma_k - \alpha) + c] \underset{H_0}{\overset{H_1}{\gtrless}} \eta''
\end{aligned} \tag{5.11}$$

where $c = \beta \log \beta - \alpha - \beta$ and $\eta'' = \eta' \beta$. As Eq. 5.10 and Eq. 5.11 have a similar form, statistical model with and without spectral subtraction will not produce significantly different VAD result as long as $\gamma_k > \alpha$. This argument is supported by the experimental results in Section 6.2.2.

Chapter 6

EXPERIMENTS

VAD algorithms are typically evaluated by comparing the VAD results on clean speech against the VAD results on noise contaminated speech [55], with performance shown on a receiver operating characteristic (ROC) curve. However, the noisy speech files in NIST SREs do not have their clean counterparts. Instead of hand labeling a large amount of speech files, this chapter uses the performance indexes of speaker verification, i.e. EER, DET, minimum DCF and minimum normalized DCF, for quantifying VAD performance.

The experiments involve nine VADs. They are (see Table 6.1)

1. AE-VAD: average-energy-based VAD,

2. ASR-VAD: ASR transcripts provided by NIST 2010 SRE Workshop,

3. AMR-VAD: the VAD in the ETSI-AMR coder (option 2),

4. SM-VAD: a statistical-model-based VAD,

5. GMM-VAD: a Gaussian-mixture-model-based VAD,

6. SS+SM-VAD: spectral subtraction followed by SM-VAD, and

7. SS+AE-VAD: spectral subtraction followed by AE-VAD.

6.1 Selection of Threshold Parameters for SS+AE-VAD

As mentioned in Section 4.3, energy-based VAD requires a decision threshold for making speech/non-speech decisions. This section describes an experiment that

VAD		Description
1	AE-VAD	Energy-based VAD with the decision governed by the combination between average magnitude of background noise and signal peaks. The combination is controlled by a weighting factor (ν in Eq. 4.23).
2	ASR-VAD	Speech segments in the Automatic Speech Recognition transcripts provided by NIST [45].
3	AMR-VAD	VAD in ETSI Adaptive Multi-Rate coder (Option2) [1].
4	SM-VAD	Sohn's statistical-model-based VAD incorporated with a fixed threshold, determined by Eq. 4.11.
5	GMM-VAD	Gaussian-mixture-model-based VAD using long-term temporal information and harmonic structure-based features in noisy speech [2] incorporated with a fixed decision threshold.
6	SS+SM-VAD	SM-VAD with spectral subtraction as a pre-processing step.
7	SS+AE-VAD	AE-VAD with spectral subtraction as a pre-processing step.

Table 6.1: The voice activity detection (VAD) methods being applied in this dissertation and their acronym.

investigates the effect of the weighting factor ν (Eq. 4.23) on the energy-based VAD.

6.1.1 Experimental Setup

NIST 2005–2008 Speaker Recognition Evaluations (SREs) were used in the experiments. NIST'05 and NIST'06 SREs were used as development data, and NIST'08 was used for performance evaluations.[1] Only male speakers in these corpora were used.

The core task (short2-short3) of NIST'08 has eight common conditions. We focus on Common Conditions 1 to 4 (CC1–CC4), because these four conditions involve interview speech. For example, CC3 reflects the performance of systems that were trained and tested on different microphones in the interview recordings. Table 6.2 summaries these four common conditions in NIST'08.

For each utterance, an energy-based VAD, the ETSI-AMR coder, and the proposed spectral subtraction energy-based VAD were used to remove the silence

[1] Hereafter, all NIST SREs are abbreviated as NIST'XX, where XX stands for the year of evaluation.

Common Condition	Train/Test Condition	No. of Targets	No. of Trials
1	All Interview speech	622	14405
2	Interview speech, same microphone type for training and test	125	731
3	Interview speech, different microphone types for training and test	622	13674
4	Interview speech for training, telephone speech for test	622	5048

Table 6.2: The training and test speech types used in Common Conditions 1 to 4 in NIST'08 (male speakers).

regions. This procedure results in three segmentation files for each utterance. For the SS+AE-VAD (see Table 6.1), different values of the weighting factor (ν in Eq. 4.23) were applied to the speech files in NIST'08. For the speech files in NIST'05 and NIST'06 used for creating the UBM [13] and Tnorm models [56], the weighting factor was set to 0.95.[2]

In feature extraction, twelfth-order MFCCs [12] plus their first derivative were extracted from the speech regions of the utterance, leading to 24-dim acoustic vectors. We used GMM-SVM [57] as target-speaker models. Specifically, interview utterances from the male speakers of NIST'05 and NIST'06 were used for creating a 512-center, gender-dependent universal background model (UBM). MAP adaptation [13], with relevance factor set to 16 was then performed for each of the target-speakers to create target-dependent GMMs. The same MAP adaptation was also applied to 300 background speakers (also from NIST'05 and '06) to create 300 impostor GMMs. The mean vectors of these GMMs were stacked to form 12288-dim GMM-supervectors [57]. For each target speaker, his target-dependent GMM-supervector and the background GMM-supervectors were used to train a GMM-SVM speaker model.

To reduce channel effects, 81 male speakers from NIST'05 and NIST'06 were

[2] The weighting factor was fixed for all speech files used for creating the UBM and Tnorm models because we assume that the optimal value of this parameter can be obtained during system development.

VAD Method	ν	EER (%)				Minimum DCF			
		CC1	CC2	CC3	CC4	CC1	CC2	CC3	CC4
AE-VAD	0.95	8.28	1.93	8.08	13.61	0.041	0.009	0.041	0.053
AE-VAD	0.99	8.14	3.57	7.71	12.68	0.043	0.015	0.041	0.048
AMR-VAD	–	8.51	1.53	8.54	11.05	0.040	**0.003**	0.040	0.042
SS+AE-VAD	0.00	16.69	10.21	16.72	19.49	0.069	0.052	0.066	0.074
SS+AE-VAD	0.80	13.29	5.51	13.40	16.36	0.055	0.012	0.055	0.054
SS+AE-VAD	0.90	9.09	1.93	9.25	12.21	0.041	0.009	0.041	0.039
SS+AE-VAD	0.95	6.94	**1.12**	7.08	9.99	0.035	0.008	0.035	**0.038**
SS+AE-VAD	0.99	**6.44**	**1.12**	**6.37**	**9.64**	**0.032**	0.007	**0.032**	**0.038**
SS+AE-VAD	1.00	8.94	1.83	8.82	13.11	0.042	0.009	0.041	0.053

