Nanyang Technological University

Noise Robust Voice Activity Detection

Pham Chau Khoa

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Master of Engineering

2012

Abstract

Voice activity detection (VAD) is a fundamental task in various speech-related applications, such as speech coding, speaker diarization and speech recognition. It is often defined as the problem of distinguishing speech from silence/noise.

A typical VAD system consists of two core parts: a feature extraction and a speech/non-speech decision mechanism. The first part extracts a set of parameters from the signal, which are used by the second part to make the final speech/non-speech decision, based on a set of decision rules. Most VAD features proposed in the literature exploit the discriminative characteristics of speech in different domains, which can be divided into five categories: energy-based features, spectral-domain features, cepstral-domain features, harmonicity-based features, and long-term features. Energy-based features are simple and can be easily implemented in hardware. Spectral-domain and cepstral-domain features are more noise robust at low SNRs, as they are beneficial from a wide class of filtering and speech analysis techniques in these domains. When SNR is around 0 dB, or when the background noise contains complex acoustical events, features relying on the harmonic structure of voiced speech, as well as ones that exploit the long-term variability of speech appear to be more robust. Next, the second part of VAD decides the speech or non-speech class for each signal segment. Existing decision making mechanisms can be divided into three categories: thresholding, statistical modelling and machine learning. The first one is the simplest, yet sufficient in many cases where the features employed possess a good discriminative power. The latter two can work well at high SNRs, but their performance decline quickly at lower SNRs.

In order to derive a state-of-the-art VAD algorithm, a comparative study has been carried out in this thesis to evaluate different VAD techniques. Traditionally, VAD algorithms are evaluated as a holistic system, from which it is hard to analyse whether performance gain is achieved from a new feature or a new decision mechanism. In this report, the author examines the use of P_e, the probability of error of two given distributions, to measure performance of a VAD feature separately from other modules in the system. The metric represents the discriminative power of a feature when used for classifying speech and non-speech. The result is a fairer comparison and a more compact performance representation. This allows a deeper analysis of VAD features, which reveals interesting trends across different SNRs.

Secondly, a new approach to VAD is proposed in this report, which tackles the cases where SNR can be lower than 0 dB and background might contain complex audible events. The proposed idea exploits the sub-regions of the speech noisy spectrum that still retain a sufficient harmonicity structure of the human voiced speech. This allows for a more robust feature, based on the local harmonicity of the spectral autocorrelation of the voiced speech, can be derived to reliably detect the heavily corrupted voiced speech segments. Experimental results showed a significant improvement over a recently proposed method in the same category.

This thesis is dedicated to Kristin,
without whose love and constant reminders
it would still be on my todo list.

Acknowledgments

I would like to express my sincerest gratitude to my supervisor, Dr. Chng Eng Siong, who has offered me the chance to pursuit this research and brought me to the exciting but challenging world of academia. His constant encouragement and mentorship has helped me tremendously not only in solving my research problems, but also in learning various aspects of life in general.

Contents

List of Figures

List of Abbreviations

ACF	Autocorrelation function
ASR	Automatic speech recognition
AUC	Area under the curve
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
GD	Gaussian distribution
GGD	Generalized Gaussian distribution
FFT	Fast Fourier Transform
LD	Laplacian distribution
LDA	Linear discriminant analysis
LDM	Linear dynamic models
LTS	Long-term stationary
LTSV	Long-term signal variability
LRT	Likelihood ratio test
LTSD	Long-term spectral diversion
MFCC	Mel-frequency cepstral coefficients

MMC	Maximum marginal clustering
PDF	Probability distribution function
ROC	Receiver's characteristic curve
SLH	Spectral local harmonicity feature
SNR	Signal-to-noise ratio
SVM	Support vector machine
VAD	Voice activity detection
WALE	Weighted autocorrelation lag energy

Chapter 1

Introduction

In the last three decades, the need for alternative methods of interaction with computer systems has been a great motivation for researchers in the speech processing community around the world to develop robust automatic speech recognition (ASR) algorithms. One major step which affects directly the performance of these systems is the detection of speech from audio stream. For example, too many false alarms, or too many non-speech segments wrongly detected as speech and used in the training can corrupt the acoustic models, and hence reduces recognition accuracy. On the other hand, during testing, if not enough speech segments are detected, the ASR algorithms will not be able to recognize the full spoken sentence.

Speech detection, or voice activity detection (VAD) is the problem of determining the existence of human speech from an audio signal. In a clean signal, or one that has very high signal-to-noise ratio (SNR), this can be solved simply by using an energy threshold. For example, in a very quiet room, human speech has a very high energy level compare to non-speech. Thus, using an energy threshold may be sufficient for detection [1, 2, 3]. However, when the signal is corrupted by noise, it can be very hard to distinguish between speech and non-speech. Depending on the surrounding environment of the recording, non-speech can be silence, noise, music, or a variety of other acoustical signals such as door knocking, coughing, paper shuffling, or even background speech. Some noise types are easy to handle, for example, stationary noise such as those produced by air-conditioners can be filtered out using simple noise reduction and speech enhancement techniques [2, 4, 5]. In more complex environments such as on the streets, in the shopping malls or at the train stations, it can be very difficult to detect the speech signal of interest, since noise level can be even higher than that of target speech [6]. Especially if the background contains babble noise, whose statistical characteristics are very similar to speech, such as in an exhibition hall, the problem is still not satisfactorily solved to date [6, 7].

A key purpose of this research is to tackle the case when SNR is very low (less than 0 dB), and noise is of types such as factory or in a shopping mall. The remainder of this chapter will provide the relevant background on the topic, formulate the problem, and introduce the typical structure of a VAD system.

1.1 Background

VAD is a fundamental task in almost any speech processing system, such as speech coding, speaker diarization and speech recognition. In speech coding [8, 9, 10], VAD helps to avoid the unnecessary coding and transmission of non-speech fragments, thus saving bandwidth and computation costs. In speaker diarization and speech recognition, VAD directly affects the purity of speaker models and, hence, the accuracy of the whole system. VAD has also been used as noise reduction in digital hearing aid devices [11, 12], and in real-time VoIP applications [13].

Several VAD algorithms have been standardized for specific applications. The most commonly mentioned includes ITU-T Recommendation G.729 Annex B [2] and ETSI AMR Option 2 VAD [4], which aim to design simple features that can be implemented in embedded applications efficiently, such as audio recording and transmission on mobile phones. In these systems, speech are often recorded with close-talk microphones, which ensures the sound level of speech is always much higher than background noise. However, this constraint is not applicable in many other applications that requires that the speakers to have the flexibility of not wearing any microphones. For example, in robotic applications [14, 15], the users generally would like to communicate with the robots freely from a distance, just like communicating to other human; or in security applications where audio and video signals are used to monitor crowd activity, the speakers are not speaking closely to the microphones. In these examples, the standard algorithms are not robust.

1.2 Structure of a VAD system

Most of the existing VAD systems follow a specific modular structure [6, 16, 5], as described in Figure 1.1 below:

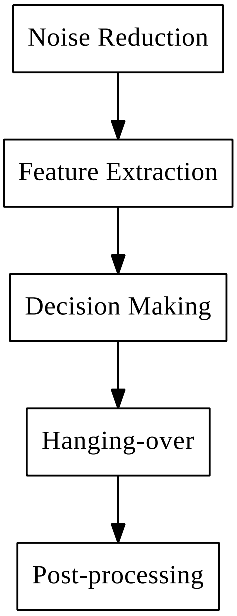


Figure 1.1: Structure of a VAD system

In this structure, an acoustic signal recorded from a microphone is first passed through a noise reduction module, which filters out noise and enhance the SNR of the signal. After that, the next module extracts the acoustic features, or a list of parameters from the signal, which are used in the decision making phase. After that, a hanging-over mechanism, which is often implemented as a finite-state machine, is employed to increase detection hits and reduce false alarms. Finally, some optional post-processing might be carried out to further fine-tune the detection results.

Among these modules, only the second and the third are considered the core of a VAD system, and will be further studied in Chapter 2.

1.3 Motivation

This research is motivated by the belief that under very low SNR, there are still some parts of the speech spectrum that can be used for detection. This can be explained by a visual inspection of the spectrum, illustrated in Figure 1.2.

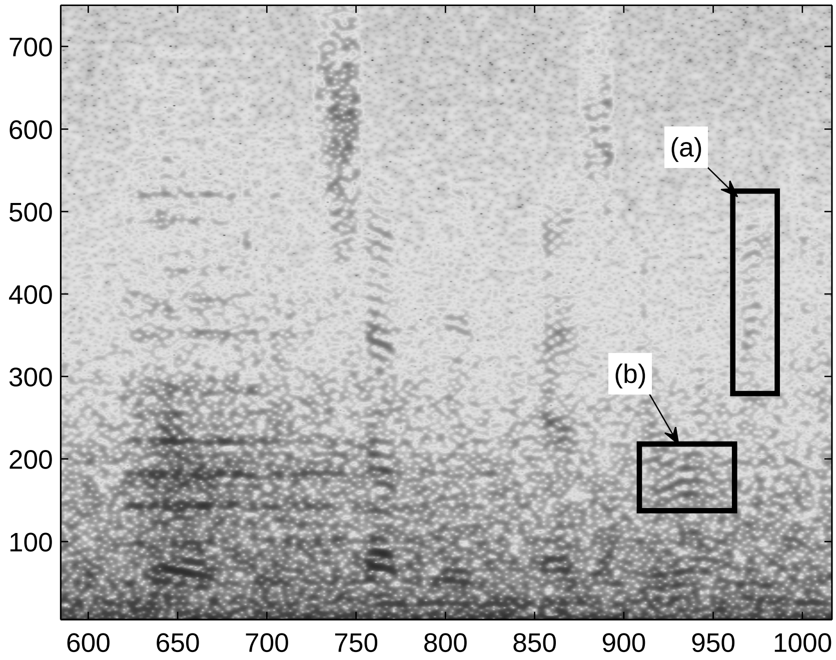


Figure 1.2: A sample speech uttered by a male speaker at −10 dB factory noise.

Take for example in Figure 1.2, the speech frame at index 976 (region (a) in the figure), it can be observed that although the majority of the frame has been corrupted by noise, the subset (a) of it is still visible and can be easily detected by human eyes. Existing VADs are not able to detect this type of corrupted frame, since the contribution of speech energy to the whole frame is insignificant. This work investigates the possibility to discover subframes such as (a) and (b), by exploiting the harmonic pattern from a subset window of the corrupted voiced frame.

1.4 Contributions

In this research, the author has contributed the following:

- A survey of existing algorithms proposed in the literature in the past fifteen years,

- An evaluation metric, P_e, which helps to compare different VAD features under various noise type and SNR configurations,
- A novel approach to detect speech which gives high precision result in very low SNR cases.

1.5 Thesis organization

The remainder of this thesis is organized as follow. Chapter 2 studies the many existing algorithms for VAD. This includes the various features and decision rules proposed in the literature, which are divided into different categories, based on their technical approaches. Their advantages and disadvantages are also analyzed. Chapter 3 then evaluate some selected VAD features, using a newly proposed evaluation metrics, which compactly represents the performance of VAD systems under many different noise configurations. Chapters 4 studies the difficult cases of highly non-stationary noise at very low SNRs, and proposes a novel approach to detect the voiced speech in such environment. Finally, Chapter 5 summarizes and concludes the thesis.

Chapter 2

Literature Survey

As a key module for many speech processing applications [11, 8, 9, 13], VAD has received considerable research interests in the last couple of decades. It started as early as in the 1970s [17], when VAD was often referred to as speech endpoint detection or word boundary detection problem. Back then, VAD algorithms often dealt with only little or no noise corruption in speech coding applications, and with separate recording utterances in speech recognition systems [18]. Up to recently [2, 1, 4], advances in various speech applications require the detection of human speech in a continuous real-time fashion, and is often corrupted by a wide variety classes of noise. Algorithms for VAD had grown accordingly over the years. Unfortunately, given the vast variety of VAD techniques proposed in the literature, to my knowledge, no single work has been found to sufficiently study and compare the state-of-the-art techniques. Examples of recent reviews include: voicing features for VAD were reviewed and compared in [19], VAD algorithms used in transmission standards are compared in [20] and [21], no extensive survey on recent VADs can be found in the literature. Therefore, it is part of this thesis's contributions to provide an up-to-date literature survey for VAD.

The core of any VAD proposed consist of two parts: a 'feature extraction' and a 'speech/ non-speech decision mechanism.' The first part extracts from the given speech signal the parameters that can represent the discriminative characteristics of speech comparing to noise. Using these parameters, the second part makes the final speech/non-speech decision, based on a set of decision rules.

The rest of this chapter reviews these two processes in more details. It first introduces different features proposed in literature, divided into categories; and then, the approaches to speech/non-speech decision mechanism are studied.

2.1 Features for VAD

The first step in any automatic voice activity detection system is the extraction of acoustic features from the signal [6, 22]. Most techniques assume that these features are statistically stationary over the interval of a few milliseconds [6]. Thus, the speech signal is divided into short overlapping segments of fixed duration using a sliding window technique, which are then treated as 'frames' of observation for feature extraction.

The choice of features is a critical design in any classification problem. In the voice activity detection context, good features must possess the following crucial properties:

Discriminative power which measures the separateness between the distributions of noisy speech frames and noise only frames. Theoretically, a good feature should have no overlapping values between speech and noise classes.

Noise robustness which ensures the performance of the classifier in real-world applications, where background noise may corrupt the speech signal greatly and thus reduce the discriminative power of the extracted features.

There have been many features proposed in the literature, which can be broadly categorized as follow:

- Energy-based features
- Spectral-domain features
- Cepstral-domain features
- Harmonicity-based features
- Long-term features

The following subsections reviews each feature category in details.

2.1.1 Energy-based features

Energy is a simple measure of the loudness of the signal. A naïve method for VAD can assume that speech is always louder than background noise, and then assign the high-energy frames to speech and lower ones to noise. However, when the loudness of speech and noise are of similar levels, for example due to the increasing background noise in the environment or during soft speech segments, the simple energy feature fails to discriminate speech and noise. Earlier work on VAD exploited the energies across different sub-bands to increase the discriminative power [23, 2]. For example, examining the spectrums of speech and noise shows that voiced speech has high energy in the low frequency bands (below 2 kHz) and unvoiced speech is more active in the high frequency bands (either 2–4 kHz or above 4 kHz), while white noise spreads its energy equally across the entire spectrum. Another way to improve noise robustness is to combine energy-based features with other features, such as zero-crossing rate (ZCR) [17, 22], or the line spectral frequency (LSF) [2].

Generally, these features work well with clean speech or high SNR conditions. However, under high noise level such as when SNR falls below 10 dB, their discriminative power drops drastically. Nevertheless, with their low computation complexity, energy-based features are still employed by some standards and various real-world applications. A typical example is the ITU-T Recommendation G.729 Annex B [2], which employs a vector of features including full-band energy, zero-crossing rate and low-band (0 to 1 kHz) energy. This standard remains the most cited work and still being used as the baseline system for performance comparison in many research. In [22], energy-based VAD is used as a preliminary event detector for further classification in an always-listening speech application.

2.1.2 Spectral-domain features

One of the most common speech processing techniques is the frequency spectrum analysis, which describes the frequency content of the signal over time [24]. This is made possible by the development of the Fast Fourier Transform (FFT) algorithm, which enables the completion of the Fourier Transform in $\mathcal{O}(n \log n)$, instead of $\mathcal{O}(n^2)$. Let

$X_k = [X_{k,1} \dots X_{k,L}]^T$, $X_k \in \mathbb{C}^L$ be the vector of coefficients resulting from applying the L-point FFT on the k-th frame $\mathbf{x}_k = [x_{k,1} \dots x_{k,N}]^T$, $\mathbf{x}_k \in \mathbb{R}^N$,

$$X_{k,w} = \sum_{n=1}^{N} x_{k,n} e^{-j\frac{2\pi}{N}wn} \quad (1 \leq w \leq L, 1 \leq n \leq N) \tag{2.1}$$

In the voice activity detection context, there have been many features derived from the spectral domain, many of which rely on some form of noise power estimation and subtraction [25]:

$$|\hat{X}_k|^2 = |X_k|^2 - |\hat{N}|^2 \tag{2.2}$$

where $|\hat{X}_k|^2$ is the estimated clean speech power spectrum at the k-th frame, $|X_k|^2$ is the corresponding noisy speech power spectrum, and $|\hat{N}|^2$ is the estimated noise power spectrum, usually calculated as the average spectrum of a sample noise segment. This is a simple yet efficient method to reduce the effects of additive noise in speech signal. It works by assuming that the noise component is additive and independent from speech in the power spectrum domain. Typically, this approach is coupled with the process of noise estimation, which can be achieved through training data or by averaging the long-term spectrum of the signal [5, 26]. The techniques related to noise estimation and spectral subtraction are, in fact, the heart of *speech enhancement* problem, which is a related research field. They are sometimes included in VAD systems either explicitly in a pre-processing step [27], or implicitly in the feature extraction step itself [5, 16].

To increase the discriminative power of the features under noisy condition, many approaches consider the relative power of speech over the estimated noise across the different frequency bands of the spectrum [5]. Effectively, this is equivalent to estimating the sub-band SNR's of the speech signal. In fact, even under very noisy condition (SNR=0 dB, Figure 2.1), it can be observed from the spectrum that the speech signal's harmonics are still distinguishable from noise in some frequency bands.

