

國立交通大學

電機與控制工程學系

博士論文

以頻帶及小波分析為基礎的強健性

語音偵測系統之研究

**A Study of Frequency Band and Wavelet Analysis for
Robust Voice Activity Detection**



研究生：王坤卿

指導教授：吳炳飛 博士

中華民國九十四年九月


以頻帶及小波分析為基礎的強健性 語音偵測系統之研究

學生：王坤卿

指導教授：吳炳飛 博士

國立交通大學電機與控制工程學系

摘要



本論文主要是針對語音偵測系統(voice activity detection)在弱的訊號與噪音比值(the signal-to-noise ratio, SNR)及劇烈性的噪音程度變動下之所面臨的問題作些探討。迄今，所提出的語音偵測系統都是假定環境噪音程度是穩定的(stationary)。然而，由於傳統演算法的特徵參數都取決於能量的估測，因此其效能易受到實際噪音程度的變動所影響。比如在車上，劇烈的噪音變動就可能因為移動、引擎運轉、車速、煞車及關車門聲而經常地產生。為了解決這個問題，我們先後提出兩種具強健性(robust)特徵參數為基礎的語音偵測系統。在第一種方法中，根據共鳴頻率(formant frequency)造成在聲音光譜圖(voice spectrogram)的帶狀性紋路(banded line)現象，我們可發現此帶狀性紋路可有效及簡單地表示出具時變特性(time-varying property)語音的存在。透過頻帶分析，我們提出一個以熵為基礎的語音偵測系統。首先，將訊號切成三十二個均勻大小的子頻帶以區隔出共振音頻

的分佈。論文中提出一個定義在子頻帶上的帶狀性頻譜熵值(banded spectrum entropy, BSE) 以充分地利用帶狀性紋路在聲音光譜圖上的固有特性。由於所切出的子頻帶可能被噪音干擾，為了增加 BSE 參數對噪音的抗雜訊能力，我們利用可適性臨界方式(adaptive threshold method)的技巧，建立一個稱作子頻帶自我擷取(subband self-extraction)的方法以能立即地擷取有效的子頻帶。但事實上，聲音光譜圖上帶狀性紋路現像只適合用來特徵有聲的語音訊號。為了要強化語音訊號的無聲部份，其低頻能量對全頻帶能量的比值(the ratio of low-band to full-band energy, RLF)可用來區隔無聲語音與背景噪音特性的差別。相較於其它方法，實驗結果可發現用以建立具強健性的語音偵測系統的 BSE 及 RLF 特徵參數可成功地特徵語音特性且不易受噪音程度變動。事實上，語音偵測技術使也在噪音估測器中扮演非常重要的角色；一般都採用語音偵測系統的技術作為判斷何時追蹤噪音頻譜變動的指示器。為了針對噪音程度極遽變動情況下，所提出的噪音估測器加入以熵為基礎的語音偵測技術並以疊代平均的方法及可調適的平滑因子為基礎。

而在另一種語音偵測系統，我們利用語音的暫態及非穩定性的特性最為擷取語音訊號的依據，採以小波作為訊號的分析。首先，離散小波轉換將輸入訊號分成四個不均勻大小的子頻帶，而在每個子頻帶上採用一種非線性(non-linear)的 Teager 能量運算(Teager Energy Operator, TEO)以有效抑制噪音在各子頻帶的影響，而另一優點就是有助於子頻帶自我相關函示(spectral auto-correlation function,

SACF)之結果。為了量化個子頻帶上的自我相關函示採用 Mean-Delta(MD)運算以估測各頻帶的週期強度，最後並相加各子頻帶的 MDSACF 參數以建立一個以小波為基礎的強健性特徵參數。為了建立完整的語音偵測系統，我們採用一個可適性臨界方式作為判斷語音偵測結果的機制。相較於其他方法，實驗結果證實了以小波為基礎的語音偵測方法可提供在可變噪音程度下的強健性且具高效率及易實現的方法。