Table 6.3: Performance on NIST 2008 SRE under common conditions (CC) 1 to 4. ν in the 2nd column is the weighting factor in Eq. 4.23 for the interview-speech files in NIST'08. **AE-VAD**: energy-based VAD without noise removal. **AMR-VAD**: VAD in AMR coder. **SS+AE-VAD**: the proposed spectral-subtraction VAD.

used for estimating the gender-dependent nuisance attribute projection (NAP) matrices [58]. Each of these speakers has at least 8 utterances. The NAP corank was set to 128. Three hundred male utterances from NIST'05 were used for creating Tnorm speaker models [56]. The same set of background speakers used for creating the target-speaker SVMs were used for creating the Tnorm SVMs.

6.1.2 Results and Discussions

Table 6.3 and Fig. 6.1 show the equal error rate (EER) and minimum decision cost function (minDCF) achieved by the three VAD methods. The results shown in Fig. 6.1 strongly suggest that preprocessing the noisy sound files by spectral subtraction is a promising idea. With SS, the VAD reduces the EER by 21% in CC1.

Table 6.3 and Fig. 6.2 also suggest that the best range of ν in Eq. 4.23 is between 0.95 and 0.99. Once this value drops below 0.95, the performance degrades rapidly. This implies that the peak amplitudes can only be used as a reference for setting the VAD decision threshold, whereas the background amplitudes are more trustworthy. However, the threshold cannot totally relies on the background

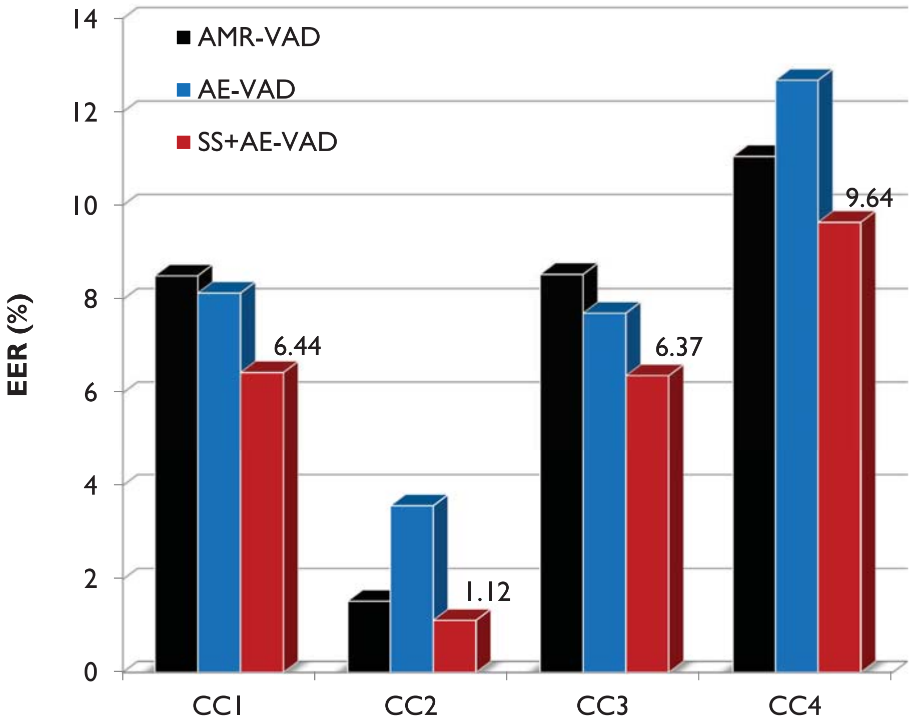


Figure 6.1: The lowest EER achieved by three different VADs in Table 6.3 under Common Conditions (CC) 1–4 in NIST'08 (male).

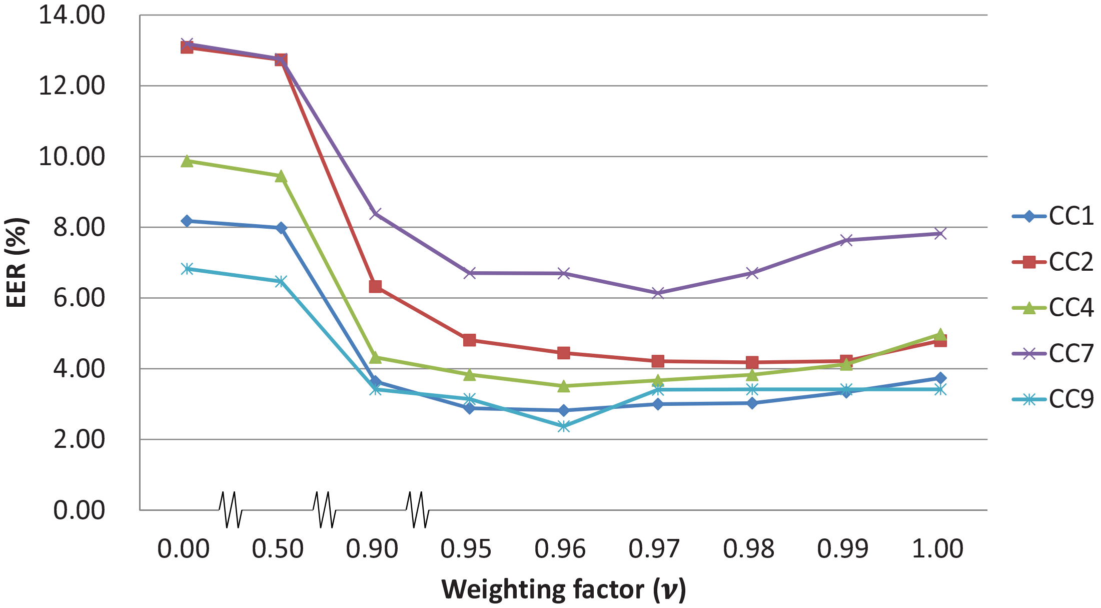


Figure 6.2: Equal error rate against weighting factor ν in Eq. 4.23 under Common Conditions (CC) 1, 2, 4, 7 and 9 in NIST'10 (male). SS+AE-VAD (see Table 6.1) was used in all cases.