To exploit this idea, Ramirez *et al.* [5] estimate the long-term upper bounding envelope of the spectrum over a set of $(2M + 1)$ contiguous frames, and then calculate the sum of all L sub-band SNRs, where M is the number of neighbour frames to include in the envelope estimation, and L is the number of FFT coefficients. The measure is called

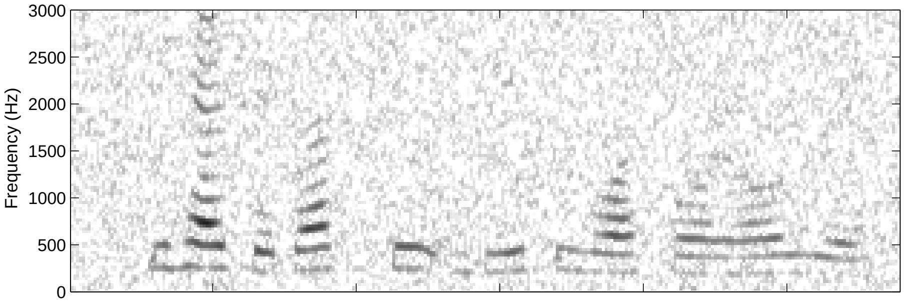


Figure 2.1: A sample spectrum of speech corrupted by white noise at SNR=0 dB

long-term spectral diversion (LTSD), whose formulation is as follow:

$$\text{LTSD}(X_k) = 10\log_{10}\left(\frac{1}{L}\sum_{i=1}^{L}\frac{|\tilde{X}_{k,i}|^2}{|N_i|^2}\right) \tag{2.3}$$

$$\tilde{X}_{k,i} = \max\{X_{k-M,i},\ldots,X_{k,i},\ldots,X_{k+M,i}\} \tag{2.4}$$

where N_i is the i-th coefficient of the average noise spectrum, and $\tilde{X}_k = [\tilde{X}_{k,1}\ldots\tilde{X}_{k,L}]^T$ is the estimated spectral envelope of the neighboring $(2M+1)$ frames (Equation 2.4). The feature works well under many noise types, as evaluated in Chapter 3.

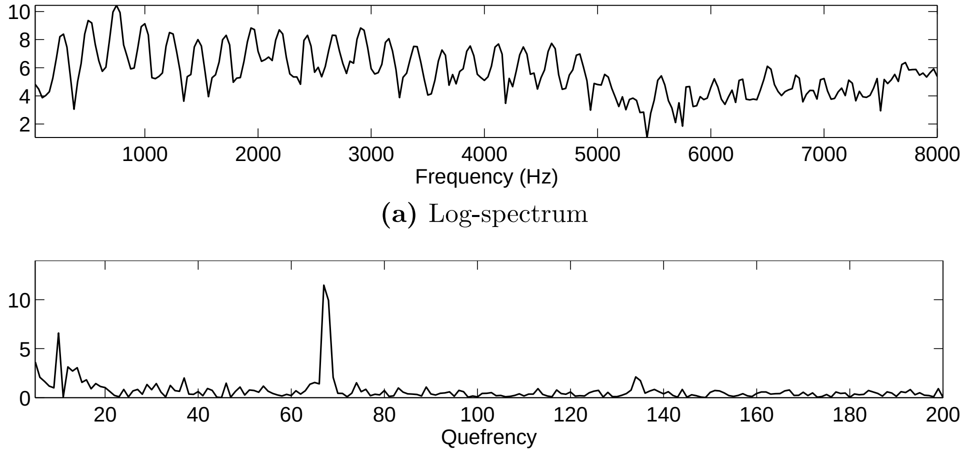
2.1.3 Cepstral-domain features

Another class of features uses a collection of nonlinear techniques known as cepstral analysis. The power cepstrum is defined as follows:

$$\mathbf{c}_k = \left|\text{DFT}\left(\log|X_k|^2\right)\right|^2 \tag{2.5}$$

Here, the power cepstrum can be viewed as the 'power spectrum of the log-power spectrum', which can be used to analyze the power spectrum of the signal. Thus, cepstral features are widely used in speech recognition research [24]. For VAD, the cepstral peaks can be used to determine the fundamental frequency of the speech signal [24, 28, 29] (Figure 2.2). This problem is often referred to as *pitch estimation*, which is another related research field of speech processing and will be discussed further in the next subsection.

Some researchers have used the Mel-frequency cepstral coefficients (MFCC) as the input feature to a supervised classifier for the speech/non speech detection [30, 31, 22]. For



(a) Log-spectrum

(b) Cepstrum

Figure 2.2: Using cepstrum for pitch estimation: (a) the log-spectrum of a sample voiced frame and (b) its cepstrum, showing a distinct spike near index 70 corresponding to its pitch.

example, Kunnunen *et al.* [30] used a feature vector that consists of MFCC and its delta and double-delta coefficients, while Fukuda *et al.* [32] proposed using delta coefficients alone for VAD. Delta cepstrum is defined as the first-order derivative of the cepstral sequence and is sometimes referred to as 'dynamic features' as it captures the dynamic changes between the cepstral frames. The delta cepstrum is defined as follows [32]:

$$\boldsymbol{\Delta}\mathbf{c}_k = \sum_{i=1}^{M} k\left(\mathbf{c}_{k+i} - \mathbf{c}_{k-i}\right) / \left(2\sum_{i=1}^{M} i^2\right) \tag{2.6}$$

In Equation 2.6, a delta window of length $2M+1$ frames is used to extract the delta cepstrum vector $\boldsymbol{\Delta}\mathbf{c}_k$ at time k. In standard ASR systems, the value of M is often set from two to four, depending on the frame size, frame rate and other parameters. The bigger value of M results in the wider temporal information across the consecutive cepstral frames being captured. The effects of such long-term window for feature extraction will be studied more in a later section.

Another technique has been proposed by Fukuda *et al.* [16] to enhance the harmonic structure of the speech by performing filtering in the cepstral domain, a process usually referred to as *liftering.* This method assumes that pitch information containing the harmonic structure is included in the middle-range cepstra. By filtering out the lower and higher cepstra, which are more likely to be corrupted by noise, the harmonic structure of the spectrum is enhanced. The results reported show a great improvement in term of noise robustness in very low SNR situations of 0 dB to -5 dB.

2.1.4 Harmonicity-based features

One of the most powerful features for VAD is based on exploiting the harmonicity of the speech signal [24]. It is well known that voiced speech has a unique structure, which contains multiple harmonics of the fundamental frequency F_0. Unlike unvoiced speech, whose characteristics are easily confusable with such as car, wind and fan noise, the harmonic structure of voiced speech are preserved even in very noisy condition [24, 19]. Observation from Figure 2.1 shows that the most distinguishable harmonics under very noisy environments are around the first formant, which contains the highest energy. However, this is not the case under certain noise types such as engine noise and babble noise (Figure 2.3), where the loudest noisy frequency range might happen to overlap the first formant, as in Figure 2.3 (a). In this case, exploiting the harmonics in other frequency ranges, as illustrated in Figure 2.3 (b), might be more beneficial.

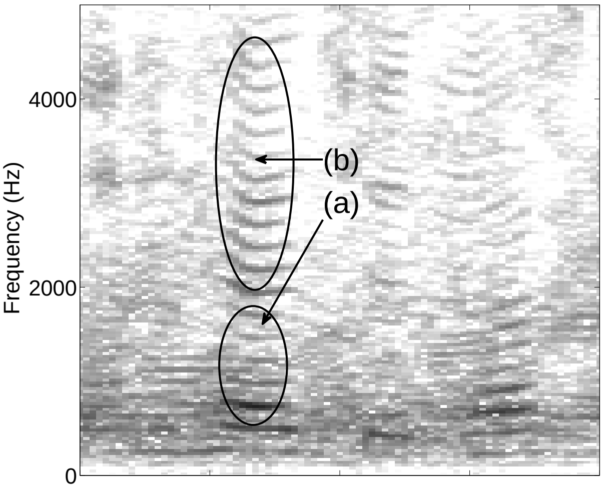


Figure 2.3: A sample spectrum of speech corrupted by babble noise at SNR=-5 dB. While the harmonics around the first formant (a) is destroyed, those in the higher frequency range (b) is still distinguishable.

Most existing algorithms [33, 19, 34] rely on heuristic cues to extract the periodicity feature of speech. This is sometimes tied to the voiced/unvoiced classification and pitch detection problems, which often involves the estimation of the fundamental frequency F_0

[35]. A number of earlier research focused on the auto-correlation of the signal to search for self-repetition components from the signal [33, 19]. This includes the *Maximum Autocorrelation Peak* [36], which finds the magnitude of the maximum peak within the range of lags that correspond to the range of pitch of male and female voices (50Hz–400Hz). Another measure is the *Autocorrelation Peak Count* [34], which counts the number of peaks found in a range of lags. Formally, the autocorrelation vector, $\mathbf{r}_k = [r_{k,1} \dots r_{k,N}]^T$, of the k-th frame $\mathbf{x}_k$, and the estimated pitch, $\hat{f}_0$ are found as follow:

$$r_{k,\tau} = \sum_{m=\tau}^{N} x_{k,m} x_{k,m-\tau} \tag{2.7}$$

$$\tau_{max} = \operatorname*{argmax}_{\tau} r_{k,\tau} \tag{2.8}$$

$$\hat{f}_0 = f_s / \tau_{max} \tag{2.9}$$

where f_s is the sampling frequency. Under environments that contain repetitive noise such as motor and car noise, auto-correlation-based features would fail because the loudest frequency component is usually the noise itself. This leads to the motivation for more noise-robust techniques in the spectral and cepstral domains.

The harmonic structure of the voiced speech appears in the spectrum as a train of spectral peaks, each of which is in multiple of the fundamental frequency. A common approach to capture this structure is by using comb filtering techniques [37] and the likes [35]. A comb filter consists of a series of regularly spaced spikes in the frequency spectrum, causing constructive interference to the speech harmonics and destructive interference to all other frequency components. In practice, comb filters are used to 'search' for the fundamental frequency by setting its lags between the comb spikes to the possible frequencies. A matched frequency would give a high correlation with the signal in the spectral domain.

Under high noise environments, however, comb filter approaches appear to be not reliable. In these cases, noise components usually corrupt most speech harmonics and causing false spikes across the spectrum, which results in the inaccurate pitch estimation and leads to the degradation of VAD performance (Figure 2.3). To make it more robust in high noise environments, it is essential to perform enhancement procedures to restore the spikes of the speech harmonics and reduce those of noise. Ichikawa *et al.* [38] proposed a technique that does not involve the estimation of F_0, named the *Local Peak Enhancement* (LPE). The method operates in the cepstral domain, which filters out the cepstra that are more likely to belong to noise (upper and lower range) and reserves those that cover the possible harmonic structures in the human voice (middle range). According to the authors the peak enhancement can achieve its optimal performance by combining with noise reduction procedures such as Wiener Filter [39] and spectral subtraction techniques [25].

Most harmonicity-based techniques mentioned above would suffer in environments containing cross-talk speech, such as babble noise or overlapped speaking. Indeed, the interference of other speakers causes the spectral frames of voiced speech to possibly contain more than one fundamental frequencies, which causes confusion to the estimator. To address this problem, another measure in the spectral domain was proposed by Krishnamachari *et al.* [40], named the *Spectral Autocorrelation Peak Valley Ratio* (SAPVR), which takes the autocorrelation of the magnitude spectrum and uses the ratio between the peak and valley of the first local maximum as the harmonicity measure. By finding the local maximum peak in the spectral autocorrelation domain, the method effectively removes the spectral peaks caused by the overlapped speech. However, the proposed technique was only studied in a clean co-channel with overlapping speech from two speakers; there was no experimental results for noisy environments such as babble noise.

One of the few statistical approaches to exploit the harmonicity of voiced speech for speech detection is the spectral entropy [40, 41, 42, 34]. By interpreting the short-time spectrum X_k as discrete random variable, its entropy h_k, according to Shannon's information theory, can measure the randomness of the frame. Due to the harmonic structure of voiced speech, it is expected that the voiced speech will have lower entropy than that of noise and unvoiced frame. Shen *et al.* [41] first proposed an entropy measure for VAD using a set of pre-trained weight factors for adjusting the different contribution of the frequencies across the spectrum:

$$p(X_{k,i}) = X_{k,i} / \sum_{j=1}^{L} X_{k,j} \tag{2.10}$$

$$h_k = -\sum_{i=1}^{L} w_i p(X_{k,i}) \log p(X_{k,i}) \tag{2.11}$$

where $p(X_{k,i})$ is a probability measure of spectral distribution of the i-th frequency bin of the k-th spectral frame; and w_i is the weight factor of the i-th frequency bin, statistically pre-estimated from training data. According to Huang and Yang [42], using spectral entropy feature alone cannot distinguish speech from periodic background noise such as babble noise and music. The authors proposed to combine this feature with the energy measure as a compromise in cases where background noise have strong harmonic structure but its energy is still less than that of speech:

$$\text{EE-Feature}(\mathbf{x}_k) = \sqrt{1 + |(\text{Energy}(\mathbf{x}_k) - E_0)(h_k - h_0)|} \tag{2.12}$$

where E_0 and h_0 are the estimated energy and entropy of recent non-speech frames, which are subtracted from the current frame measures to reduce the effects of noise. Reported result showed an improvement over using energy or entropy feature alone. Unfortunately, it was only tested with recordings at SNR=15 dB; no further experimental result has been used to confirm its robustness under lower SNR conditions. Nevertheless, this feature combination has a reasonably low complexity, which can be easily implemented in hardware applications, such as in-car entertainment systems, and its robustness can be supported by using close-talk microphones to achieve a sufficient SNR level.

2.1.5 Long-term features

Human speech is commonly represented as a non-stationary signal. A person with average speaking rate produces approximately 10–15 phonemes per second [43], each of which has different spectral distribution, causing the statistics of speech to vary greatly over time. On the other hand, most noise encountered in everyday conversation are stationary (white noise, machinery noise), or at least their degrees of variation are much lower than speech's. This suggests that the analysis over a longer window for exploiting the degree of non-stationarity of the signal might be beneficial for distinguishing speech from noise.

Long-term temporal information has been the focus of research in the speech community recently [44, 45, 46]. Psychophysical study [47] also showed that temporal information from both short windows (20–40ms) and long windows (150–250ms) are important to understand spoken language. For VAD, most techniques exploit the long-term information by either extending the processing windows [5] or performing post-processing over the features extracted from a set of contiguous frames [32, 7]. Ramirez *et al.* [5] finds the long-term spectral envelope, from which a measure is derived from the sub-band SNRs (Equations 2.3 and 2.4). Fukuda *et al.* [32, 45, 16] proposed the long-term dynamic feature for VAD, derived from fusing the delta cepstrum of the $2K + 1$ neighbor frames (Equation 2.6). Ghosh *et al.* [7] calculate the sample variance of the long-term sub-band entropies, showing a significant improvement in both stationary and non-stationary noise conditions.

2.1.6 Summary

This section has reviewed several main classes of features that have been proposed in the literature. Each of these classes of features have shown certain advantages and disadvantages when applied to the problem of voice activity detection, in which the main goal is to maximize the discriminative power and noise robustness. Energy-based features is simple and can be easily implemented in hardware applications, but they are not noise robust. Domain-specific features such as those derived from the spectrum and cepstrum are beneficial from a wide class of filtering techniques to reduce the effects of noise. Under very low SNR conditions such as below 0 dB, however, most statistical information in the spectral and cepstral domains are badly affected. Finally, including the long-term temporal information of the signal can exploit the different degrees of variability of speech and noise, which results in an improved discriminative power of the features.

2.2 Decision rules for VAD

After a set of feature vectors is extracted from the original signal, the next step is to determine their categories (speech/non-speech). In a pattern classification problem such as VAD, a decision rule is achieved by defining a set of *decision boundaries* that partition the feature space into different regions, each of which is assigned to a single class.

In the past couple of decades, there has been many techniques proposed for VAD in the literature, which can be categorized as the following:

- Thresholding
- Statistical modelling approaches

- Machine learning approaches

The remaining of this section will review these techniques in more details.

2.2.1 Thresholding

Thresholding is the simplest form of decision boundary in which a line (or a set of lines) is used to define the regions for each classes in the feature space. If the extracted feature at frame k-th is a scalar $y_k \in \mathbb{R}$, a single threshold η is used to divide the $\mathbb{R}$ space into two part. We assign the frames that have $y_k \geq \eta$ to the speech class, and those with $y_k < \eta$ to the non-speech classs [48, 7], as shown in Equation 2.13. In cases where each feature is a vector $\mathbf{y}_k \in \mathbb{R}^N$, a vector of thresholds $\boldsymbol{\eta}$ can be used to divide each dimension of $\mathbb{R}^N$ in a similar manner [49, 5, 23].

$$y_k \underset{noise}{\overset{speech}{\gtrless}} \eta \tag{2.13}$$

Some papers employed more than one threshold to detect both activation and deactivation transitions [42, 17]. A threshold η_1 represents the maximum non-speech parameter values and detects the non-speech-to-speech transitions (or the onset transitions); while another threshold η_2 represents the minimum speech parameter values and detect the speech-to-non-speech transitions (or the offset transitions). The duration between these two transitions is then assigned to be speech.