A STUDY OF FREQUENCY BAND AND WAVELET ANALYSIS FOR ROBUST VOICE ACTIVITY DETECTION

Student: Kun-Ching Wang

Advisor: Dr. Bing-Fei Wu

Department of Electrical and Control Engineering

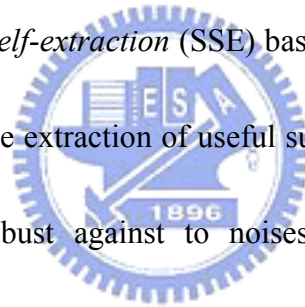
National Chiao Tung University



ABSTRACT

This dissertation mainly addresses the problem of a voice activity detection (VAD) failed in poor signal-to-noise ratio (SNR) and in dynamically time-varying background. So far, the commonly used VAD algorithms always assume that the background noise level is stationary. Since the feature extractions from conventional algorithms are closely depended on the estimation of energy level, the corresponding performances are easily contaminated by the variable noise-level. For example, may usually exit in car due to movements, engine running, speed change, braking, slam, etc. To solve the problem, the VAD algorithms based on two types of robust feature parameters are proposed in turn. In the first presented approach, it is found that the

nature of banded line is highly efficient, compact representation for the time-varying characteristics of speech signals according to the appearance of banded line on voice spectrogram resulted from formant frequency. For frequency band analysis, an entropy-based VAD is presented herein. First, the input signal is decomposed into 32 uniform subbands to locate the formant frequency bands. A measure of entropy defined in subband domain, regarded as *banded spectrum entropy* (BSE) parameter, is then proposed to sufficiently exploit the inherent nature of banded lines on voice spectrogram. Due to that the some decomposed subbands can be contaminated by noise, a strategy of *subband self-extraction* (SSE) based on adaptive threshold skill is presented herein to execute the extraction of useful subbands with time and is further used to let the BSE be robust against to noises. The banded lines on voice spectrogram, in practice, are only suitable for characterizing voiced speech. In order to enhance the part of unvoiced speech, *the ratio of low-band energy to full-band energy* (RLF) is presented to discriminating the unvoiced sound from background noises. Compare to other VAD approaches, experimental results shown that the two BSE and RLF parameters used for determining voice activity successfully exploit the characteristic of speech signal and is nearly robust against variable noise level. A technology of VAD, in practice, plays an essential role in noise spectrum estimator. The VAD scheme is frequently employed into noise spectrum estimator as an



indicator of updating noise spectrum. Enclosed herein the proposed noise spectrum estimation employs an entropy-based VAD above mentioned as an indicator of updating noise spectrum. In addition, a recursive averaging-based formula and an adaptive smoothing factor are then involved herein for quickly adapting to variable level of noise.

In the alternative VAD method, wavelet analysis is used for extracting speech signals to further exploit the transient components and non-stationary property. First, we divide the input signal into four non-uniform subbands via discrete wavelet transform (DWT). In addition, a nonlinear Teager energy operator (TEO) is then utilized into each subband signals. We show that the TEO can decrease the influence of noise on subbands significantly. Besides, the other advantage is suitable for the result of subband auto-correlation function (SACF). To obtain the amount of periodicity, a Mean-Delta (MD) operator is then applied into SACF on each subband. Summing up the all MDSACFs derived from each decomposed subband, a robust wavelet-based feature parameter is then proposed. Finally, we adopt an adaptive threshold method as VAD decision to form a complete VAD. The simulation result shows the wavelet-based VAD is robust against changing noise level and is an efficient and simple approach as comparing with other methods.

ACKNOWLEDGEMENT

回顧博士班研究的過程，在這漫長的的求學中，曾經歷過很多挫折的經驗，更曾經有過放棄追求博士學位的念頭。幸有指導教授 吳炳飛老師的體諒及支持，讓學生得以順利完成學位。除了教授的嚴謹教學態度，他對於學生的幫助更是義不容辭。而教授的求學生涯、成長經歷以及奮鬥過程，更是學生一面鏡子，值得終身學習。

首先感謝口試委員 王小川教授、王逸如教授、吳俊德教授、張森嘉博士以及陸儀斌教授，給予學生論文相當寶貴的意見。

特別謝謝彭昭暉 學長，謝謝你在我低潮的時候給予我忠懇的建議及鼓勵，並感謝CSSP實驗室的所有學長及學弟妹們。

我的內人 陳昱卉女士，謝謝妳辛苦地陪我走過14年求學的生涯，更感謝為我生個可愛又好動的兒子 王翊丞及乖巧的女兒 王芊涵。在這段時間，讓我回味好久不曾的家的溫馨。

母親 張秀吻女士，在我的心目中，您一直無怨無悔地的付出。因為有您的付出，使我可以專注於學業上。感謝岳父 陳國華先生、岳母 黃月華女士，因為有您們分擔照顧小孩子，減低我與內人不少的辛勞!!