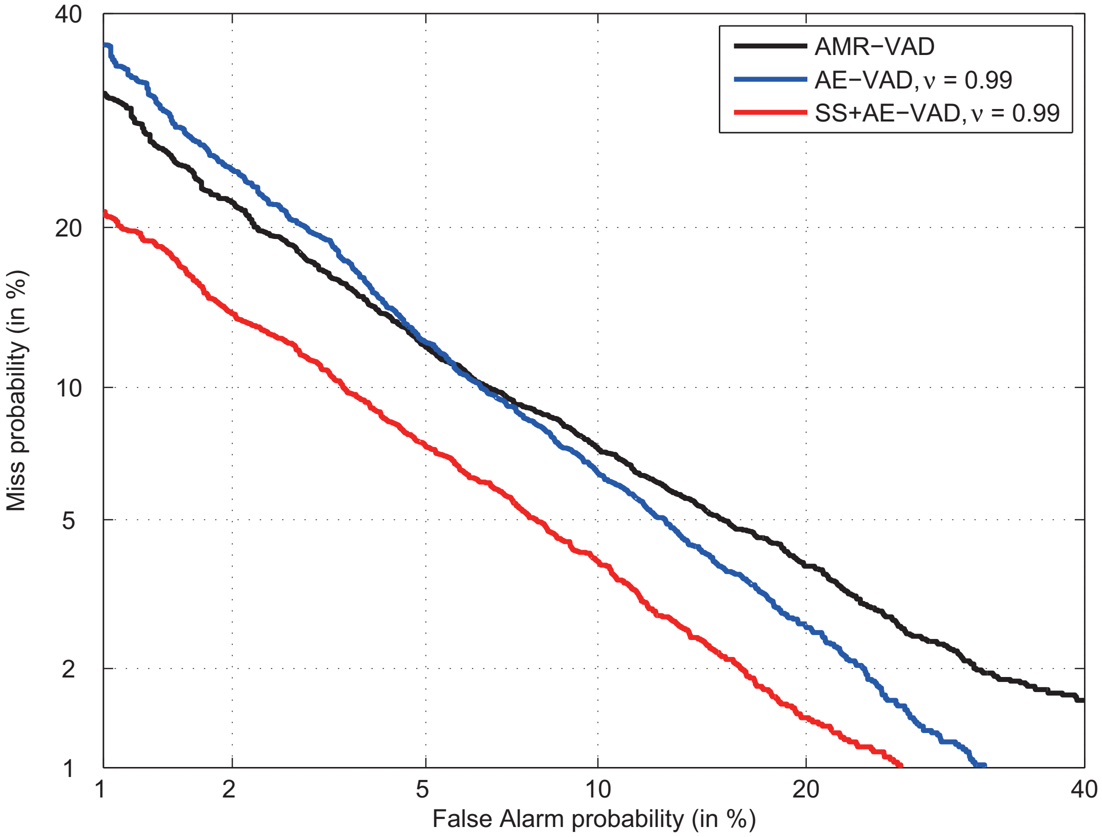


Figure 6.3: DET performance of 3 VADs under Common Condition 1 in NIST'08 (male).

amplitude, because the EER and minDCF increase when ν increases from 0.99 to 1.0.

Fig. 6.3 shows the DET performance (under CC1) of the three VAD methods. The results show that SS+AE-VAD achieves a significant lower error rates than the ETSI-AMR coder for a wide range of operating points.

6.2 Comparison of Different VADs

Different VADs mentioned in Table 6.1 were applied to justify the effectiveness of the proposed SS+AE-VAD.

6.2.1 Experimental Setup

NIST 2005–2010 SREs were used in the experiments. NIST'05–08 SREs were used as development data, and NIST'10 was used for performance evaluations. Only male speakers in these corpora were used. The core task of NIST'10 is divided into 9 common conditions. Conditions 1, 2, 4, 7 and 9 were considered because interview speech and telephone speech collected by different microphones are involved in these five conditions. Detail descriptions of these five conditions can be found in the Section 4 of [59].

Note that in VADs 6–7 of Table 6.1, spectral subtraction was used as a preprocessing step to remove the background noise. By comparing the speaker verification performance obtained by these VADs against the ones without spectral subtraction, we can observe the contribution of spectral subtraction to the VAD performance.

The weighting factor ν in Eq. 4.23 was set to 0.95 and 0.96 for AE-VAD and SS+AE-VAD, respectively. For SM-VAD, SS+SM-VAD, and GMM-VAD, ν in Eq. 4.11 was set to 0.993.

We extracted 12 MFCCs [12] and their first derivatives from the speech regions of the utterances to create 24-dim acoustic vectors. Cepstral mean normalization [60] was applied to the MFCCs, followed by feature warping [61].

The target-speakers were modeled by GMM-SVM [57]. In the modeling process, a gender-dependent universal background model (512-center) was created by using the interview utterances of NIST'05–06. MAP adaptation [13], with relevance factor set to 16, was then performed for each of the target-speakers to create target-dependent GMMs. The same MAP adaptation was also applied to 300 background speakers (also from NIST'05–06) to create 300 impostor GMMs.

The mean vectors of these GMMs were stacked to produce 12288-dim GMM-supervectors [57]. Finally, a GMM-SVM speaker model for each target speaker is trained by using his target-dependent GMM-supervector and the background GMM-supervectors.

The utterances of 144 male speakers from NIST'05–08 were used for estimating the gender-dependent NAP matrices [58] to reduce channel effects (NAP corank was set to 128). Each of these 144 speakers has at least 8 utterances. For the T-norm speaker models [56], 300 male utterances from NIST'05 were used. The same set of background speakers used for creating the target-speaker SVMs were used for creating the T-norm SVMs.

6.2.2 Results and Discussions

Table 6.4 shows the equal error rate (EER) and minimum normalized decision cost function (minNDCF) achieved by seven VAD methods. The results strongly suggest that preprocessing the noisy sound files by spectral subtraction is a promising idea. After applying SS, the AE-VAD and SM-VAD reduce the overall EER by 56% and 5% respectively.