The thresholds are often pre-estimated empirically, based on the distribution of the extracted features in the training data set [49, 35]. In applications where noise condition keeps changing, however, a fixed value of theshold won't work well [7]. Thus, adaptive threshold schemes are needed. Ramirez *et al.* [5] and Evangelopoulos *et al.* [50] employed a linear calibration curve to adjust the threshold η as a function of signal energy E, whose values vary between E_0 and E_1, the minimum and maximum energies measured from the training data. If the threshold estimated at these two extreme conditions are η_0 and η_1 respectively, the threshold for a certain energy E is given as follow:

$$\eta = \begin{cases} \eta_0 & E \leq E_0 \\ \frac{\eta_0 - \eta_1}{E_0 - E_1} E + \eta_0 - \frac{\eta_0 - \eta_1}{1 - E_1/E_0} & E_0 < E < E_1 \\ \eta_1 & E \geq E_1 \end{cases} \tag{2.14}$$

Most other research [21] employed a running average scheme in which a learning factor α is used to control the threshold updating rate. Given a new noise parameter n_k at frame k-th, the threshold representing the onset transition is updated as:

$$\hat{\eta} = \alpha\eta + (1-\alpha)n_k \tag{2.15}$$

In this equation, a smaller α causes the threshold to be more sensitive to the noise fluctuation, which can be beneficial in situations where the background noise constantly changes. On the other hand, a larger α causes the threshold to be more conservative to changes, and might be desirable in conditions that have infrequent changes of background. Ghosh *et al.* [7] proposed a modified version of Equation 2.15:

$$\hat{\eta} = \alpha \min(\mathcal{Y}) + (1-\alpha)\max(\mathcal{N}) \tag{2.16}$$

Here, the threshold is updated based on the newest speech and noise values extracted from the audio stream, which are stored in $\mathcal{Y}$ and $\mathcal{N}$, respectively. The learning factor α is used to control the learning rate in the same manner. This seems to be a better method, as it takes into account the fact that the feature values for speech and noise has a certain overlapped region, which is marked by the minimum of speech values and the maximum of noise values – assuming that speech values are higher than noise in the mean sense.

If the features extracted from the previous section has good discriminative power and its feature space is linearly separable, using thresholds might be sufficient. However, when the speech signal is corrupted by noise, its linear separability is decreased. To some certain degree of noise, the linear decision boundaries can no longer separate the class regions. Therefore, nonlinear methods such as the statistical model-based and machine learning-based approaches introduced below are preferred.

2.2.2 Statistical modelling approaches

A statistical decision rule deriving from the generalized likelihood ratio test (LRT) was first proposed in [49]. In this method, noisy speech and background noise are assumed to be independent Gaussian random processes. Thus, their DFT coefficients can be modelled as Gaussian random variables that are asymptotically independent. The method considers two hypotheses, H_N and H_S, for non-speech and speech, respectively. Given the k-th spectral frame $X_k = [X_{k,1} \dots X_{k,L}]^T$, the probability density functions conditioned on each hypothesis are computed as:

$$p(X_k|H_N) = \prod_{i=1}^{L} \frac{1}{\pi\lambda_{N,i}} \exp\left(-\frac{|X_{k,i}|^2}{\lambda_{N,i}}\right) \tag{2.17}$$

$$p(X_k|H_S) = \prod_{i=1}^{L} \frac{1}{\pi\left(\lambda_{N,i} + \lambda_{S,i}\right)} \exp\left(-\frac{|X_{k,i}|^2}{\lambda_{N,i} + \lambda_{S,i}}\right) \tag{2.18}$$

where i indicates the spectral bin index, and $\lambda_N = [\lambda_{N,1} \dots \lambda_{N,L}]^T$ and $\lambda_S = [\lambda_{S,1} \dots \lambda_{S,L}]^T$ denote the variances of the estimated spectrums of noise and speech, respectively. These parameters can be obtained by applying the noise estimation and noise subtraction techniques on the training data set. After that, the likelihood ratio for the i-th frequency band Λ_i, and the final decision statistic Λ is obtained as follow:

$$\Lambda_i = \frac{p(X_{k,i}|H_S)}{p(X_{k,i}|H_N)} \tag{2.19}$$

$$\log\Lambda = \frac{1}{L}\sum_{i=1}^{L} \log\Lambda_i \underset{H_N}{\overset{H_S}{\gtrless}} \eta \tag{2.20}$$

In [51], the author pointed out that this test is biased towards H_S, due to the left-hand side of Equation 2.20 being a positive value. A new method is thus proposed which uses the decision-directed (DD) method [52] to reduce the fluctuation of the estimated likelihood ratios during noise-only periods. Later, Cho *et al.* [53] identified the problem of this method [52] having frequent detection errors at the speech to non-speech transition regions, due to the delayed noise variance term in the DD estimator for calculating the *a priori* SNR. To overcome this problem, the authors proposed the smoothed LRT method, which reduced the abrupt changes of the likelihood ratio at the speech to non-speech transition regions and thus improved the performance. In [54], Ramirez *et al.* extended the method even further by incorporating multiple observation from the neighbouring frames. The proposed method, dubbed the multiple observation likelihood ratio test (MO-LRT), computes the joint probabilities conditioned on H_N and H_S of the $2m + 1$ contiguous frames, including the past and future m frames, and tests them against a threshold. It was reported that the method improved the decision robustness against acoustic noise present in the environment. However, by considering the m-frame neighbourhood, this test implicitly introduces a non-controllable hangover mechanism, as pointed out by Ramirez *et al.* [55]. The authors then proposed a revised method that tests all the 2^{2m+1} possible hypotheses defined over the multiple observation vectors, thus eliminates the implicit hangover and leaves rooms for further improvement.

Conventionally, it is assumed that the distribution of clean speech and noise spectra can be modelled by the Gaussian densities (GD, Equation 2.21) [49, 51, 53, 54, 55]. In order to improve the performance of the statistical modelling VAD techniques, several researchers tried to find a model that fits the distribution of speech better. Martin [56] reported that the Laplacian and Gamma distributions (LD, Equation 2.22, and γD, Equation 2.23, respectively) can better model the DFT coefficients of the clean speech and noise, respectively. Instead of separating clean speech and noise in different distributions, Chang *et al.* [57, 58] later proposed to use the complex Laplacian distribution [57] and the generalized Gaussian distribution (GGD, Equation 2.24) [58] to model the DFT coefficients of the noisy speech directly and reported improved results. To compare the fitness of each distribution for speech modelling, Gazor and Zhang [59] performed the Komogorov-Smirmov test [60] on the various widely known speech distributions across different domains. The results showed that speech signal during active intervals can be best modelled by a Laplacian distribution in the time domain, as well as in various decorrelated domains such as the Karhunen-Loeve Transform (KLT) and Discrete Cosine Transform (DCT), while the edge frames of the speech segments favor a gamma distribution.

$$\text{GD:}\quad f_{\mathbf{x}}^{g}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{2.21}$$

$$\text{LD:}\quad f_{\mathbf{x}}^{L}(x) = \frac{1}{2a} \exp\left(-\frac{|x|}{a}\right) \tag{2.22}$$

$$\gamma\text{D:}\quad f_{\mathbf{x}}^{\gamma}(x) = |x|^{h} \exp\left(-\frac{|x|}{a}\right) \frac{1}{2h!a^{h+1}} \tag{2.23}$$

$$\text{GGD:}\quad f_{\mathbf{x}}^{G}(x) = \frac{\upsilon\alpha(\upsilon)}{2\sigma\Gamma(1/\upsilon)} \exp\left(-\left[\alpha(\upsilon)\frac{|x|}{\sigma}\right]^{\upsilon}\right) \tag{2.24}$$

$$\text{G}\Gamma\text{D:}\quad f_{\mathbf{x}}^{\Gamma}(x) = \frac{\gamma\beta^{\eta}}{2\Gamma(\eta)}|x|^{\eta\gamma-1} \exp\left(-\beta|x|^{\gamma}\right) \tag{2.25}$$

As speech signals are often affected by different speaker-generated random factor, such as the speaker's temper, emotion, health, gender, age, etc., recent works on the distribution of speech [61, 62, 63] shows that using a more generic model that can be adapted to different distributions is preferred. Recently, Shin *et al.* [61] proposed the Two-sided Generalized Gamma Distribution (GΓD, Equation 2.25) for modelling the speech spectra. Using a similar test as mentioned above, the authors showed that GΓD can model the distribution of the speech spectra more accurately compared to the previously mentioned distributions. Due to its highly generic form, GΓD can be configured to cover different distributions. For example, when $\gamma = 1$, GΓD defines a gamma pdf; when $\eta\gamma = 1$, it becomes a GGD. If $\gamma = 2$ and $\eta = 0.5$, it represents the GD; and if $\gamma = 1$ and $\eta = 1$ the LD [61].

Chang *et al.* [64] extends the concept by introducing a switching mechanism to automatically chooses a suitable model that best fits the speech distribution. For every K frames, the proposed technique performs an online Komogorov-Smirmov test on a set of preferred distributions and adopts the best one to use in a LRT-based decision rule for the next frames. Towards this direction, Petsatodis *et al.* [63] suggested to use not only the auto-switching mechanism to choose different statistical models for different situations, but also to combine multiple distributions for modelling the speech signal in each frame. This can be done using a convex combination of multiple models [63], whose influence on the final speech model at any given time is determined by a set of weights. The outcome showed an improve modelling capability, which better captures the statistic of speech even under adverse noise conditions and reverberation effects.

2.2.3 Machine learning approaches

Recently, there has been a large interest [65]–[66] in employing machine learning techniques as the decision rules for VAD. Many methods have been proposed in the last decade such as Support Vector Machine, Maximum Marginal Clustering, Neural Networks, Linear Discriminant Analysis, Boosting algorithms and Genetic algorithms. This section gives an overview of the machine learning techniques have been applied to VAD, their advantages and disadvantages are also discussed.

Support Vector Machine (SVM) is a non-probabilistic binary classifier, which has been used successfully in many computational problems such as handwritten character recognition, face detection, image classification, speaker verification [67]. It aims to construct a hyperplane in the feature space that maximize the margin between the two classes. In [65], SVM was applied directly to VAD, leveraging the same set of features recommended by the G.729B standard. The authors reported 4% absolute improvement of detection accuracy. In a similar manner, Ramirez *et al.* [68, 69] applied SVM to the LTSD features [5], and Kinnunen *et al.* [30] the MFCC features. Both have reported improved VAD performance comparing to other decision rules on the same feature set. These results suggest that the marginal performance was due to SVM alone.

One weakness of SVM is that it is sensitive to noise. In theory, if the training data for all noisy conditions are available and match to test conditions, the SVM can lead to minimal classification error. In practice, however, this is impossible under noisy scenarios, causing SVM performance to degrade [70]. The Maximum Marginal Clustering (MMC) technique [71, 72] extends the idea of SVM, but remove the dependency on training labels by finding not only the maximum margin hyperplane in the feature space, but also the optimal label vector that maximizes the margin among all possible label vectors [71]. Recently, Wu and Zhang [70] have proposed using MMC as the decision rule for VAD. Although the results showed little improvement to existing approaches using SVM, it has no dependency on the training labels, and is therefore preferred to SVM.

Martin *et al.* [31, 73] proposed the use of Linear Discriminant Analysis (LDA) for VAD. Intuitively, LDA aims to find a linear combination of features from the feature vector, which best characterizes the speech events. Effectively, LDA acts as both a dimensionality reduction process and a fusing technique, which can be beneficial when multiple feature vectors are combined. For example, Martin *et al.* [31] employed LDA on MFCC+energy features, while Padrell *et al.* [74] on spectral-based features. Unfortunately, their results were only tested with the speech coder standards (GSM, AFE), and no recent VAD techniques were included in their comparison.

There were several other machine learning techniques proposed in the literature for VAD. Kwon and Lee [75] used AdaBoosting algorithm [76] to combine a set of weak classifiers to generate a stronger one. Boosting is a successful technique in several other applications such as face detection [77] and music classification [78], which can achieve high accuracy while preserving a low complexity. Usukura and Mitsuhashi [79] applied AdaBoost on both long-term and short-term features and showed superior results comparing to the statistical model approach [51]. Other natural inspired machine learning techniques such as neural networks [80, 81] and genetic algorithms [82, 66] were also proposed for VAD decision making. All of these works reported improved VAD performance over the G.729B system [2], but did not carry out any extensive comparison on modern VAD techniques [5, 6]. Their actual performance comparing to state-of-the-art VAD systems remains unanswered.

2.2.4 Summary

This section has reviewed many decision rules for VAD proposed in the literature, which can be categorized into three classes: thresholding, statistical modelling and machine learning approaches. The thresholding decision rules appeared to be the most commonly used technique, due to their simplicity and low complexity. These techniques often require the features extracted to be sufficiently discriminative and noise robust. However, when SNR drops, the feature space is no longer linearly separable, causing the performance to degrade drastically. More sophisticated non-linear approaches are therefore required.

In the statistical modelling approaches, many researchers have worked extensively to find a distribution that best captures the characteristics of the speech signal. The Two-sided Generalized Gamma Distribution has been shown to be a good model for speech in multiple domains such as time-domain, KLT and DCT, due to its highly generic form. Others have proposed to employ multiple distributions to model speech, either separately in an online switching machine, or together in a mixture of models. However, these models were studied using the entire frame spectrum. Their application to other features such as sub-band energies, harmonicity or long-term features is still not clear.

Finally, many machine learning approaches were proposed for VAD in the last decade. Among these methods, SVM and MMC appear to be the most studied and have the best performance. Other techniques such as LDA, boosting, neural networks and genetic algorithms were proposed and showed improved performance over the standard VADs [2, 4]. However, their performance against more modern VAD systems are not measured in these works [5, 6, 7].

2.3 Conclusion

As a key module in many speech processing applications, VAD has received considerable research interests recently. This chapter has provided an up-to-date literature survey of VAD research in the past fifteen years. In can be concluded from this survey that existing VAD features and decision rules can work sufficiently well under high SNRs, or when the background noise is relatively stationary. When SNR drops to lower than 5 dB, or when noise contains complex audible events, such as babble noise in a restaurant or in a factory, existing VADs cannot yet give reliable results. Therefore, it is part of this thesis's motivations to improve VAD performance in these cases.

Chapter 3

Evaluation of VAD Features

Given the numerous methods proposed for VAD in the literature, it is clear that a proper bench-marking strategy is important for their evaluation. Since VAD consists of multiple modules, in order to reveal the performance gain contributed by different algorithms for each module, a fair evaluation should consider comparing each module independently. For example, comparing two VAD algorithms that use different features and decision rules would not reveal the insights of whether the performance improvement was achieved by the new feature or decision rule.

Another key requirement with evaluating VAD algorithms is to measure the performance under different noise types and SNR conditions. This helps to reveal the environments with which a VAD algorithm can work best, since there are VAD algorithms purposely designed to work only on a certain application. The traditional way of presenting these results are not compact. For example, a thorough comparison of results on N different noise types at K different SNR levels will result in $N \times K$ different configurations. Using the traditional receiver's characteristic curves (ROC) for each configuration requires $N \times K$ figures to be shown and analyzed, which is just not practical. On the other hand, averaging the results on all configurations hides the potential insights on the design of each algorithm and might lead to shallower analysis.

This chapter deals with both problems stated when evaluating existing VAD features. The discriminativeness of the features are compared separately from decision making and other modules, using P_e, the distribution overlapping area, as the performance metric, which compactly represents the discriminative power of the features for each noise configuration. The remainder of the chapter is organized as follow: The first section prepares the data for the experiments by choosing a proper dataset, extracting reference labels, and adding noise artificially. In Section 3.2, the performance metric is introduced and then tested through an experiment described in Section 3.3. Finally, conclusions are provided in Section 3.4.

3.1 Data Preparation

3.1.1 Dataset

There are many speech copora used for VAD evaluation in the literature. The choice of dataset is sometimes dependent on the speaking language of the authors and the nature of the research project they had work on. For example, Ramirez *et al.* tends to use Spanish datasets such as the SpeechDat-Car [83]; while Ishizuka *et al.* prefers the AURORA-2J Japanese dataset [84]. For English language, TIMIT [85] and AURORA-2 [86] databases are among the most commonly used. For the evaluation of VAD features, TIMIT is preferred since it provides manual transcription down to word and phoneme-level. This is tremendously helpful when evaluating the distribution of the features, assuming that the manual transcription is of acceptable accuracy. For example, word-level transcription can be used to plot the distribution of speech against non-speech, and phoneme-level transcription for voiced speech against unvoiced speech plus noise.

3.1.2 Reference labels

A preliminary examination of TIMIT transcription showed that they were not perfectly labelled. For example, segment (b) in Figure 3.1 was labelled with speech, but its energy is insignificant and is close to silence energy profile. Such phenomenon can sometimes cover more than 10% the total number of frames of an utterance, for eg. utterance FJEM0/SA2, which may bring a noticable disturbance to the purity of the features being compared. On the other hand, the energy measure alone cannot determine the noise segments made by the speakers such as breath and click sounds (Figure 3.1). Thus, a combination of both manual label and energy is proposed to estimate the ground truth of each speech utterance.