願將這份榮耀與您們一起分享!!

最後，也獻給已往生將近一年的父親 王文豐先生，希望父親在天有知，與兒子分享這份喜悅!!

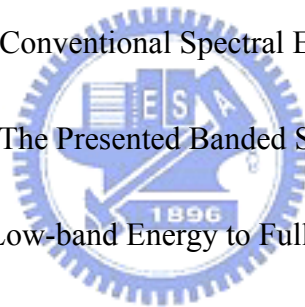
坤卿 於交通大學電機與控制系CSSP實驗室 2005/10/6

CONENTS

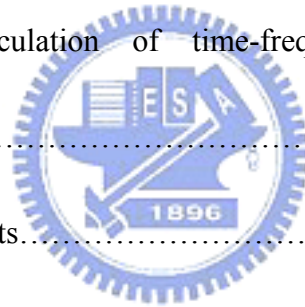
ABSTRACT (Chinese)	i
ABSTRACT (English)	iii
ACKNOWLEDGEMENT	v
TABLE OF CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	xii

1 INTRODUCTION	1
1.1 Motivation.....	1
1.2 The Classification of Speech Signal.....	2
1.3 The Categories of Background Acoustical Noises.....	3
1.4 Prior Art.....	5
1.5 Objectives.....	8
1.5.1 Frequency Band Analysis for Voice Activity Detection Using A Measure of Banded Spectral Entropy.....	8
1.5.2 Recursive Noise Estimation with A VAD for Highly Non-stationary Environments.....	9

1.5.3	Voice Activity Detection Based on Wavelet Analysis Using a Measure of Auto-Correlation Function and Teager Energy Operator.....	10
1.6	Organization of This Dissertation.....	12
2	FREQUENCY BAND ANALYSIS FOR VOICE ACTIVITY DETECTION USING A MEASURE OF BANDED SPECTRAL ENTROPY.....	13
2.1	Introduction.....	14
2.2	The Robust Feature Parameter.....	18
2.2.1	Motivation.....	18
2.2.2	A Measure of Conventional Spectral Entropy (SE).....	20
2.2.3	A Measure of The Presented Banded Spectral Entropy (BSE).....	21
2.2.4	The Ratio of Low-band Energy to Full-band Energy (RLF).....	26
2.3	The Proposed Entropy-based VAD Algorithm.....	27
2.3.1	An Adaptive Threshold Method.....	29
2.3.2	The Strategy of Subband Self-Extraction (SSE).....	30
2.3.3	The Block Diagram.....	32
2.4	Evaluation.....	34
2.4.1	Artificially Added Noise.....	35
2.4.2	Recordings in a Car.....	37
2.5	Discussion & Future Work.....	38



3	A SINGLE CHANNEL NOISE ESTIMATOR WITH RAPID ADAPTATION IN VARIABLE-LEVEL OF NOISY ENVIRONMENTS.....	55
3.1	Introduction.....	56
3.2	Proposed Noise Estimation Algorithm.....	59
3.2.1	A modified entropy-based VAD.....	59
3.2.2	Update of noise spectrum during voice absent frames.....	60
3.2.3	Update of noise spectrum during voice active frames.....	60
3.2.3.1	Tracking the local minimum.....	61
3.2.3.2	The calculation of time-frequency dependent smoothing factor.....	61
3.3	Experimental Results.....	62
3.4	Discussion.....	64
4	VOICE ACTIVITY DETECTION BASED ON WAVELET ANALYSIS USING A MEASURE OF AUTO-CORRELATION FUNCTION AND TEAGER ENERGY OPERATOR.....	70
4.1	Introduction.....	71
4.2	Wavelet Transform.....	74
4.3	Teager Energy Operator (TEO).....	76
4.4	The Robust Feature Extraction Derived From MDSACF.....	78