The overall values of EER and minNDCF in Table 6.4 were then plotted for the seven VADs in Fig. 6.5. Evidently, the proposed SS+AE-VAD attains the lowest EER and minNDCF among all seven VADs.

Fig. 6.4 shows the DET performance based on the seven VADs. The results show that SS+AE-VAD achieves a significant lower error rates than the ETSI-AMR coder, ASR transcripts and the simple energy-based VAD for a wide range of operating points. In the plot, SM-VAD performs even better than GMM-VAD. We notice that both SS and SM work well for the interview speech in NIST 2010 SRE. The error rates achieved by SS+AE-VAD, however, are slightly lower than that achieved by SM-VAD.

Comparing the results of AE-VAD and SS+AE-VAD reveals that SS has significant contribution to the conventional energy-based VAD. However, the performance of SS+SM-VAD is better than SM-VAD by a small margin only. This

VAD Method	CC1	CC2	CC4	CC7	CC9	Overall
AE-VAD	6.57	11.72	7.23	12.28	7.44	10.30
ASR-VAD	5.15	8.58	7.74	12.81	5.74	8.88
AMR-VAD	4.44	8.05	9.44	12.85	5.98	9.61
GMM-VAD	3.64	5.68	5.71	8.93	4.27	6.28
SM-VAD	3.23	4.68	4.49	9.48	3.06	5.03
SS+SM-VAD	2.83	4.45	4.04	7.58	2.56	4.80
SS+AE-VAD	**2.82**	**4.44**	**3.51**	**6.70**	**2.37**	**4.55**

(a) EER (%)

VAD Method	CC1	CC2	CC4	CC7	CC9	Overall
AE-VAD	0.84	0.99	0.96	0.84	0.97	0.97
ASR-VAD	0.78	0.85	0.74	0.88	0.77	0.90
AMR-VAD	0.81	0.85	0.80	0.77	0.55	0.90
GMM-VAD	0.71	0.72	0.72	0.63	0.45	0.82
SM-VAD	0.66	0.68	0.70	0.65	0.38	0.77
SS+SM-VAD	**0.62**	0.61	0.70	**0.59**	0.42	0.76
SS+AE-VAD	0.70	**0.58**	**0.62**	0.64	**0.17**	**0.72**

(b) minNDCF

Table 6.4: Performance on NIST 2010 SRE achieved by 7 VADs. (a) EER and (b) minimum normalized DCF under Common Conditions (CC) 1, 2, 4, 7 and 9. **AE-VAD**: energy-based VAD without noise removal; **ASR-VAD**: VAD segmentation from NIST provided ASR transcripts; **AMR-VAD**: VAD in ETSI-AMR coder; **GMM-VAD**: Gaussian-mixture-model-based VAD [2]; **SM-VAD**: Sohn's VAD [4] incorporated with the proposed fixed thresholds (Eq. 4.11); **SS+SM-VAD**: SM-VAD with spectral subtraction; **SS+AE-VAD**: the proposed spectral-subtraction VAD.

suggests that SS is not vital to the statistical-model-based VAD. The reason is that in SM-based VADs, the background spectrum has already been taken into account in the scoring function. As pre-processing the noisy speech by spectral subtraction is another approach to using the background spectrum, therefore in SS+SM-VAD, the background spectrum has been used twice. As a result, the gain of applying SS to SM-VAD is not very significant.

Note that SS+AE-VAD and SM-VAD use the background spectrum in a different manner. For the former, the background spectrum is used for spectral subtraction, whereas for the latter it is used for computing the likelihood ratio scores. This difference enables us to make better use of the background spectrum in SS+AE-VAD. Specifically, to remove as much background noise as possible, we

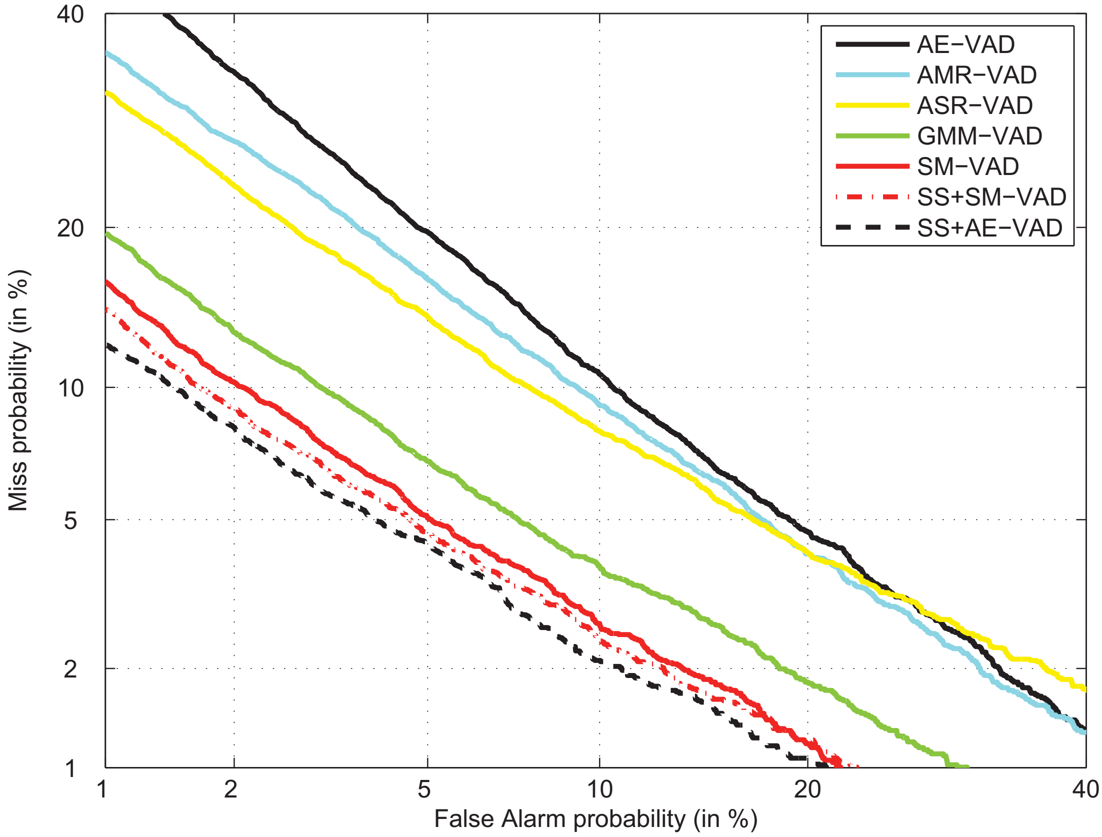


Figure 6.4: DET performance of all trials combining common conditions 1, 2, 4, 7, and 9 in NIST'10 (male). Labels in the legend are arranged in descending EER.

may apply a large upper-limit for the over-subtraction factor (α_{max}) and a small lower-limit for the noise floor (β_{min}).[3] The over-subtraction factor $\alpha(m)$ is a linear function of the *a posteriori* SNR for certain range of SNR and is bounded by the lower- and upper-limit when the SNR is beyond this range. As a result, more background noise will be removed in low SNR region whereas more speech content will be retained in high SNR region. The SM-VAD, on the other hand, does not have such property because the background spectrum is assumed constant for both low and high SNR.