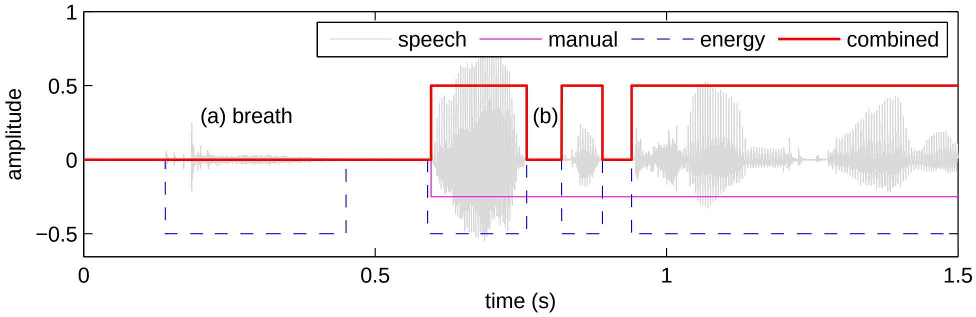


Figure 3.1: Combining manual labels and energy-based VAD to determine referenced labels for TIMIT utterences. Manual label helps to remove breath regions (a), while energy-based VAD helps to remove very low energy regions (b). (Utterance FECD0/SI788)

Another problem of TIMIT dataset is that the utterances are short (about 3.5s), most of which are speech, which may introduce a bias when comparing the distributions of speech and noise. To reduce this effect and make it closer to real-world scenarios, silence frames are added at the beginning and ending such that it covers 60% of the total utterance length. After that, all the ten[1] utterances of each speaker are joined together, giving a long utterance of about 60s per speaker.

3.1.3 Noise addition

To compare the discriminative power of VAD features, noise signals from the NOISEX-92 database [87] are added artificially to each speech utterance at different SNRs. This process is described as follow:

(i) Calculate the average power of speech signal, P_x, at the normalized speech frames.

(ii) Calculate the average power of normalized noise frames, P_n.

(iii) Calculate the desired gain for speech, $g = \sqrt{\frac{\hat{P}_x}{P_x}}$, where the desired speech power is: $\hat{P}_x = 10^{SNR/10} P_n$.

(iv) Add noise to the clean signal: $\hat{x} = gx + n$.

Note that it is important to ensure a constant noise level across all utterances by applying a gain to the speech signal instead of the noise signal. This is to make sure the distribution of noise for gain-sensitive features the same in all utterances.

[1] In TIMIT database, each speaker has 10 utterances.

3.2 Evaluation metrics

This section reviews the existing evaluation metrics in the literature, analyses their weaknesses, and introduce a metric to evaluate VAD feature. There are currently several metrics used in the literature for VAD evaluation, which often come in pairs: precision and recall, speech detection rate and nonspeech detection rate, and the VAD performance parameter set proposed by Freeman *et al.* [8] and Beritelli *et al.* [9].

VAD is a two class classification problem in which there exists trade-offs between the different metrics. Depending on different applications, improvement of one metric may be more desirable, even at a reasonable decrement of the other metric. For example, *precision* and *recall* can be used to measure VAD performance [88]. Precision measures the rate of accurate detection among the detected speech frames, while recall measures the amount of speech frames detected. Typically, when the possible parameters of a VAD algorithm is tuned to improve speech detection precision, its recall metric is worsen, and vice versa. This pair of metrics can be used in various application when knowledge of the relative cost of misdetection and incorrect detection is known. For example, for the speech recognition application, it is crucial to have most of the spoken speech frames detected so that the whole sentence can be recognized. However, in a speaker diarization and speaker identification problem, precise detection is preferred to enhance the purity of speaker models [89]. In other applications where the relative cost of speech detection and *non*-speech detection is more important, the *sensitivity* and *specificity* metric pair is often used [5]. Sensitivity is the same as recall, which measure the true positive detection rate or hit rate, while specificity measure the true negative detection rate.

A common approach to determine the performance for VAD while using these pairs of metrics is via the Receiver Operating Characteristics (ROC) curve [5, 6, 51], in which one's aim is to maximize the area under the curve (AUC). This evaluation strategy shows clearly the trade-off in performance of different VAD algorithms, showing which settings each algorithm can perform best. However, it is not a compact metric form for benchmarking the algorithms in multiple noise conditions. For example, to compare the VAD performance in 8 different noise types, at 7 common SNRs, from 20 dB to -10 dB, one needs to show 56 different ROC curves, which can be overwhelming to draw a useful comparison.

Beritelli *et al.* [9] proposed a set of parameters for evaluating VAD performance, as listed below:

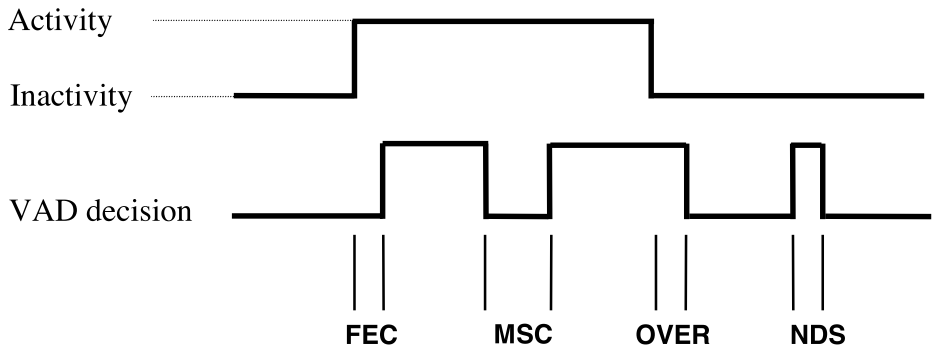


Figure 3.2: VAD performance parameter set proposed by Beritelli *et al.*. Image reproduced from [9]

.

- **CORRECT**: Correct decisions made by the VAD.
- **FEC** (Front end clipping): Clipping due to speech being misclassified as nonspeech in the onset regions.
- **MSC** (Mid speech clipping): Clipping due to speech being misclassified as nonspeech in the speech regions.
- **OVER** (Carry over): Nonspeech interpreted as speech in the offset regions.
- **NDS** (Noise detected as speech): Nonspeech interpreted as speech in the nonspeech regions.

FEC and MSC are performance indicators similar to recall, while OVER and NDS are the break-down of precision. This break-down of evaluation metrics are helpful in analysing the performance of a VAD algorithm in each class of decision regions. For bench-marking many algorithms under various noise conditions, however, the number of parameters become too many to draw useful conclusions. Averaging each parameter over all SNRs is a common solution for this, but this will lose interesting performance patterns across the different SNRs. For example, one algorithm might be designed to perform best at SNR$>$ 5 dB, which is the case for many in-door speech applications. Comparing the average performance will not reveal such nature.

3.2.1 Proposing metrics

This section examines a new performance metric for VAD features. It is designed to be compact so that it can easily reveal the performance patterns across many noise conditions. Keeping in mind that VAD consists of two separate modules, the proposing metric only focuses on the performance of the feature extraction process, i.e. the discriminative power of the extracted feature across the different SNRs.

The discriminative power of a VAD feature can be measured by the separateness of its distributions for noise and speech. Existing research [5, 7] only showed the plots of these two distributions for each feature, no formal measure was used to compare the separateness across different features. In this chapter, the author proposes to use the *histogram intersection* as a distance measurement between the two distributions. The idea is to measure the overlapping area of the two normalized histograms, which approximate the PDFs of the distributions of the feature values for speech and noise, as illustrated in Figure 3.3. Numerically, the histogram intersection distance, $d_{Intersect}$, of two histograms $H_X(x)$ and $H_N(n)$ of speech and nonspeech frames, respectively, is given as follow:

$$d_{Intersect} = -\ln \sum_{i=1}^{n} \min\left\{H_X(x \in X_i), H_N(n \in X_i)\right\} \tag{3.1}$$

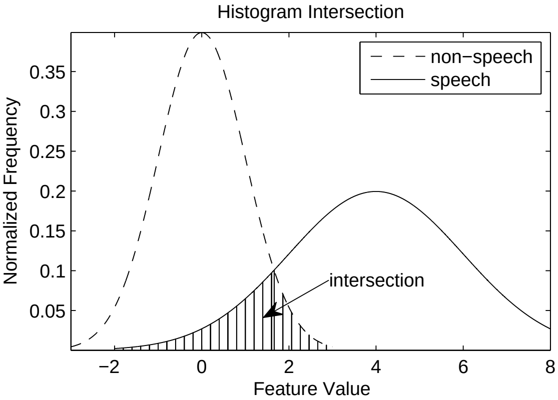


Figure 3.3: Histogram intersection

In Equation 3.1, the x-axis is divided into n bins $\{X_1, \ldots, X_n\}$, and the probabilities $f_X(x \in X_i)$ and $f_N(n \in X_i)$ are calculated based on the normalized histograms $H_X(x)$ and $H_N(n)$, respectively. The minus sign was used so that a higher value means the two PDFs have less overlapped area; or in other words, they are more separated.

3.3 Feature bench-marking

This section evaluates a selected set of VAD features under various noise condition settings. Experimental results using both the proposed metric and the traditional precision-recall ROC curve shows that their outcomes are consistent with each other. In addition, in section 3.3.3, it is shown that hangover plays an important part in evaluating VAD's overall performance.

3.3.1 Using histogram intersection distance

In this experiment, 4 selected features [51, 7, 5] (as listed in Table 3.1) are evaluated using the proposed histogram intersection distance, which calculates the separateness of the distributions of speech and noise. The experiment runs on 100 TIMIT utterances uttered by 100 unique speakers, each of which is 60 seconds long, resulting in approximately 100 minutes of test data. A selected subset of 7 noise samples (as listed in Table 3.2) from the NOISEX-92 dataset are resampled to 16KHz and added at 7 diferent levels of SNRs, ranging from 20 dB down to -10 dB. Note that in this experiment, the speech/non-speech decision is purely based on the ground truth described in Section 3.1.2, no decision rule is involved. This is to emphasize the independence of the proposing metric from the underlying VAD system, i.e. it purely reflects the seperateness of the frame-based features.

Table 3.1: List of features used in the experiment

`energy`	Short-term log-energy feature
`ramirez04`	The long-term spectral diversion (LTSD) feature (page 10)
`sohn99`	Log-likelihood ratio vector of the statistical model (page 19)
`ghosh11`	Long-term signal variability (LTSV) feature (page 16)

Figure 3.4 shows the histogram intersection distance between the distributions of speech and noise, in which a higher value shows a better feature. From this figure, it appears that Ramirez's LTSD feature outperformed the other features in most noise condition settings. Sohn's feature ranked the second in most high-SNR settings. Ghosh's LTSV feature performs very well on the low-SNR conditions, but become saturated at

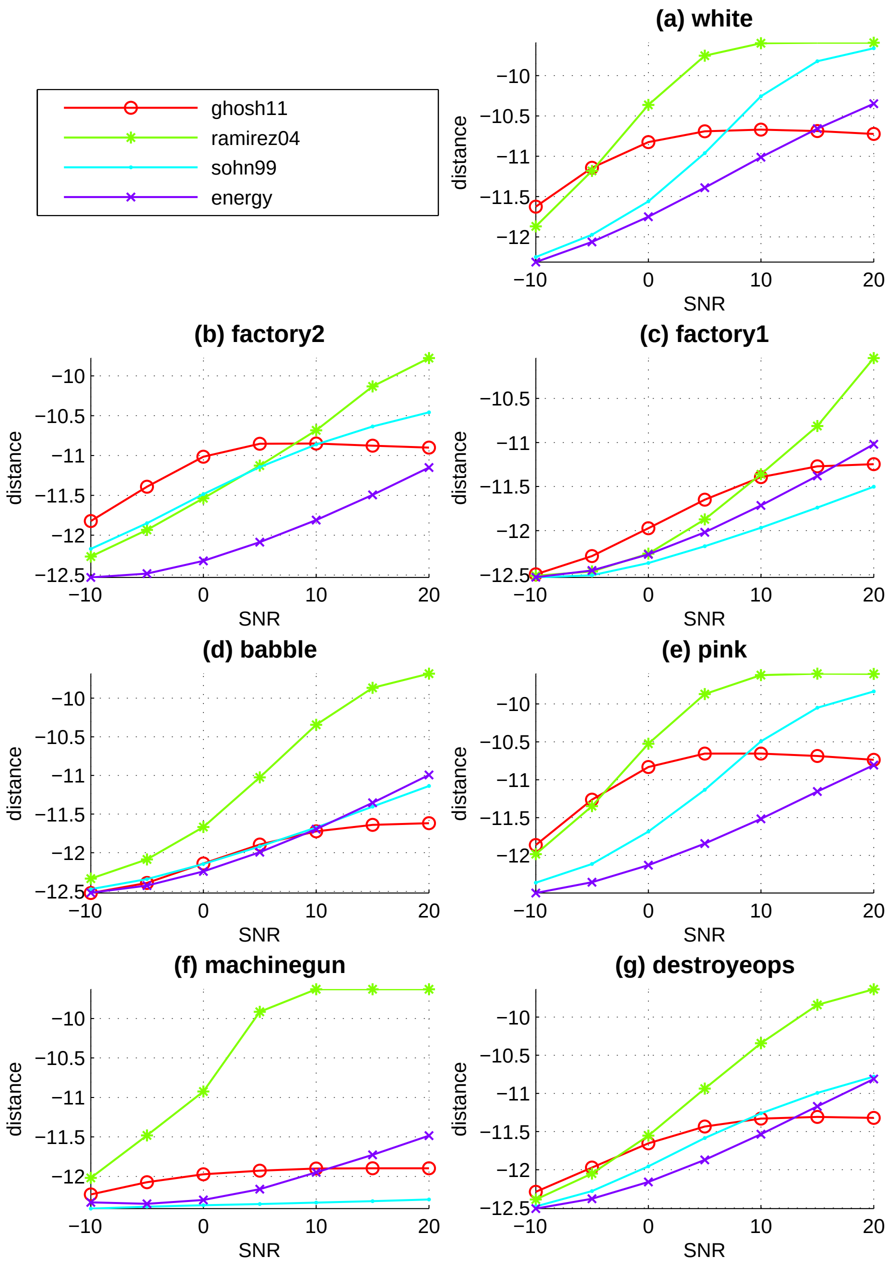


Figure 3.4: Histogram intersection distance between the distribution speech and nonspeech of various VAD features under various noise conditions. Selected features are: Ghosh 2011's LTSV (R=30, M=20), Ramirez 2004's LTSD (M=9), Sohn 1999's posterior probability, and short-term log-energy,

Table 3.2: List of noise types used in the experiment

`white`	white noise
`factory2`	factory noise
`factory1`	factory noise
`babble`	babble noise
`pink`	pink noise
`machinegun`	gun shot noise
`destroyeops`	airplane operation noise

higher SNRs. Lastly, the log-energy feature seems to perform sufficiently well at high SNRs, but degrade quickly and ranked last at lower SNRs. In non-stationary noise conditions ('babble' and 'destroyer operations'), Ramirez's feature performed significantly higher than the others. This can be explained by the implicit energy-based noise estimation and noise reduction that the LTSD feature possess, making it possible to discriminate speech from noise.

The performance of Ghosh's LTSV feature can be explained as follow. At first, it uses a very long window to estimate the averaged spectrogram, using the Barlett-Welch's method, as well as for computing the long-term entropies of each frequency band. This result in a great improvement at very low SNRs, at the cost of much false alarm at the activity edges. However, the authors reported [7] that this behaviour can be overcome in the post-processing module, resulting in a much better performance.

Sohn's statistical method performed well under most stationary noise, but loses its discriminative power significantly under non-stationary noise such as babble noise and factory noise. This is because the feature consider only a single frame, no temporal information betwen frame is considered, and thus it cannot detect the different between speech and non-speech in these cases. This also explains the similar phenomenon for log-energy feature.

It can be concluded from this experiment that the performance of the various features is dependent on the SNR level and on the noise type of the environment. It also shows that the usage of the suggested distance measure can compactly capture the performance of the evaluating VAD features in the various noise conditions, which helped to reveal interesting performance patterns.

3.3.2 Using precision/recall ROC curve without hanging-over

The previous section showed that the proposed distance measure compactly shows the goodness of VAD features in various noise conditions. This section aims to verify the consistency of the proposed distance measure with the traditional precision-recall metrics by plotting the ROC curve of the same set of features and noise types. Due to the non-compactness of the ROC plots, this section only consider the cases where SNR= 0 dB.

Figure 3.5 shows the ROC curves of the precision-recall metric pair of the selected features at various noise condition at 0 dB SNR. Comparing the total area under the curves in each subplots, it appears that this ranking is consistent with the ranking of features at 0 dB in Figure 3.4.

3.3.3 Using precision/recall ROC curve with hanging-over

This section further examine the performance of VAD features after being post-processed by a simple hanging-over scheme. The same experiment as the previous section is carried out, with an additional hanging-over of VAD decisions before calculating the precision and recall values. The hanging-over scheme used for the experiment is as follow. The speech/non-speech decision for frame k should be toggled if there exist frames $k-m$ and $k+n$, ($m, n \leq M$ where M is the hanging-over factor), such that the decisions for these frames are opposite to frame k's, as illustrated in Figure 3.6.

Figure 3.7 shows an interesting performance improvement of some of the features. For example, under white noise at 0 dB, the log-energy feature and Sohn's feature improved greatly and achieved a comparable performance with the others. However, the degrees of relative improvement given by the hanging-over process achieved by each features are different. The most obvious improvement is Sohn's feature, while the least one is Ghosh's LTSV feature. This can be explained by the fact that LTSV and LTSD features use long-term frames, while the other two are based on individual ones. Thus, the former two do not benefit much from the hanging-over process.