The results show that using the ASR transcripts provided by NIST SRE Workshop as VAD leads to poor speaker verification performance, suggesting that the

[3] Note that musical noise is not a concern because the denoised speech is only used for VAD, not for speaker recognition.

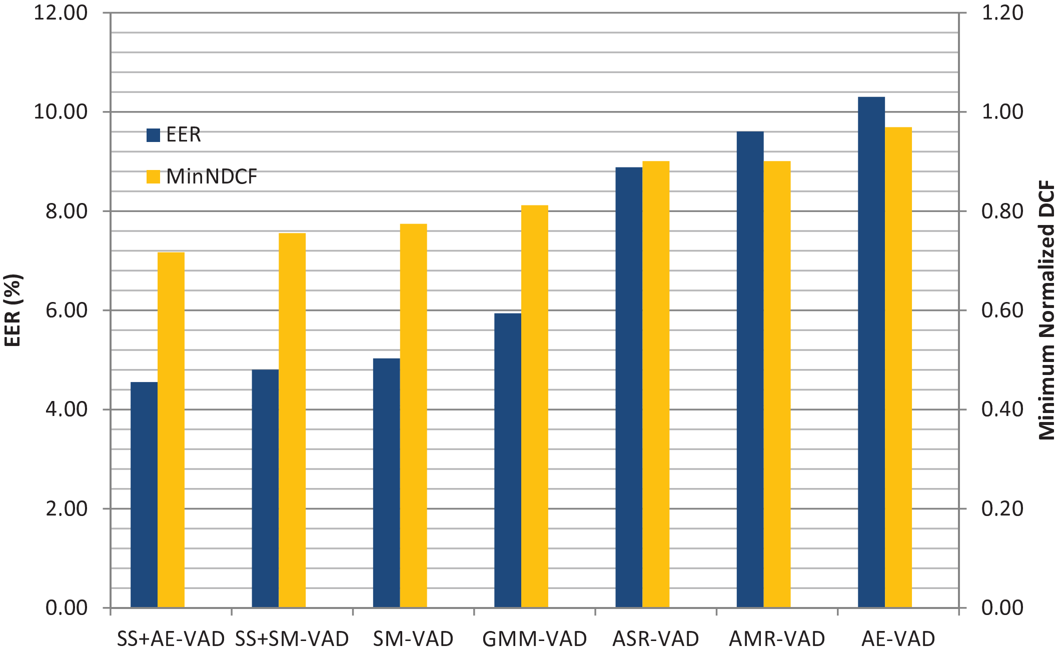


Figure 6.5: EER and minimum normalized DCF based on all trials in Common Conditions 1, 2, 4, 7 and 9 achieved by the 7 VADs in Table 6.1.

ASR transcripts do not produce accurate speech/non-speech segmentations. The VAD in ETSI-AMR coder also performs poorly. This is mainly caused by the overestimation of both the speech onset and offset regions. To ensure the intelligibility of the encoded speech, it is important for the VAD in a speech coder to include speech onsets and offsets. However, this overestimation is not appropriate for speaker verification, as excessive amount of non-speech will be used for verification.

Chapter 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

A voice activity detector specially designed for extracting speech segments from the interview-speech files in NIST SREs has been proposed and evaluated under the NIST 2008 and 2010 SREs protocols. Several conclusions can be drawn from this work:

1. noise reduction is of primary importance for VAD under extremely low SNR;

2. it is important to remove the sinusoidal background noise found in NIST SRE sound files as this kind of background signal could lead to many false detection in energy-based VAD;

3. a reliable threshold strategy is required to address the spiky (impulsive) speech signals, and;

4. our proposed spectral subtraction VAD outperforms the segmentations derived from the ASR transcripts provided by NIST, the VAD in the advanced speech coder (ETSI-AMR, Option2), the state-of-the-art statistical-model-based VAD, and Gaussian-mixture-model-based VAD in speaker verification.

7.2 Future Work

The GMM-based VAD can be improved in two aspects: (1) Selection of Noise Robust Features and (2) Determination of better decision thresholds.

7.2.1 Selection of Noise-robust Features for GMM-based VAD

MFCCs have been widely used in speaker recognition due to their acceptable performance under moderate noisy conditions. However, MFCC-based systems are susceptible to the acoustic mismatch in training and testing conditions. More recently, researchers have started to investigate a new feature for speaker recognition [3]. The new feature promises to be more robust to noise and is capable of capturing speaker identity conveyed in speech signals [3]. It has been found that feature parameters obtained from the temporal envelope of a gammatone filterbank can achieve a significantly higher recognition accuracy than that of MFCCs [3]. This motivates extensive research efforts in the use of such feature for training GMM-based speaker recognition systems [5].

Fig. 7.1 depicts the procedure for extracting the new acoustic feature for GMM-based VAD. The speech signal $s(n)$ is first filtered using a 32-channel gammatone filterbank to simulate the effect of auditory filtering [62]. Hilbert transform is then applied to the temporal envelope of the j-th channel $s(n, j)$ of the filter to obtain the Hilbert envelope $e(n, j)$. To determine the amplitude of the temporal envelope at frame t, the sample means are calculated:

$$E(t, j) = \frac{1}{N} \sum_{n=0}^{N-1} w(n) e(n, j), \tag{7.1}$$

where $w(n)$ denotes the Hamming window and N is the frame size. Natural logarithm is then applied to the envelope $E(t, j)$ to compress the dynamic range. The compressed envelopes are normalized for each channel by their long-term average:

$$E_n(t, j) = \frac{E_{log}(t, j)}{\frac{1}{T} \sum_{t=1}^{T} E_{log}(t, j)}. \tag{7.2}$$

Finally, DCT is applied to decorrelate the normalized features to produce 32-dimensional spectral vectors called mean Hilbert envelope coefficients (MHEC).

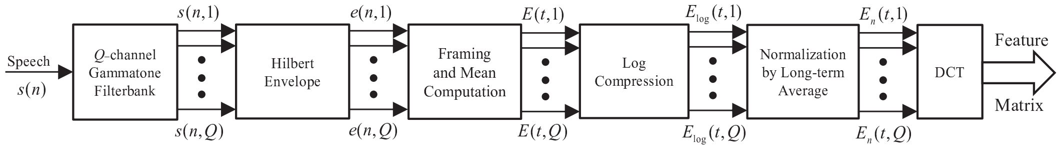
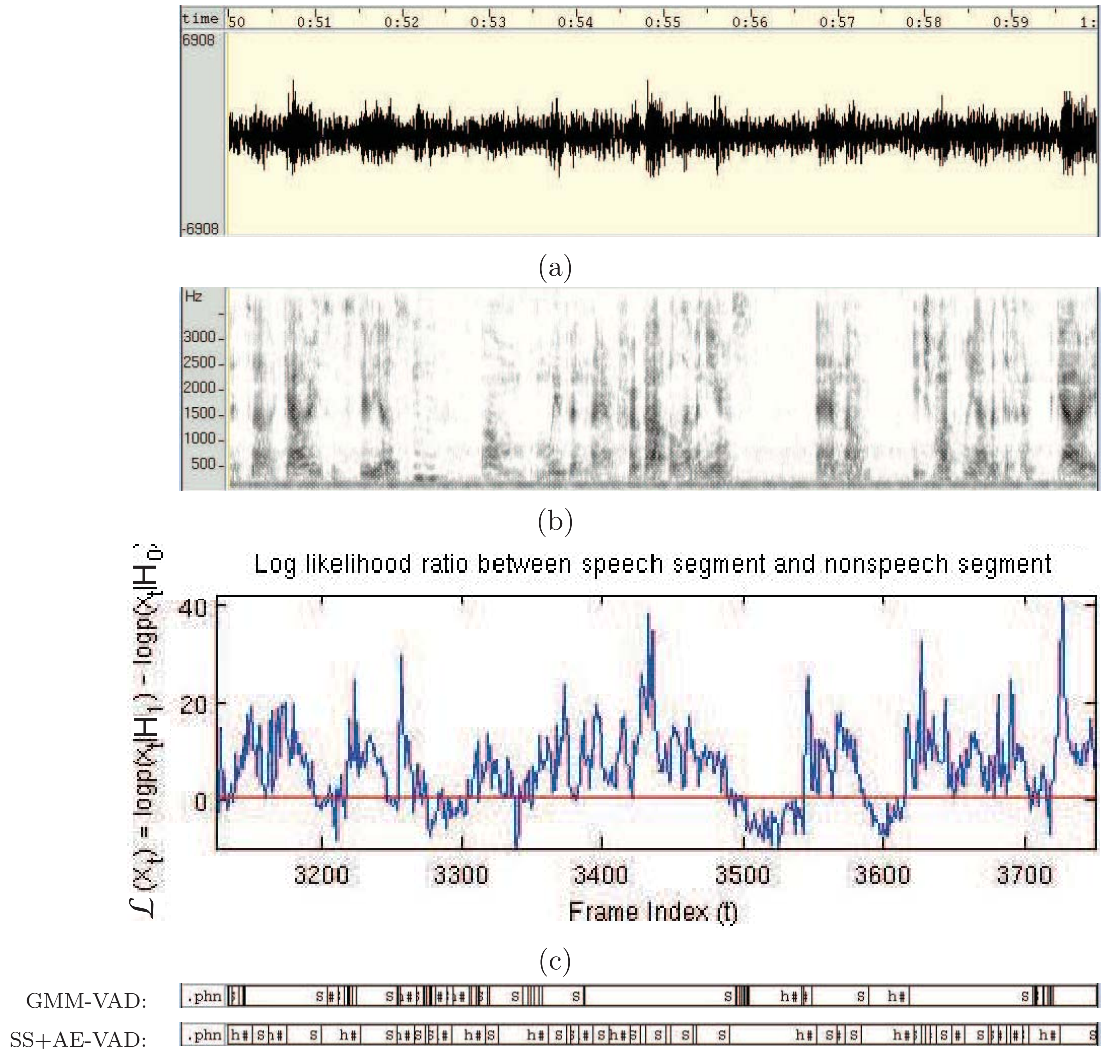


Figure 7.1: Block diagram of the Sadjadi's feature extraction scheme [3].

7.2.2 Threshold Determination for GMM-based VAD

In this dissertation, a fixed decision threshold for GMM-based VAD is determined by a linear combination of the GMM-scores of background frames and signal-peak frames. However, this threshold determination strategy may not be appropriate for GMM-based VAD because overestimation of speech segments may occur in some situations, as exemplified in Fig. 7.2(d). The log likelihood ratios in Fig. 7.2(c) suggests that the speech and non-speech segments are highly distinguishable, however, the fixed threshold cannot make good use of this characteristic. Therefore, a better threshold determination strategy is required for the GMM-based VAD.



(a)

(b)

(c)

(d)

Figure 7.2: (a) A short segment of interview speech in NIST 2010 SRE. (b) The spectrogram of the same segment. (c) The log likelihood ratios (blue curve) of the same segment obtained from GMM-based VAD with a fixed decision threshold (red line). (d) Speech/non-speech decisions (`S` for speech and `h#` for silence) made by GMM-based VAD (GMM-VAD) and the energy-based VAD with spectral subtraction as a pre-processor (SS+AE-VAD).

BIBLIOGRAPHY

[1] ETSI, *Voice activity detector VAD for adaptive multi-rate (AMR) speech traffic channels, ETSI EN 301 708 v7.1.1*, 1999.

[2] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 834–844, October 2010.