From this experiment, it is observed that hanging-over plays an important part in designing a VAD feature. One can implicitly embed this process within the feature extraction itself by employing a long window, which often results in a high preliminary

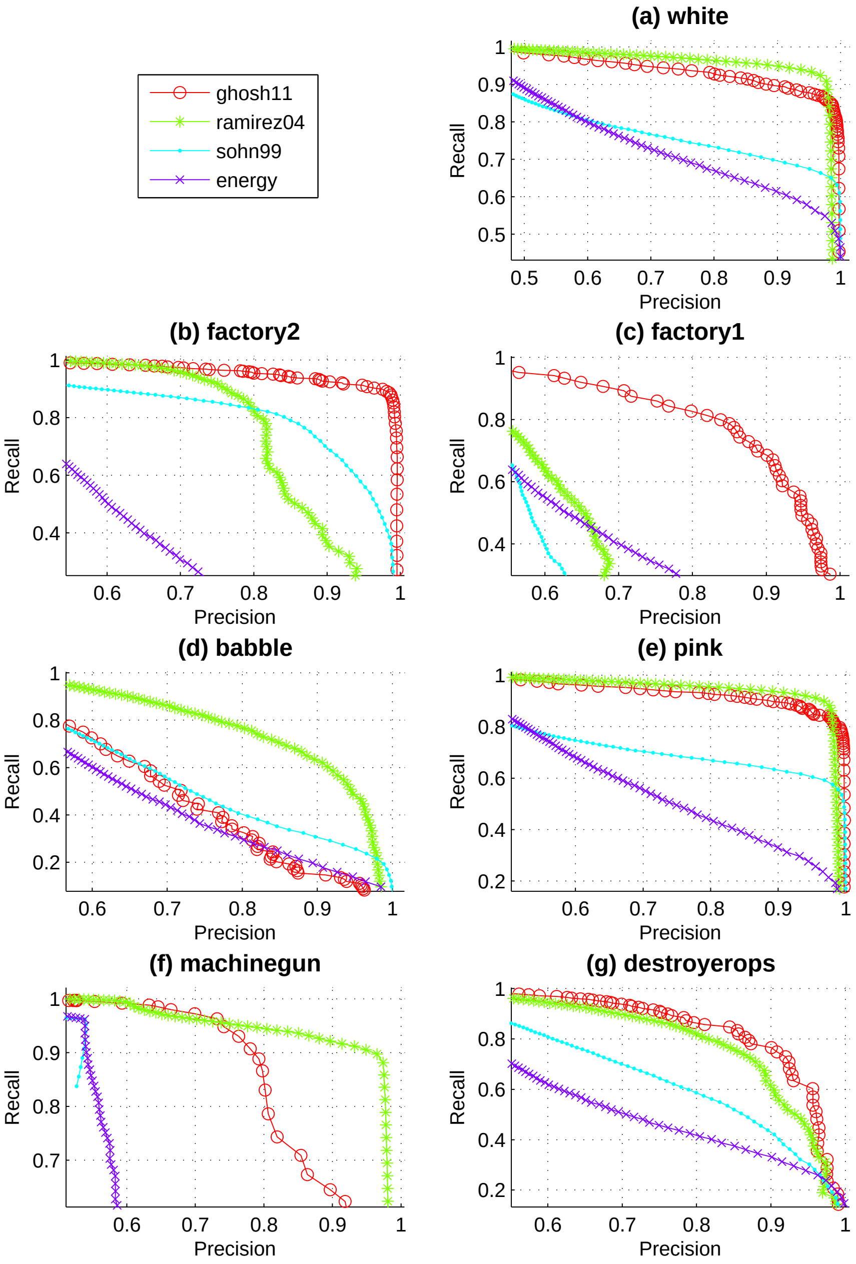


Figure 3.5: Precision/recall ROC curves in various noise types at 0 dB SNR. Selected features are: Ghosh 2011's LTSV (R=30, M=20), Ramirez 2004's LTSD (M=9), Sohn 1999's posterior probability, and the short-term log-energy,

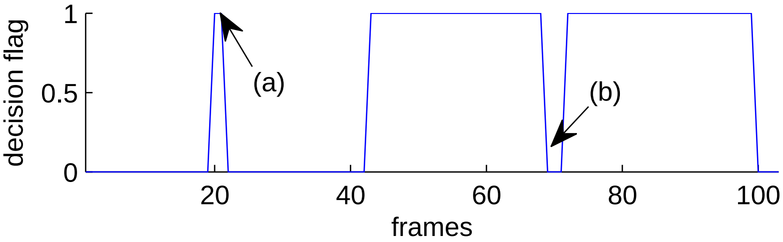


Figure 3.6: Hanging-over scheme, the short onset segment at (a) and offset segment at (b) should be toggled by the hanging-over process.

performance, but suffers from determining precise decision edges. On the other hand, using a shorter window in feature extraction decouples the hanging-over process, leaving rooms for easy improvement in the post-processing module.

3.4 Conclusion

This chapter studied the evaluation of VAD features. The traditional metrics used in existing research is not compact, making it difficult to bench-mark the features against various noise conditions. A proposed metric, called the histogram intersection distance, is introduced to solve this problem. This new metric shows the performance of each feature in various noise types and SNR conditions, revealing many interesting trends and telling the design purpose of each features. The result is a more compact and fairer comparison than the traditional averaging method. Through an experiment, it was shown that this new metric is consistent with the relative ranking given by the area under the ROC curves.

In another experiment, it was observed that although hanging-over is a separate process from feature extraction, it has an considerable effect on the performance of VAD features. To some features, it is a significant performance improvement, while to some others, it is negligible.

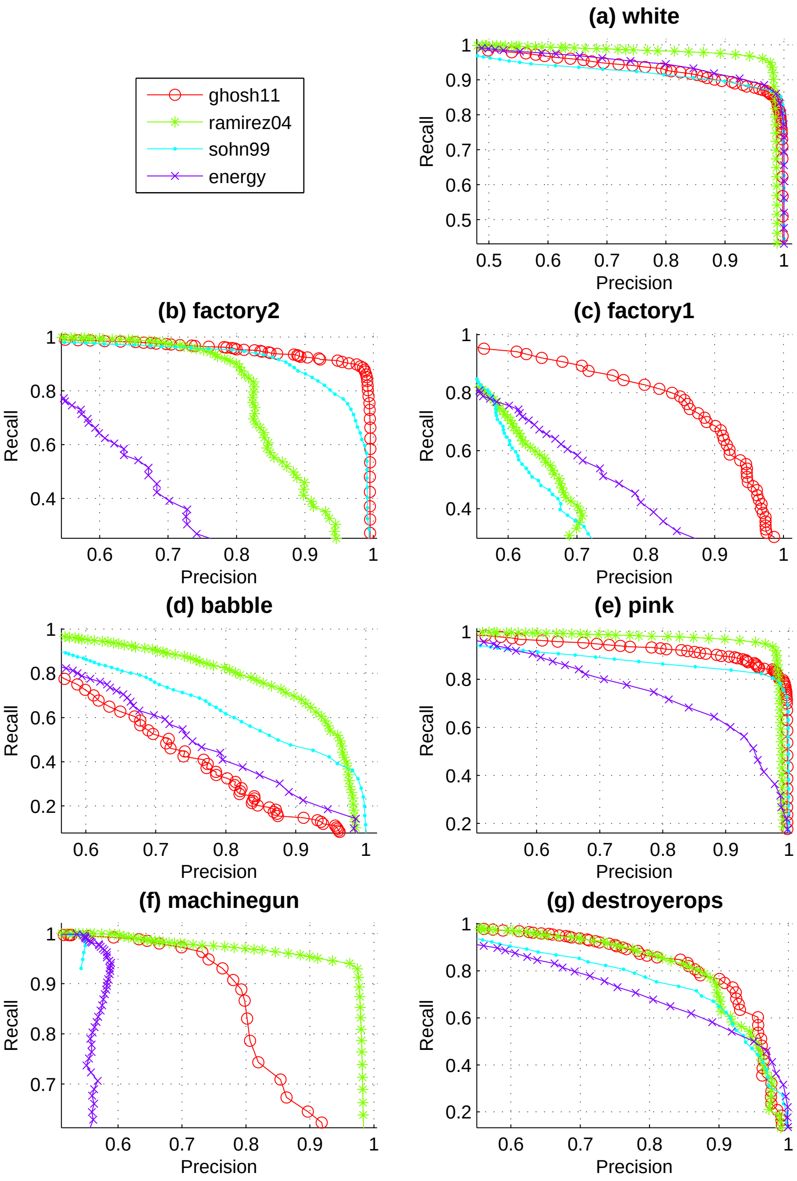


Figure 3.7: Precision/recall ROC curves in various noise types at 0 dB SNR, after being post-processed by a simple hanging-over scheme. Selected features are: Ghosh 2011's LTSV (R=30, M=20), Ramirez 2004's LTSD (M=9), Sohn 1999's posterior probability and the short-term log-energy

Chapter 4

Spectral Local Harmonicity Feature for Voice Activity Detection

The evaluation of VAD features in the previous chapters has revealed a problem of all existing algorithms: they do not work well if SNR is too low (less than 0 dB). At -10 dB SNR, the feature distribution of speech and non-speech are almost fully overlapped, making the detection results no longer reliable. Even at 0 dB, if the background noise contains complex noise events, especially babble and factory noise, existing VADs are not robust [7, 5, 51, 53, 6].

This chapter examines a possible solution to distinguish speech from complex background noise, under very low SNRs. It is based on the observation from the noisy speech spectrum that even when the speech frames are heavily corrupted by noise, there are still sub-regions on the spectrum that show the existence of voiced speech. The idea is further examined and discussed in the remainder of this chapter, which is organized as follow: Section 4.1 discusses the idea further and how it can be realized. Section 4.2 formulates the problem and proposed an algorithm following this idea. Experiments are carried out in Section 4.3, which is discussed further in Section 4.4. Finally, Section 4.5 concludes the chapter and discusses the future work.

4.1 Performance of existing features on heavily corrupted voiced frames

Most of the existing feature extraction methods utilize the entire spectral frame to extract features that measure speech/non-speech likelihood [7, 5, 51, 53, 6]. The problem of such

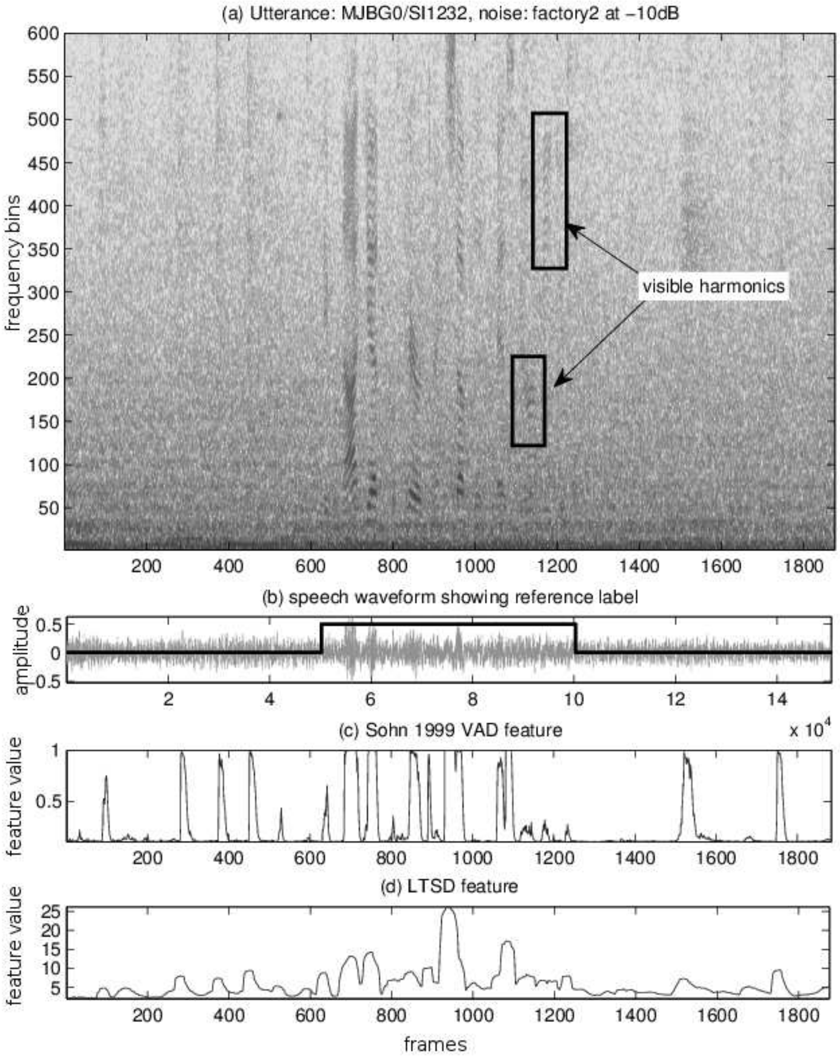


Figure 4.1: Performance of Sohn and LTSD features on very low SNR (−10 dB) on factory noise show that the two features no longer can distinguish between speech and non-speech, even at voiced frames where clear harmonics still visible (a).

approaches is that in low SNR condition, most information contained in the frame is corrupted by noise. This includes the harmonicity structure contained in the voiced speech frames. In many cases where the existing features failed to give a distinguishable measure, this structure is still visible by inspection on the frequency spectrum. For example, in Figure 4.1, there are at least 5 harmonics can be detected visually at frame index 1100–1200, as specified in the figure, but the tested features [51, 5] give very low values for the corresponding frames, which makes them indistinguishable from non-speech frames.

Further examining these frames showed that the energy contributed by this subframe is negligible from the total energy of the spectral frame, which is mainly contributed by noise. This effect can be reduced by filtering out the frequency bands that are heavily corrupted by noise. However, this requires prior knowledge of noise, which makes the solution not flexible and, in some applications, not reliable. Under cases where there are many different noise events, which cover different frequency bands, such approach is not feasible.

Instead of relying on the prior knowledge of noise, one can exploit the prior knowledge of voiced speech itself. Many voiced-based features have been proposed in the literature to exploit the harmonic structure of the voiced frames in severe noise conditions [19]. However, all existing voiced-based features often require the frame to retain many harmonics to be distinguishable from unvoiced and non-speech frames.

In the next section, I propose a new approach to this problem, which aims at detecting voiced frames that have only a little number of harmonics left, which are not detectable using the existing features.

4.2 Problem formulation

To formulate a new VAD feature, the relevant terminations and notations of a VAD system is briefly described as follow. The noisy speech signal $s_y(n)$ is segmented into overlapped frames $y_k(n)$ by using a sliding window function, where k denotes the frame number, and n denotes the sample number, $n \in [1 \dots N]$, where N is the frame length. For simplicity, the remainder of this section only analyzes a single frame at constant k, and thus it is omitted from the following equations. Let $Y(\omega)$ and $N(\omega)$ be the log-magnitude spectrum of the noisy speech and noise only frames, respectively, where $\omega \in [1 \dots N_{\text{DFT}}]$ denotes the frequency bin of the spectrum, and N_{DFT} is the number of coefficients used in the discrete Fourier transform. In practice, $N(\omega)$ can be estimated from a sample segment of the noise signal. The stationary noise components in $Y(\omega)$ can be reduced by using a simple spectral subtraction technique [25], this results in a fairly cleaner spectrum $\hat{Y}(\omega)$:

$$Y(\omega) = \log_{10}\left|\sum_{n=0}^{N-1} y(n)e^{-2\pi\omega n/N}\right| \tag{4.1}$$

$$\hat{Y}(\omega) = Y(\omega) - N(\omega) \tag{4.2}$$

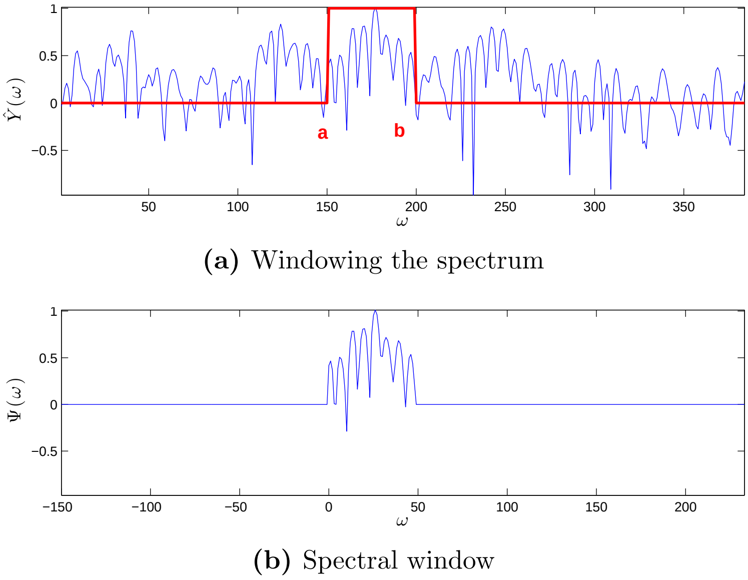
4.2.1 Optimal spectral local feature searching

Let us define a spectral window spanning from frequency bins $[a, b]$ of the spectrum $\hat{Y}(\omega)$ as a parameterized function $\Psi(\omega; a, b)$:

$$\Psi(\omega; a, b) = \hat{Y}(\omega + a)w_r\left(\frac{\omega}{b-a}\right) \tag{4.3}$$

$$w_r(\omega) = \begin{cases} 1 & \omega \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

where parameters a and b denote the outer frequency bins of the spectral window, $1 \leq a < b \leq N_{\text{DFT}}$, and $w_r(\omega)$ is the rectangular window function. In other words, $\Psi(\omega; a, b)$ is calculated by first shifting $\hat{Y}(\omega)$ left by a, and then windowed from 0 to $b-a$ (Figure 4.2).