[3] S. O. Sadjadi and J. H. L. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *INTERSPEECH-2010*, Makuhari, Japan, Sep. 2010, pp. 2138–2141.

[4] 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, 1999.

[5] D. A. Reynolds, "Automatic speaker recognition using Gaussian mixture speaker models," *Lincoln Laboratory Journal*, vol. 8, no. 2, pp. 173–192, 1995.

[6] "Financial success for biometrics?," *Biometric Technology Today*, vol. 13, no. 4, pp. 9–11, 2005.

[7] "ABN AMRO to roll out speaker verification next term system for telephone banking," *Biometric Technology Today*, vol. 14, no. 7-8, pp. 3–4, July-Aug 2006.

[8] "Speaker verification finds its voice in Australia," *Biometric Technology Today*, vol. 17, no. 6, pp. 4, June 2009.

[9] "T-mobile trials speaker verification," *Biometric Technology Today*, vol. 2009, no. 11, pp. 2–3, Nov-Dec 2009.

[10] T. Kinnunena and H. Z. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[11] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *International Conference on Speech and Computer (SPECOM 2005)*, 2005, vol. 1, pp. 191–194.

[12] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, no. 4, pp. 357–366, August 1980.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[14] J. Ramirez, J. M. Gorriz, and J. C. Segura, *Robust Speech Recognition and Understanding*, chapter Voice activity detection. Fundamentals and speech recognition system robustness, pp. 1–22, I-Tech, Vienna, Austria, June 2007.

[15] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/Fuzzy voice activity detectors," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 85–88, March 2002.

[16] 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*, Glasgow, UK, May 1989, vol. 1, pp. 369–372.

[17] R. Tucker, "Voice activity detection using a periodicity measure," *Communications, Speech and Vision, IEE Proceedings I*, vol. 139, no. 4, pp. 377–380, August 1992.

[18] A. Benyassine, E. Shlomot, and H. Y. 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"*, International Telecommunication Union, September 1997.

[19] R. Chengalvarayan, "Robust energy normalization using speech/non-speech discriminator for German connected digit recognition," in *Proc. EUROSPEECH 1999*, Budapest, Hungary, 1999, pp. 61–64.

[20] K. H. Woo, T. Y. Yang, K. J. Park, and C. Y. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, Jan 2000.

[21] 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, Feb 2002.

[22] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transcations on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, Mar 2001.

[23] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Transcations on Speech and Audio Processing*, vol. 8, no. 4, pp. 478–482, 2000.

[24] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, "VAD techniques for real-time speech transmission on the internet," in *IEEE International Conference on High-Speed Networks and Multimedia Communications*, July 2002, pp. 46–50.

[25] K. Itoh and M. Mizushima, "Environmental noise reduction based on speech/non-speech identification for hearing aids," in *Acoustics, Speech, and Signal Processing (ICASSP'97)*, Munich, Germany, April 1997, vol. 1, pp. 21–24.

[26] L. Rabiner and M. Sambur, "Voiced-unvoiced-silence detection using the itakura LPC distance measure," in *ICASSP*, May 1977, pp. 323–326.

[27] J. C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," in *Eurospeech'91*, 1991, pp. 1371–1374.

[28] P. Pollak, P. Sovka, and J. Uhlir, "The noise suppression system for a car," in *Eurospeech93*, Berlin, Germany, September 1993, pp. 1073–1076.

[29] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoint of isolated utterances," *The Bell Sys. Tech. J.*, vol. 54, no. 2, pp. 297–315, 1975.

[30] ETSI, *Universal Mobile Telecommunication Systems (UMTS); Mandatory Speech Codec speech processing functions, AMR speech codec; Voice Activity Detector VAD, ETSI TS 126 094 V4.00*, 2001-03.

[31] E. Cornu, H. Sheikhzadeh, R. L. Brennan, H. R. Abutalebi, E. C. Y. Tam, P. Iles, and K. W. Wong, "ETSI AMR-2 VAD: evaluation and ultra low-resource implementation," in *Acoustics, Speech, and Signal Processing (ICASSP'03)*, Hong Kong, April 2003.

[32] A. D. L. Torre, J. Ramirez, C. Benitez, J. C. Segura, L. Garcia, and A. J. Rubio, "Noise robust model-based voice activity detection," in *Interspeech*, Pittsburgh, Pennsylvania, 2006, pp. 1954–1957.

[33] J. H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.

[34] J. Ramirez, J. C. Segura, J. M. Gorriz, and L. Garcia, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Trans. Audio Speech Language Processing*, vol. 15, no. 8, pp. 2177–2189, 2007.

[35] ETSI, *Digital cellular telecommunications system (phase 2+); Adaptive multi-rate (AMR) speech transcoding, GSM 06.90 v7.2.0 (draft ETSI EN 301 704)*, 1998.

[36] ETSI, *Digital cellular telecommunications system (phase 2+); Voice activity detector (VAD) for full rate speech traffic channels, GSM 06.32 (ETIS EN 300 965 v7.0.1)*, 1998.

[37] ETSI, *Digital cellular telecommunications system (phase 2+); Voice activity detector (VAD) for full rate speech traffic channels, GSM 06.42 (draft ETSI EN 300 973 v8.0.0)*, 1999.

[38] ETSI, *Digital cellular telecommunications system; Voice activity detector (VAD) for enhanced full rate (EFR) speech traffic channels, GSM 06.82 (ETS 300 730)*, March 1997.

[39] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[40] B. Andrassy, D. Vlaj, and C. Beaugeant, "Recognition performance of the Siemens front-end with and without frame dropping on the Aurora 2 database," in *Eurospeech01*, Aalborg, Denmark, 2001, vol. 1, pp. 193–196.

[41] T. Kinnunen, J. Saastamoinen, V. Hautamaki, M. Vinni, and P. Franti, "Comparing maximum a posteriori vector quantization and Gaussian mixture models in speaker verification," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, Taipei, April 2009, pp. 4229–4232.

[42] V. Hautamaki, M. Tuononen, T. Niemi-Laitinen, and P. Franti, "Improving speaker verification by periodicity based voice activity detection," in *Proc. 12th Int. Conf. Speech and Computer (SPECOM'2007)*, Moscow, October 2007, vol. 2, pp. 645–650.