(a) Windowing the spectrum

(b) Spectral window

Figure 4.2: An example of the spectral windowing process showing (a) the noise reduced frequency spectrum, and (b) an example window of the spectrum.

Let $\Lambda(a, b)$ be an objective function that measures the local feature of the spectral window $\Psi(\omega; a, b)$. The proposing approach searches for all the potential windows, $\Psi(\omega; a, b)$, and find the one that produces the maximum value, ξ, for the objective function. This will be used as the feature for discriminating noisy voiced speech against noise:

$$\xi = \max_{\substack{a,b \in [1; N_{\mathrm{FFT}}] \\ a<b}} \Lambda(a,b) \tag{4.5}$$

Such features henceforth will be referred to as the *spectral local features.* The next subsection presents one example of such features, which exploits the local harmonicity of the spectrum.

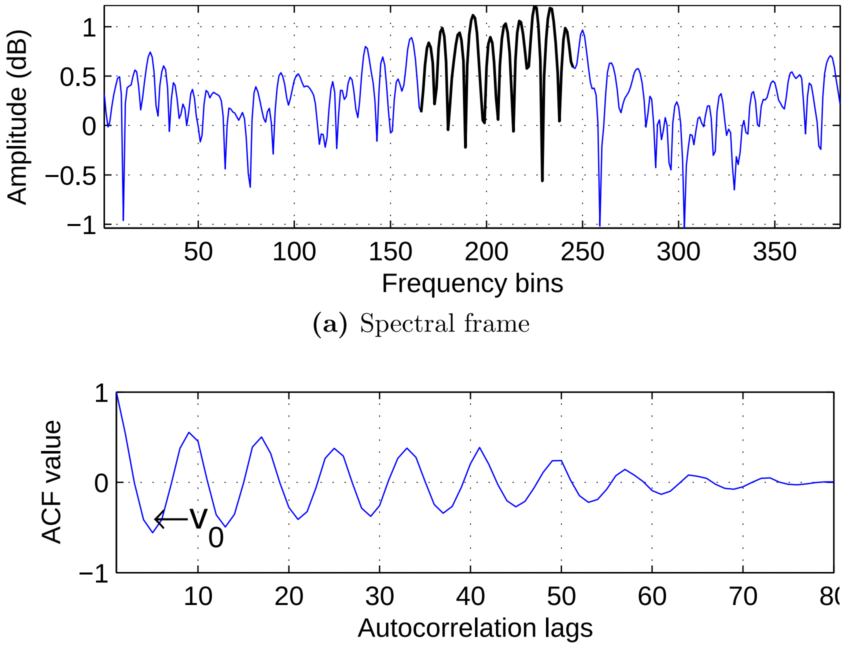
4.2.2 Spectral local harmonicity feature

The harmonicity property of the voiced speech frame results in the local maxima (i.e. spectral peaks) at equal intervals in the frequency spectrum. Most traditional features either use an impulse train [6, 90] to search for the set of peaks in multiples of F_0, or uses the autocorrelation-inspired functions to find the self repetition of the signal in the time- or frequency-domain [33, 19]. One problem appears when searching for an optimal spectral window to extract local harmonicity feature is that the window length can vary. Thus, a long noise frame might contain a set of random peaks which might appear at the multiples of an acceptable fundamental frequency. On the other hand, a window of a noisy speech frame might contain enough clear harmonics, but might have a very short frame length. Therefore, traditional approaches don't work well in balancing between the frame length and its harmonicity measure.

A new measure for the local harmonicity of the spectral window, based on the shape of the spectral autocorrelation is given as follow. Let us first define the normalized zero-mean sample autocorrelation function of the spectral window as follow: [91]

$$r(m;a,b) = \frac{1}{r_0}\frac{1}{M}\sum_{\omega=1}^{M-m}\left(\Psi(\omega;a,b)-\bar{\Psi}\right)\left(\Psi(\omega+m;a,b)-\bar{\Psi}\right) \tag{4.6}$$

where $m = [0, 1 \ldots M]$ is the autocorrelation lag, $M = b - a$ is the maximum lag, r_0 and $\bar{\Psi}$ are the energy and mean of the spectral window, respectively. Theoretically, this autocorrelation function has the following inherent properties: (i) zero mean, (ii) its envelope decreases gradually from 1 at lag zero towards 0 at the final lag, and (iii) it has a period corresponds to the lag of the fundamental frequency. These properties are similar to that of a damped cosine function. Thus, given an arbitrary autocorrelation function, we can test for these properties by fitting an envelope-weighted cosine curve to it. This effectively measures how 'harmonic-like' a curve is (Figure 4.3b).



(a) Spectral frame

(b) Spectral window ACF

Figure 4.3: An example of the spectral sub-window autocorrelation function, showing (a) a noisy spectral frame (0 dB factory noise), together with the selected sub-window (the thicken line), and (b) the autocorrelation of the selected spectral window. The first autocorrelation peak, v_0, is also annotated.

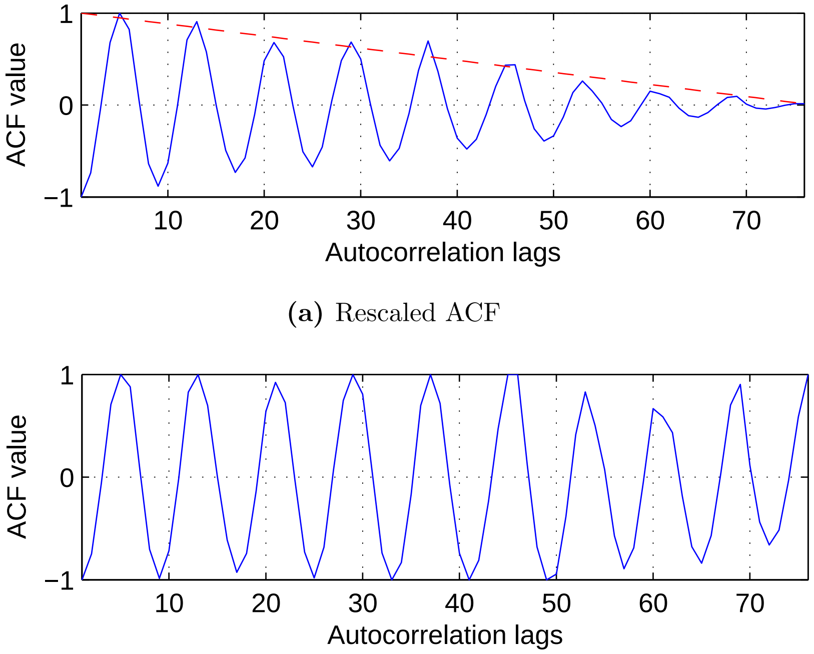
Due to the effect of noise, the ACF curve of many spectral windows might have a very high peak at lag zero, its energy, destroying the typical decreasing envelope of property (ii). To avoid this effect, the algorithm only processes the curve from the first autocorrelation valley (i.e. local minimum, Figure 4.3b) onwards[1], and rescales it to bring back to the $[-1, 1]$ peak-to-peak range:

$$v(f(x)) = -1 + 2\frac{f(x) - \min(f(x))}{\max(f(x)) - \min(f(x))} \tag{4.7}$$

$$\dot{r}(m; a, b) = v\left(r(m + m_{\mathrm{v}_0}; a, b)\right) \tag{4.8}$$

[1] Experiments showed that this lead to better results than choosing from the second peak onwards

where $\upsilon(f(x))$ is the scaling function, which brings any function $f(x)$ to the $[-1, 1]$ peak, $\dot{r}(m; a, b)$ is the modified ACF curve, computed by first shifting $r(m; a, b)$ left by m_{v_0}, the lag corresponding to its first valley, and then scaling up using $\upsilon(\cdot)$ (Figure 4.4a).



(a) Rescaled ACF

(b) Envelope weighted ACF

Figure 4.4: (a) $\dot{r}(m; a, b)$, the [-1,1] peak-to-peak rescaled version of the ACF curve, together with $\epsilon_r(m)$, the ACF envelope (dashed line), (b) $\ddot{r}(m; a, b)$, the result of weighing $\dot{r}(m; a, b)$ against the reciprocal of $\epsilon(m)$.

To find the closest cosine curve that matches the ACF curve, we apply the discrete cosine transform (DCT) over $\ddot{r}(m; a, b)$, a scaled up version of the curve, by weighing it against the reciprocal of the ACF envelope, $1/\epsilon_r(m)$. Effectively, this removes the (ii) property of the ACF function and brings it even closer to a single cosine component (Figure 4.4b), thus reduces noise in the DCT spectrum (Figure 4.5a):

$$\ddot{r}(m; a, b) = \dot{r}(m; a, b)/\epsilon_r(m) \tag{4.9}$$

$$R(k; a, b) = \kappa(k) \sum_{m=1}^{M} \ddot{r}(m; a, b) \cos \frac{\pi(m + 0.5)k}{2M} \tag{4.10}$$

$$\kappa(k) = \begin{cases} \sqrt{\frac{1}{M}} & k = 0 \\ \sqrt{\frac{2}{M}} & k > 0 \end{cases} \tag{4.11}$$

where $R(k;a,b)$ are the DCT coefficients, $k=[1..M]$, $\kappa(k)$ is the DCT scaling factor [92], and $\epsilon_r(m)$ is the autocorrelation envelope scaling function:

$$\epsilon_r(m) = \frac{M-m}{M} \tag{4.12}$$

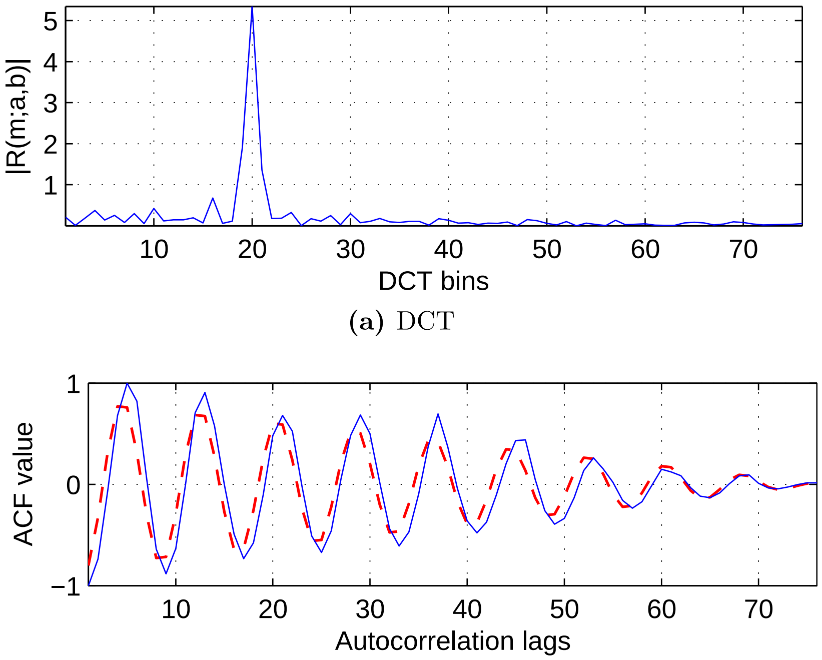
Fitting a cosine curve to the ACF function can now be done by selecting the single highest energy sinusoidal component k_0 from the DCT spectrum:

$$k_0 = \underset{k\in[k_{\min},k_{\max}]}{\arg\max} |R(k;a,b)| \tag{4.13}$$

where $k_{\min}$ and $k_{\max}$ are the acceptable range of DCT component, which can be derived from the range of F_0 by the following mapping:

$$k_0 = \frac{2M}{m_0} \tag{4.14}$$

$$m_0 = F_0 N_{FFT}/F_S \tag{4.15}$$



(a) DCT

(b) Reconstructed ACF

Figure 4.5: Its DCT is shown in (a), and the reconstructed ACF, $\hat{r}(m;a,b)$, is shown as the dashed line in (b).

And finally, the objective function associated with each spectral window $[a,b]$ is based on the mean absolute residual error of the ACF curve, $\dot{r}(m;a,b)$, and $\hat{r}(m;a,b)$, its reconstructed version from the cosine fitting :

$$\Lambda(a,b) = 1 - \frac{1}{M}\sum_{m=1}^{M} |\dot{r}(m;a,b) - \hat{r}(m;a,b)| \tag{4.16}$$

where $\hat{r}(m;a,b)$ is reconstructed using the inverse DCT, and weighted by the ACF envelope (Figure 4.5b):

$$\hat{r}(m;a,b) = \sqrt{\frac{2}{M}} R(k_0;a,b) \cos\frac{\pi(m-0.5)k_0}{2M} \epsilon_r(m) \tag{4.17}$$

Note that in Equation 4.16, we are using the energy-removed ACF curve, $\dot{r}(m;a,b)$, instead of the weighted version to keep the ACF's gradually reduced magnitude shape. When calculating the residual error, this shape effectively implies smaller weights being applied to the ending lags, which tends to be biased during the ACF calculation due to the small number of data points used.

Since the ACF is normalized to energy 1 at lag zero, Equation 4.16 can give high value for noise window with very low energy. Thus, we propose to also use the local mean energy, $E(a,b)$, of the window $\Psi(\omega;a,b)$ to extract a final parameter for VAD. $E(a,b)$ is calculated as follow:

$$E(a,b) = \frac{1}{b-a+1}\sum_{\omega=1}^{b-a+1} \Psi(\omega;a,b) \tag{4.18}$$

The final parameter, dubbed the *Spectral Local Harmonicity* (SLH) feature, is extracted using Algorithm 1. The entire process to extract the SLH feature for each frame can be summarized into the following steps:

(i) Find the log magnitude spectrum, $Y(\omega)$, and perform noise subtraction to give $\hat{Y}(\omega)$.

(ii) Let SLH $= 0$.

(iii) Take a sub-window $[a,b]$ of the spectrum, using the window function $w_r(\omega)$, which gives $\Psi(\omega;a,b)$

(iv) Compute the autocorrelation $r(m;a,b)$

(v) Find the first valley location m_{v_0}

Algorithm 1 SLH feature

```
$\Lambda^* \leftarrow 0$
SLH $\leftarrow 0$
**for** $a \leftarrow 1, N_{\text{DFT}}$ **do**
    **for** $b \leftarrow a + 1, N_{\text{DFT}}$ **do**
        Calculate $\Lambda(a, b)$ according to Equation 4.16
        **if** $\Lambda^* < \Lambda(a, b)$ **then**
            $\lambda \leftarrow \Lambda(a, b) * E(a, b)$
            **if** $\lambda >$ SLH **then**
                $\Lambda^* \leftarrow \Lambda(a, b)$
                SLH $= \lambda$
            **end if**
        **end if**
    **end for**
**end for**
```

(vi) Compute the modified ACF curve, $\dot{r}(m; a, b)$, by shifting $r(m; a, b)$ left by m_{v_0} lags, and then scaling to $[-1, 1]$ peak-to-peak value using the scaling function $\upsilon(\cdot)$

(vii) Reshape the curve by weighing against the reciprocal of the autocorrelation shape function $\epsilon_r(m)$, this gives $\ddot{r}(m; a, b)$

(viii) Compute the DCT of $\ddot{r}(m; a, b)$, $R(k; a, b)$

(ix) Find the maximum cosine component, k_0, in an acceptable range according to the range of F_0

(x) Compute the reconstructed ACF curve, $\hat{r}(m; a, b)$

(xi) Compute the value of the objective function $\Lambda(a, b)$

(xii) Use Algorithm 1 to update the SLH parameter.

(xiii) Repeat steps 3–12 for all possible $[a, b]$ pairs.

4.3 Experiments

An experiment was carried out to evaluate the proposed feature. To make a fair comparison, a recently published voicing feature called the Windowed Autocorrelation Lag Energy (WALE) [19] was chosen to compare the performance in the voiced-based feature category. WALE is an autocorrelation-based feature, whose performance has been evaluated against various other voicing features and has shown to perform better in most cases. As a reference to all other feature categories, Ramirez's LTSD feature [5] was also used in the comparison. In the first experiment, the LTSD feature was configured to used only N=1 frame, i.e. using no long-term information. In the second experiment, SLH feature is compared against LTSD with N=9 as proposed in [5] to inspect the effect of long-term information on the detection of voiced speech.

The experiment ran on 50 speech utterances randomly chosen from the TIMIT database, artificially corrupted by 4 selected noise types from the NOISEX-92 database. The selected noise types are `white`, `pink`, `factory1` and `machinegun`. Of these four noise types, the former two are stationary color noise, which represent the cases that can be considered as easy for most existing VADs. The latter two, however, contain various noise events that make it difficult to detect speech, and can be considered as difficult cases for most existing VADs. The final noisy signal is sampled at 16 KHz, segmented into short frames using a sliding Hanning window of length 30 ms and shifting length 10 ms.

Figure 4.6 shows the result of the proposed solution on a male speaker, under 0 dB factory noise, which contains complex stationary and non-stationary mixtures. Figure 4.6a shows the noisy log-magnitude spectrum, together with the ground truth voiced label, extracted from the database. It can be observed from Figure 4.6b that the proposed method gives high value for the voice segments around frame 10, whose harmonics of the fundamental frequency are only visible in higher frequency range (around 1500–2000Hz). On the other hand, the corresponding result of WALE (Figure 4.6c) is not sufficient to discriminate this frame against noise.

Figure 4.7 shows the distributions of the comparing features for voiced speech and noise at 3 different SNRs. It can be observed that the probability density function (PDF) curves of the proposed feature for voiced speech and noise have better separation than WALE's in all cases. For most noise types, the performance difference is not significant at 5 dB (Figure 4.7b and 4.7d), as WALE feature performs relative well at higher SNRs. But as SNR getting lower, for example, at 0 dB `white` noise (Figure 4.7b) or -5 dB `factory1` noise (Figure 4.7a), it is clearer that the proposed feature results in better feature distribution separation. The PDF curve of `machinegun` noise for both features

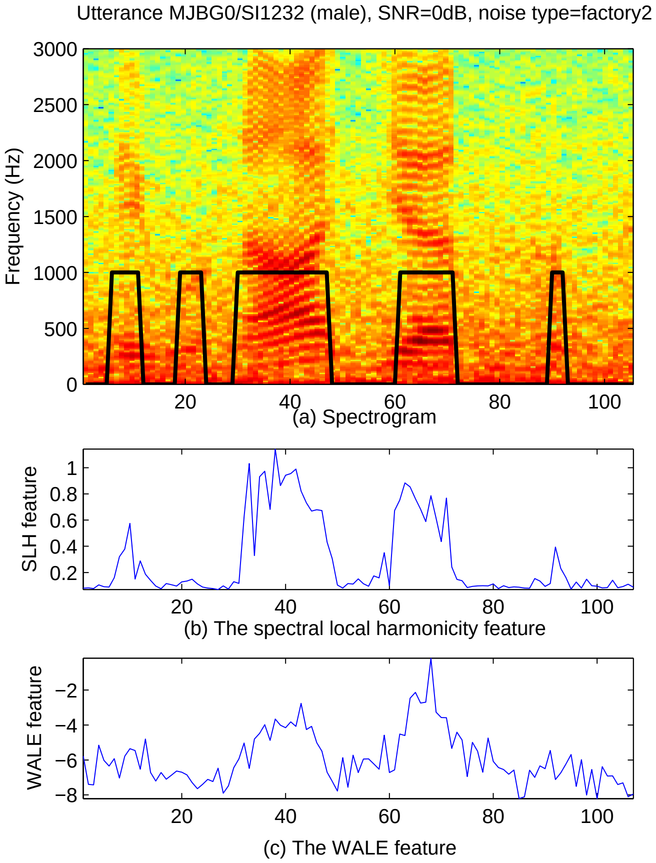
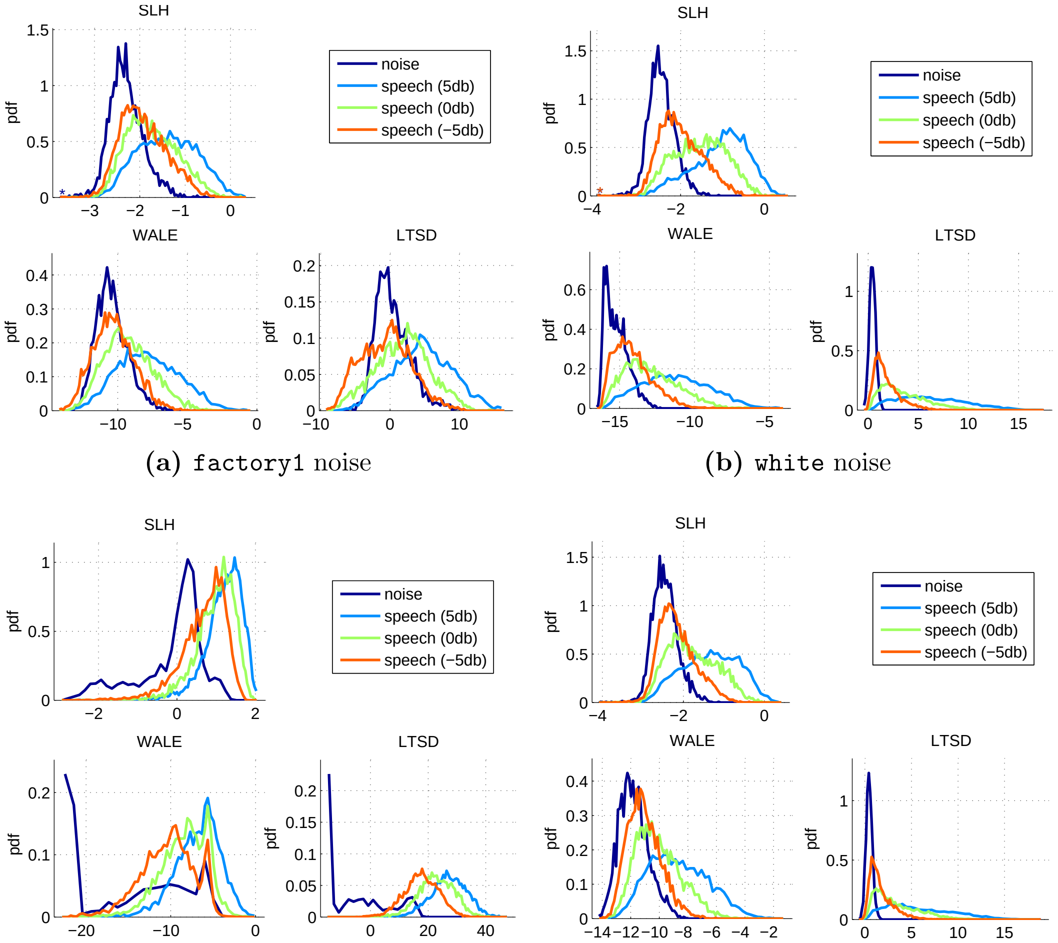


Figure 4.6: Result on an utterance at 0 dB factory noise. Notice the peak value of the proposed feature around frame 10.

(Figure 4.7c) have a very distinct shape. This is due to the frequent gun shots in the environments, which results in an additional peak in the distribution. It can be observed that the peaks of speech and noise distributions are well separated for the proposed feature, while they are completely overlapped for WALE feature. This can be explained by the fact that the frequent high energy gun shots had heavily corrupted the harmonicity structure in the autocorrelation domain, thus affected WALE's performance. On the other hand, SLH feature searches for the remaining harmonic structure in the spectrum, thus it is less affected by the gun shot noise, since in most cases, the voiced speech still retains some harmonics in the higher frequency range.

The LTSD feature performed well in most cases comparing to the proposed feature. This can be explained by the fact that LTSD used a sub-band energy-based approach,



(a) `factory1` noise

(b) `white` noise

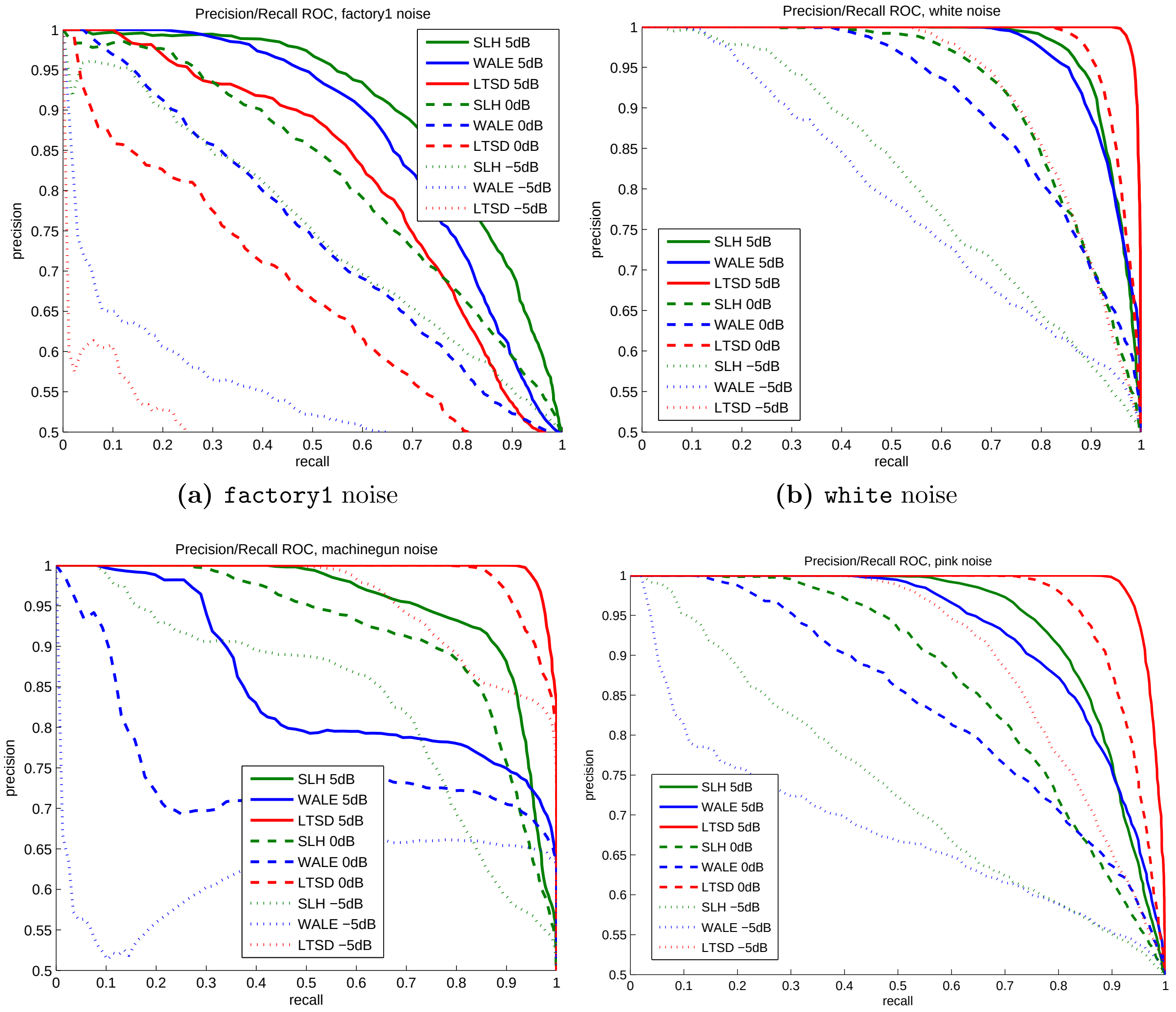
(c) `machinegun` noise

(d) `pink` noise

Figure 4.7: Distributions of voiced speech and noise of SLH, WALE and LTSD

which works well in stationary noise types such as **`white`** and **`pink`**. In these cases, the high-energy harmonics of voiced speech in some sub-bands make it effective to use sub-band energy-based features. But when the noise contains complex high-energy noise events such as `factory1` noise, the proposed feature appears to perform better, especially in the lower SNR cases (Figure 4.7a).

Figure 4.8 shows the same result data plotted in a precision-recall ROC graph. Each plot is the result of varying a threshold in the previous PDF plot, and measure the detection precision and recall. Each threshold value results in a single point on the ROC curve. Thus, the plot shows the trade off between the precision and recall of the actual voiced speech detection. The further away from the origin a point on the curve is, the



(a) `factory1` noise

(b) `white` noise

(c) `machinegun` noise

(d) `pink` noise

Figure 4.8: The precision-recall curves for SLH, WALE and LTSD (N=1) features at different SNRs

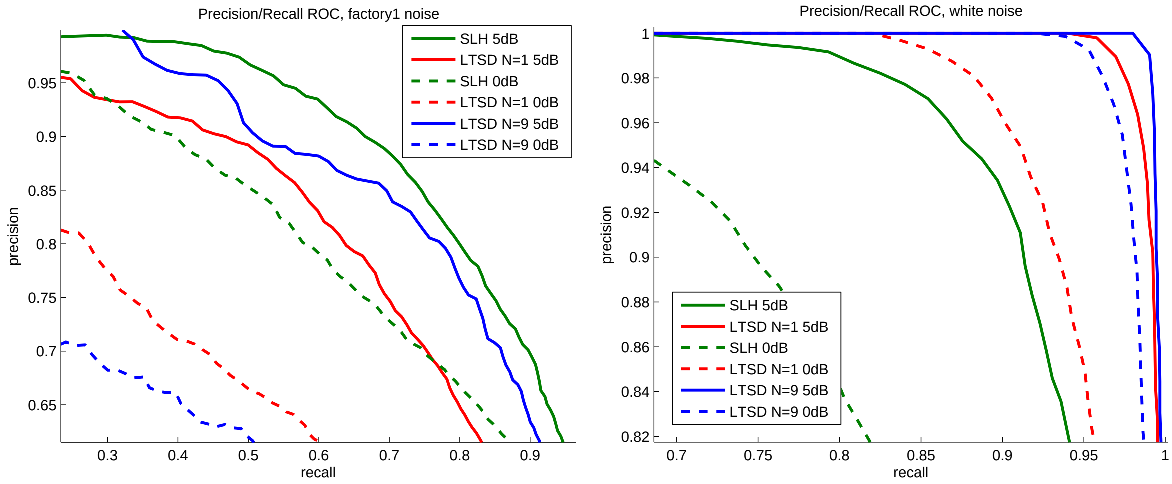
better overall performance the VAD system is. This also means that a larger area under the curve (AUC) shows a better performance of the corresponding feature.

It can be observed from Figure 4.8 that the new feature give larger area under the curve than WALE in all 12 cases. Specifically, for easier noise types such as `white` (Figure 4.8b) and `pink` (Figure 4.8d), the ROCs for the proposed feature are always slightly on the upper right of the corresponding curves for WALE feature. This means that the resulting AUCs of the proposed features are slightly larger than WALE's, thus better performances. However, for more difficult noise cases such as `factory1` and `machinegun`, the difference are much more noticeable. For `factory1` noise (Figure 4.8a), the performance of the SLH feature at -5 dB is slightly better than WALE's performance at 0 dB. Under `machinegun` noise (Figure 4.8c), the ROCs and PDFs of WALE suggest its insufficient discriminative power and thus poor performance, possibly due to the frequent high energy gun shot noise events. Still, the ROCs of the proposed feature SLH is relatively comparable with its performance under other noise cases at the corresponding SNRs. This suggest the SLH feature performs more stably at different noise types and SNRs, or in other words, it is more noise robust.

Figure 4.8 also shows that the LTSD feature performed surprisingly well in most cases, even no long-term information was used. As discussed previously, only in the `factory1` case has the proposed feature surpassed LTSD's performance. Figure 4.8a shows that under `factory1` noise, where the environment contains many high-energy noise events, the SLH feature performed significantly higher comparing to LTSD and WALE. Especially at -5 dB, when both LTSD and WALE shows the insufficient discriminative power (their ROCs are very close to the lower left corner), SLH's curve is much further away to the right. This proves the proposed feature has the desired characteristics as stated in the beginning of this chapter, i.e. noise robust under lower SNR and complex noise events.

4.3.1 The effect of long-term information

The previous experiment only evaluate the proposed feature against other single-frame feature. To further inspect the effect of long-term information for VAD features, another experiment was carried out to compare SLH against LTSD feature configured with N=9, i.e. the LTSD feature at each frame is produced by using data from the 9 previous adjacent frames. The result plotted in Figure 4.9b shows that the long-term information has greatly improved LTSD feature's performance under `white` noise, bringing its 0 dB curve (blue dashed line) to much further away from the 5 dB curve of SLH (green solid line). On the other hand, under `factory1`, while the extra long-term information improved the LTSD performance at 5 dB, SLH's performance is still better (Figure 4.9a). At lower SNRs such as 0 dB, however, the long-term information has a reverse effect on LTSD's performance, lowering its AUC significantly (blue and red dashed lines).



(a) `factory1` noise

(b) `white` noise

Figure 4.9: The precision-recall curves for SLH, LTSD (N=1) and LTSD(N=9) features at different SNRs

4.4 Discussion

Although many features have been proposed in the literature to exploit the harmonicity of voiced speech, to the best of the author's knowledge, they all rely on the entire frame spectrum. Many of these features work sufficiently well in high SNR cases such as from 10 dB and above. When facing with lower SNR cases, or when the background noise contains complex audible events appearing occasionally, such as the train noise at a train station, and babble noise and machinery noise in a factory, there will be cases when most of the speech spectrum are corrupted, which destroy the overall statistical as well as structural properties of the speech signal. The proposed approach which searches for the best local feature ensures that voiced speech can be detected, so long as only a subset of the spectrum contains enough harmonicity information. This results in the higher recall and precision rates, as observed in the previous experiment.

One major drawback of this approach is that it operates in $\mathcal{O}(nm^2)$ time, where n is the number of frames, and m is the number of bins in the frequency transforms. In the experiments, the time complexity was reduced by only considering the spectral windows that cover at least 4 harmonics of the highest F_0, and at most 8 harmonics of the lowest F_0. Also, only the windows beginning and ending at the valleys of the spectrum were considered. For example, if $V = \{v_1 \dots v_K\}$ is the set of locations of valleys of the spectrum $\hat{Y}(\omega)$, only the windows $[a, b]$ having $a, b \in V, a < b$ were processed. This allowed the experiments to be completed in a reasonable amount of time, but is still far from real-time. Nevertheless, the method can be used in applications where real-time execution is not a major concern.

4.5 Conclusion

This chapter has introduced a new method for VAD feature extraction. A sliding window was employed in the speech frequency spectrum to search for the optimal part of the spectrum where harmonicity is still well preserved. Features extracted from such sub-window are called the Spectral Local Features. The author has also developed one such feature, called the Spectral Local Harmonicity feature, which aims at measuring how well a spectral ACF curve can be represented by a weighted cosine curve. The new feature showed significant improvements over an earlier reported voicing feature in the literature in both stationary and non-stationary noise cases, which may contain various complex audible events. Experiments showed that the proposed feature has better discriminative power and is more noise robust.