[43] M. W. Mak and H. B. Yu, "Robust voice activity detection for interview speech in NIST speaker recognition evaluation," in *Proc. APSIPA ASC 2010*, Singapore, 2010, pp. 64–71.

[44] E. Dalmasso, F. Castaldo, P. Laface, D. Colibro, and C. Vair, "Loquendo - politecnico di torino's 2008 NIST speaker recognition evaluation system," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, Taipei, April 2009, pp. 4213–4216.

[45] A. Martin and C. Greenberg, Eds., *NIST SRE10 workshop*, Brno, Czech Republic, June 2010. NIST Multimodal Information Group.

[46] R. Padmanabhan, Sree Hari Krishnan P., and Hema A. Murthy, "A pattern recognition approach to VAD using modified group delay," in *Proc. 14th National conference on Communications*, February 2008, pp. 432–437.

[47] J. R. Deller Jr, J. G. Proakis, and J. H. L. Hansen, *Discrete-time Processing of Speech Signals*, Macmillan Pub. Company, 1993.

[48] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[49] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[50] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International, Inc., 1993.

[51] J. Ramirez, J. C. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, October 2005.

[52] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, Salt Lake City, UT , USA, 2001, vol. 1, pp. 125–128.

[53] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 2, pp. 254–272, 1981.

[54] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 443–445, 1985.

[55] F. Basbug, S. Nandkumar, and K. Swaminathan, "Robust voice activity detection for DTX operation of speech coders," in *IEEE Workshop on Speech Coding*, 1999, pp. 58–60.

[56] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[57] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.

[58] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP'06*, 2006, vol. 1, pp. 97–100.

[59] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Interspeech*, Japan, September 2010, pp. 2726–2729.

[60] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[61] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, 2001, pp. 213–218.

[62] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," *Auditory Physiology and Perception*, pp. 429–446, 1992.

Appendix A

PSEUDO CODE OF THE PROPOSED SS+AE-VAD

```
Input:              $Y(1), \ldots, Y(M)$ - original noisy speech frames
                    $\alpha_{\min}, \alpha_{\max}$ - limits of over-subtraction factor (refer to Eq. 4.22)
                    $\beta_{\min}, \beta_{\max}$ - limits of spectral floor factor (refer to Eq. 4.22)
                    $\nu$ - combination weight (refer to Eq. 4.23)
Local parameter:    $X(1), \ldots, X(M)$ - enhanced speech frames
                    $X(b_1), \ldots, X(b_K)$ - background frames (typically $K = 0.05M$)
                    $X(p_1), \ldots, X(p_L)$ - peak frames (typically $L = 0.01M$)
                    $a(1), \ldots, a(M)$ - amplitude of denoised frames
                    $a(b_1), \ldots, a(b_K)$ - amplitude of background frames
                    $a(p_1), \ldots, a(p_L)$ - amplitude of peak frames
                    $z(1), \ldots, z(M)$ - zero-crossing rate of denoised frames
                    $z(b_1), \ldots, z(b_K)$ - zero-crossing rate of background frames
                    $\eta$ - VAD decision threshold
Output:             $S(1), \ldots, S(N)$ - speech frames, where $N < M$
beginproc
    // Denoise input signal using spectral subtraction, refer to Eq. 4.21
    $[X(1), \ldots, X(M)] = \text{SpectralSubtraction}([Y(1), \ldots, Y(M)], \alpha_{\min}, \alpha_{\max}, \beta_{\min}, \beta_{\max});$
    // Remove DC offset
    $[X(1), \ldots, X(M)] = \text{RemoveDCOffset}([X(1), \ldots, X(M)]);$
    // Find the background frames by searching for $K$ frames with the lowest amplitude
    // among the $M$ frames in the denoised speech
    $[X(b_1), \ldots, X(b_K)] = \text{FindBkgFrames}([X(1), \ldots, X(M)]);$
    // Find the peak frames by searching for $L$ frames with the largest amplitude among
    // the $M$ frames in the denoised speech
    $[X(p_1), \ldots, X(p_L)] = \text{FindPeakFrames}([X(1), \ldots, X(M)]);$
    // Determine VAD threshold $\eta$ based on the mean of background frames and
    // the minimum amplitude of peak frames
    $[a(b_1), \ldots, a(b_K)] = \text{Amplitude}([X(b_1), \ldots, X(b_K)]);$
    $[a(p_1), \ldots, a(p_L)] = \text{Amplitude}([X(p_1), \ldots, X(p_L)]);$
```

```
        $\bar{a}_b = \text{mean}([a(b_1), \ldots, a(b_K)])$;
        $\eta = \nu\bar{a}_b + (1-\nu)\text{min}([a(p_1), \ldots, a(p_L)])$;
        **if** $(\eta == 0 \mid\mid \eta > 0.2 * \text{mean}([a(p_1), \ldots, a(p_L)]))$
          $\eta = 0.2 * \text{mean}([a(p_1), \ldots, a(p_L)])$;
        **endif**
        // *Detect speech frames by comparing the smoothed amplitude of* $[X(1), \ldots, X(M)]$ *with threshold* $\eta$
        // *Consider frames with extremely low zero-crossing rate as non-speech*
        $[a(1), \ldots, a(M)] = \text{Amplitude}([X(1), \ldots, X(M)])$;
        $[a(1), \ldots, a(M)] = \text{MovingAverage}([a(1), \ldots, a(M)])$;
        $[z(1), \ldots, z(M)] = \text{MovingAverage}([z(1), \ldots, z(M)])$;
        $n = 1$;
        **for** $m = 1, \ldots, M$
          **if** $(a(m) > \eta \;\&\&\; z(m) > 0.1 * \text{mean}([z(b_1), \ldots, z(b_K)]))$
            $S(n) = X(m)$;
            $n = n + 1$;
          **endif**
        **endloop**
**endproc** // *End of SS+AE-VAD algorithm*
```

Appendix B

AUTHOR'S PUBLICATIONS

B.1 International Conference Papers

1. M. W. Mak and H. B. Yu, Robust voice activity detection for interview speech in NIST speaker recognition evaluation, in *Proc. APSIPA ASC 2010*, pp. 64-71, December 2010, Singapore.

2. H. B. Yu and M. W. Mak, Comparison of voice activity detectors for interview speech in NIST speaker recognition evaluation, in *Interspeech'11*, pp. 2353-2356, August 2011, Florence.