In another experiment, it has been shown that sub-band energy-based feature such as LTSD works sufficiently well for many noise cases, even without any long-term information was used. However, under more difficult noise type where there are many complex audible noise events such as `factory1`, LTSD didn't work well and completely lost its discriminative power at -5 dB. It can also be concluded from this experiment that long-term information can greatly improve VAD performance, which lays a hint for further improvement of the proposed feature.

Chapter 5

Conclusions

This thesis has studied the different features proposed for VAD in the literature. Existing algorithms can work sufficiently well with noisy environment having SNR$\geq$ 5 dB. At lower SNRs, or when the background noise contains complex acoustic events, however, their discriminative power starts to drop quickly. At lower than 0 dB, most features can no longer distinguish speech from noise reliably. In this thesis, the author has proposed a novel approach to deal with very low SNR ($<$ 0 dB). Instead of relying on the entire spectrum, only subsets of the spectrum is used to extract discriminative features. The proposed feature, called the Spectral Local Harmonicity feature, showed a significant improvement over a recently reported VAD feature in the same category, both in term of discriminative power and noist robustness.

In this research, the author has contributed the following:

- A thorough literature survey of the VAD algorithms proposed in the past fifteen years. Most work for VADs can be grouped into 2 categories: new features and new decision mechanisms. Further sub-categories have been defined and studied in Chapter 2.

- An evaluation metric for comparing the discrimination power of VAD features was given in Chapter 3. The new metric allows the compact presentation of VAD performance under various noise type and SNR configurations, showing interesting trends across different SNRs.

- In Chapter 4, a novel approach to detecting voiced speech in heavy non-stationary noise has been proposed. The proposed method uses only the sub-regions of the spectrum that still retain a sufficient harmonic structure. The new feature, dubbed the Spectral Local Harmonicity feature, chooses the most harmonic sub-window of the spectrum by fitting a weighted cosine to its spectral autocorrelation curve and measuring the residual.

5.1 Future Work

Future work of this thesis can be carried out in several directions. Firstly, Chapter 3 can be expanded to cover more VAD features, and to compare the suggested measure against other distance measures such as the Kullback-Leibler distance and the Bhattacharyya distance. Secondly, it has been observed in Chapter 4 that long-term information can bring a significant performance improvement for VAD. Incorporating long-term information to the proposed feature is therefore expected to improve its performance, especially in the stationary noise types and higher SNRs. And lastly, the proposed algorithm to extract the local harmonicity is a brute-force approach. A better algorithm is needed to bring the execution time closer to real-time.

5.2 Publications

The following papers have been published during the course of this research:

- **Pham Chau Khoa**, Chng Eng Siong, "Spectral Local Harmonicity feature for voice activity detection," in *International Conference on Audio, Language and Image Processing (ICALIP),* 2012.

- Kannu Mehta, **Pham Chau Khoa**, Chng Eng Siong, "Linear dynamic models for voice activity detection," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2011. *(As attached in the appendix)*

References

[1] W. Gardner, P. Jacobs, and C. Lee, “QCELP: A variable rate speech coder for CDMA digital cellular,” *Kluwer International Series in Engineering and Computer Science*, pp. 85–85, 1993.

[2] A. Benyassine, E. Shlomot, H.-Y. Su, and E. Yuen, “A robust low complexity voice activity detection algorithm for speech communication systems,” in *Speech Coding For Telecommunications Proceeding, 1997, 1997 IEEE Workshop on*, p. 97, 1997.

[3] TIA/EIA/IS-127, “Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems,” 1996.

[4] ETSI, “Universal mobile telecommunication systems mandatory speech codec speech processing functions, AMR speech codec; voice activity detector (3GPP TS 26.094 version 4.0.0 release 4),” 2001.

[5] J. Ramirez, J. Segura, C. Benitez, A. De La Torre, and A. Rubio, “Efficient voice activity detection algorithms using long-term speech information,” *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[6] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, “Noise robust voice activity detection based on periodic to aperiodic component ratio,” *Speech Communication*, vol. 52, no. 1, pp. 41–60, 2010.

[7] P. Ghosh, A. Tsiartas, and S. Narayanan, “Robust voice activity detection using long-term signal variability,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 600–613, 2011.

[8] D. Freeman, G. Cosier, C. Southcott, and I. Boyd, “The voice activity detector for the Pan-European digital cellular mobile telephone service,” in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pp. 369–372, IEEE, 1989.

[9] F. Beritelli, S. Casale, and A. Cavallaero, “A robust voice activity detector for wireless communications using soft computing,” *Selected Areas in Communications, IEEE Journal on*, vol. 16, no. 9, pp. 1818–1829, 1998.

[10] D. Enqing, Z. Heming, and L. Yongli, “Low bit and variable rate speech coding using local cosine transform,” in *TENCON’02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*, vol. 1, pp. 423–426, IEEE, 2002.

[11] K. Itoh and M. Mizushima, “Environmental noise reduction based on speech/non-speech identification for hearing aids,” in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 1, pp. 419–422, IEEE, 1997.

[12] J. Festen, J. Van Dijkhuizen, and R. Plomp, “The efficacy of a multichannel hearing aid in which the gain is controlled by the minima in the temporal signal envelope.,” *Scandinavian audiology. Supplementum*, vol. 38, p. 101, 1993.

[13] A. Sangwan, M. Chiranth, H. Jamadagni, R. Sah, R. Venkatesha Prasad, and V. Gaurav, “VAD techniques for real-time speech transmission on the internet,” in *High Speed Networks and Multimedia Communications 5th IEEE International Conference on*, pp. 46–50, IEEE, 2002.

[14] R. Brueckmann, A. Scheidig, and H. Gross, “Adaptive noise reduction and voice activity detection for improved verbal human-robot interaction using binaural data,” in *Robotics and Automation, 2007 IEEE International Conference on*, pp. 1782–1787, IEEE, 2007.

[15] H. Kim, K. Komatani, T. Ogata, and H. Okuno, “Two-channel-based voice activity detection for humanoid robots in noisy home environments,” in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pp. 3495–3501, IEEE, 2008.

[16] T. Fukuda, O. Ichikawa, and M. Nishimura, “Long-term spectro-temporal and static harmonic features for voice activity detection,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, p. 834, 2010.

[17] L. Rabiner and M. Sambur, “An algorithm for determining the endpoints of isolated utterances,” *Bell System Tech. Jour.*, vol. 54, no. 2, pp. 297–315, 1975.

[18] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, “An improved endpoint detector for isolated word recognition,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 4, pp. 777–785, 1981.

[19] T. Kristjansson, S. Deligne, and P. Olsen, “Voicing features for robust speech detection,” in *Conference of the International Speech Communication Association*, pp. 369–372, 2005.

[20] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, “Performance evaluation and comparison of G. 729/AMR/fuzzy voice activity detectors,” *Signal processing letters, IEEE*, vol. 9, no. 3, pp. 85–88, 2002.

[21] S. Gerven and F. Xie, “A comparative study of speech detection methods,” in *Fifth European Conference on Speech Communication and Technology*, 1997.

[22] N. Cho and E.-K. Kim, “Enhanced voice activity detection using acoustic event detection and classification,” *Consumer Electronics, IEEE Transactions on*, vol. 57, no. 1, p. 196, 2011.

[23] K. Woo, T. Yang, K. Park, and C. Lee, “Robust voice activity detection algorithm for estimating noise spectrum,” *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[24] L. Rabiner and R. Schafer, *Digital processing of speech signals.* Prentice Hall, 1978.

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

[26] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *proc. Eurospeech*, pp. 1887–1890, 2001.

[27] J. Ramirez, J. Segura, C. Benitez, d. la, and A. Rubio, "Voice activity detection with noise reduction and long-term spectral divergence estimation," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 2, p. ii, 2004.

[28] A. M. Noll, "Cepstrum pitch determination," *The journal of the acoustical society of America*, vol. 41, pp. 293–309, 1967.

[29] S. Ahmadi and A. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 333–338, 1999.

[30] T. Kinnunen, E. Chernenko, M. Tuononen, P. Frnti, and H. Li, "Voice activity detection using MFCC features and support vector machine," *Int. Conf. on Speech and Computer*, vol. 2, pp. 556–561, 2007.

[31] A. Martin, D. Charlet, and L. Mauuary, "Robust speech/non-speech detection using LDA applied to MFCC," *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP'01). 2001 IEEE International Conference on*, vol. 1, pp. 237–240, 2001.

[32] T. Fukuda, O. Ichikawa, and M. Nishimura, "Phone-duration-dependent long-term dynamic features for a stochastic model-based voice activity detection," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[33] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 25, no. 1, pp. 24–33, 1977.

[34] S. Basu, "A linked-HMM model for robust voicing and speech detection," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1, pp. I–816–I–819, 2003.

[35] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA2006), September*, pp. 65–70, 2006.

[36] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM spine evaluation system," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–53–I–56, May 2002.

[37] H. Tolba and D. O'Shaughnessy, "Robust automatic continuous-speech recognition based on a voiced-unvoiced decision," in *Fifth International Conference on Spoken Language Processing*, 1998.

[38] O. Ichikawa, T. Fukuda, and M. Nishimura, "Local peak enhancement combined with noise reduction algorithms for robust automatic speech recognition in automobiles," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4869–4872, 2008.

[39] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, fourth ed., 2002.

[40] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions," in *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 710–713, 2000.

[41] J.-L. Shen, J.-W. Hung, and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," 1998.

[42] L.-S. Huang and C.-H. Yang, "A novel approach to robust speech endpoint detection in car environments," vol. 3, pp. 1751–1754, 2000.

[43] A. Liberman, *Speech: A special code.* The MIT Press, 1996.

[44] B. Chen, Q. Zhu, and N. Morgan, “Learning long-term temporal features in LVCSR using neural networks,” in *Proc. ICSLP*, pp. 612–615, Citeseer, 2004.

[45] T. Fukuda, O. Ichikawa, and M. Nishimura, “Short-and long-term dynamic features for robust speech recognition,” in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[46] H. Hermansky and S. Sharma, “Temporal patterns (TRAPS) in asr of noisy speech,” in *icassp*, pp. 289–292, IEEE, 1999.

[47] D. Poeppel, “The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time’,” *Speech Communication*, vol. 41, no. 1, pp. 245–255, 2003.

[48] J. Haigh and J. Mason, “Robust voice activity detection using cepstral features,” in *TENCON'93. Proceedings. Computer, Communication, Control and Power Engineering. 1993 IEEE Region 10 Conference on*, no. 0, pp. 321–324, IEEE, 1993.

[49] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1, p. 365, 1998.

[50] G. Evangelopoulos and P. Maragos, “Speech event detection using multiband modulation energy,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[51] J. Sohn, N. Soo, and W. Sung, “A statistical model-based voice activity detection,” *Signal Processing Letters, IEEE*, vol. 6, no. 1, p. 1, 1999.

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

[53] Y. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 2, p. 737, 2001.

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

[55] J. Ramirez, J. Segura, J. Gorriz, and L. Garcia, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, p. 2177, 2007.

[56] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–253, IEEE, 2002.

[57] J. Chang and N. Kim, "Voice activity detection based on complex Laplacian model," *Electronics Letters*, vol. 39, no. 7, pp. 632–634, 2003.

[58] J. Chang, J. Shin, and N. Kim, "Voice activity detector employing generalised gaussian distribution," *Electronics Letters*, vol. 40, no. 24, pp. 1561–1563, 2004.

[59] S. Gazor and W. Zhang, "Speech probability distribution," *Signal Processing Letters, IEEE*, vol. 10, no. 7, pp. 204–207, 2003.

[60] R. Reininger and J. Gibson, "Distributions of the two-dimensional DCT coefficients for images," *Communications, IEEE Transactions on*, vol. 31, no. 6, pp. 835–839, 1983.

[61] J. Shin, J. Chang, and N. Kim, "Statistical modeling of speech signals based on generalized gamma distribution," *Signal Processing Letters, IEEE*, vol. 12, no. 3, pp. 258–261, 2005.

[62] J. Shin, J. Chang, and N. Kim, “Voice activity detection based on statistical models and machine learning approaches,” *Computer Speech & Language*, vol. 24, no. 3, pp. 515–530, 2010.

[63] T. Petsatodis, C. Boukis, F. Talantzis, Z. Tan, and R. Prasad, “Convex combination of multiple statistical models with application to VAD,” *Audio, Speech, and Language Processing, IEEE Transactions on*, no. 99, pp. 1–1, 2011.

[64] J. Chang, N. Kim, and S. Mitra, “Voice activity detection based on multiple statistical models,” *Signal Processing, IEEE Transactions on*, vol. 54, no. 6, pp. 1965–1976, 2006.

[65] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, “Applying support vector machines to voice activity detection,” in *Signal Processing, 2002 6th International Conference on*, vol. 2, pp. 1124–1127, IEEE, 2003.

[66] M. Farsinejad and M. Analoui, “A new robust voice activity detection method based on genetic algorithm,” in *Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian*, pp. 80–84, IEEE.

[67] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* Wiley-Interscience, second ed., 2001.

[68] J. Ramirez, P. Yélamos, J. Górriz, and J. Segura, “SVM-based speech endpoint detection using contextual speech features,” *Electronics letters*, vol. 42, no. 7, pp. 426–428, 2006.

[69] J. Ramírez, P. Yélamos, J. Górriz, J. Segura, and L. García, “Speech/non-speech discrimination combining advanced feature extraction and SVM learning,” in *Ninth International Conference on Spoken Language Processing*, 2006.

[70] J. Wu and X. Zhang, “Maximum margin clustering based statistical VAD with multiple observation compound feature,” *Signal Processing Letters, IEEE*, no. 99, pp. 1–1, 2011.

[71] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, “Maximum margin clustering,” *Advances in neural information processing systems*, vol. 17, pp. 1537–1544, 2005.

[72] F. Wang, B. Zhao, and C. Zhang, “Linear time maximum margin clustering,” *Neural Networks, IEEE Transactions on*, vol. 21, no. 2, pp. 319–332, 2010.

[73] A. Martin and L. Mauuary, “Robust speech/non-speech detection based on LDA-derived parameter and voicing parameter for speech recognition in noisy environments,” *Speech communication*, vol. 48, no. 2, pp. 191–206, 2006.

[74] J. Padrell, D. Macho, and C. Nadeu, “Robust speech activity detection using LDA applied to FF parameters,” in *Proc. ICASSP*, vol. 1, pp. 557–560, 2005.

[75] O. Kwon and T. Lee, “Optimizing speech/non-speech classifier design using adaboost,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*, vol. 1, pp. I–436, IEEE, 2003.

[76] R. Schapire and Y. Singer, “Improved boosting algorithms using confidence-rated predictions,” *Machine learning*, vol. 37, no. 3, pp. 297–336, 1999.

[77] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” 2001.

[78] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, “Aggregate features and AdaBoost for music classification,” *Machine Learning*, vol. 65, no. 2, pp. 473–484, 2006.

[79] T. Usukura and W. Mitsuhashi, “Voice activity detection using AdaBoost with multi-frame information,” in *Signal Processing and Communication Systems, 2008. ICSPCS 2008. 2nd International Conference on*, pp. 1–8, IEEE.

[80] M. Farsinejad, M. Mohammadi, B. Nasersharif, and A. Akbari, “A model-based voice activity detection algorithm using probabilistic neural networks,” in *Communications, 2008. APCC 2008. 14th Asia-Pacific Conference on*, pp. 1–4, IEEE.

[81] T. Pham, C. Tang, and M. Stadtschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," in *Computing and Communication Technologies, 2009. RIVF'09. International Conference on*, pp. 1–8, IEEE.

[82] P. Estevez, N. Becerra-Yoma, N. Boric, and J. Ramirez, "Genetic programming-based voice activity detection," *Electronics Letters*, vol. 41, no. 20, pp. 1141–1143, 2005.

[83] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speechdat-car: A large speech database for automotive environments," in *Proceedings of the II LREC Conference*, vol. 1, 2000.

[84] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, *et al.*, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," *IEICE transactions on information and systems*, pp. 535–544, 2005.

[85] J. Garofolo, *DARPA TIMIT: Acoustic-phonetic Continuous Speech Corps CD-ROM.* US Dept. of Commerce, National Institute of Standards and Technology, 1993.

[86] D. Pearce, H. Hirsch, *et al.*, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP00*, vol. 4, pp. 29–32, Citeseer, 2000.

[87] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 CD-ROMs," *The NOISEX-92 study on the effect of additive noise on automatic speech recognition*, 1992.

[88] K. Mehta, C.-K. Pham, and E.-S. Chng, "Linear dynamic models for voice activity detection," in *Conference of the International Speech Communication Association*, 2011.

[89] T. Nguyen, E. Chng, and H. Li, "T-test distance and clustering criterion for speaker diarization," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[90] C. Shahnaz, W. Zhu, and M. Ahmad, “Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme,” *Audio, Speech, and Language Processing, IEEE Transactions on*, no. 99, pp. 1–1, 2012.

[91] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control.* Prentice-Hall, 3rd ed., 1994.

[92] N. Ahmed, T. Natarajan, and K. Rao, “Discrete cosine transform,” *Computers, IEEE Transactions on*, vol. 100, no. 1, pp. 90–93, 1974.