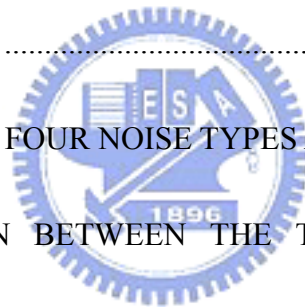


4.5	Proposed Voice Activity Detection (VAD) Algorithm.....	80
4.6	EXPERIMENTAL RESULTS.....	84
4.6.1	Test Environment and Noisy Speech Database.....	84
4.6.2	Evaluation in Stationary Noise.....	85
4.6.3	Evaluation in Non-stationary Noise.....	87
4.7	Discussion & Future Work.....	88
4.7.1	Comparison.....	89
4.7.2	Future Work.....	92
5	CONCLUSIONS.....	100
	BIBLIOGRAPHY.....	103
	VITA.....	111
	PUBLICATION LIST.....	112



LIST OF TABLES

TABLE 2-I	COMPARISON BETWEEN THE PROPOSED ENTROPY-BASED VAD AND ATF-BASED VAD [7] UNDER VARIOUS NOISE CONDITIONS	41
TABLE 2-II	COMPARISON BETWEEN THE PROPOSED ENTROPY-BASED VAD AND ATF-BASED VAD [7] UNDER VARIOUS NOISE CONDITIONS	41
TABLE 3-I	SEGERR FOR FOUR NOISE TYPES AND LEVELS.....	65
TABLE 4-I	COMPARISON BETWEEN THE THREE VAD ALGORITHMS TESTING IN VARIOUS NOISE CONDITIONS	93
TABLE 4-II	SUBJECTIVE EVALUATION OF LISTENING TEST	93
TABLE 4-III	ILLUSTRATION OF EFFICIENCY FOR THE FOUR VAD.....	93
TABLE 5-I	COMPARISON BETWEEN PROPOSED ENTROPY-BASED VAD AND WAVELET-BASED VAD	102



LIST OF FIGURES

Fig. 2-1	Inherent characteristic of banded lines (formant traces) only appears on voice-active spectrogram	42
Fig. 2-2	Illustration of the banded lines existing in various types of noises	43
Fig. 2-3	Nature of banded lines on voice-active spectrogram.....	44
Fig. 2-4	Power distributions of all 32 uniform subbands with the same entropy (logarithmic BSE=21.9601).....	45
Fig. 2-5	Illustration of characterizing speech signals by using entropy-based feature parameter.....	46
Fig. 2-6	Different types of noises focusing on different frequency bands	47
Fig. 2-7	Relation between the number of useful subbands and NMinBE parameter	48
Fig. 2-8	Illustration of the efficiency of NMinBE parameter for applying in BSE parameter.....	49
Fig. 2-9	An adaptive threshold method for VAD decision	50
Fig. 2-10	Flowchart of SSE strategy for automatically extracting the useful subbands	51

Fig. 2-11	Block diagram of the proposed entropy-based VAD algorithm.....	51
Fig. 2-12	Result of RLF measure tested in recorded speech sentence /start/	52
Fig. 2-13	Measurement of the two BSE and RLF parameters.....	53
Fig. 2-14	Comparison between different feature parameters for VAD algorithm testing an utterance with musical background noise inside a car	54
Fig. 3-1	Flow diagram of the proposed algorithm for updating noise estimate	66
Fig. 3-2	Flow diagram of a modified entropy-based VAD	67
Fig. 3-3	Sigmoid function for computing a time-frequency smoothing factor	68
Fig. 3-4	Tracking capability of the noise estimator tested in noisy speech with a sudden increase in the level of factory noise	68
Fig. 3-5	S/N ratio in each frequency subband	69
Fig. 4-1	Discrete wavelet transform (DWT) using filter banks.....	94
Fig. 4-2	18-tap Daubechies Wavelets	94
Fig. 4-3	Structure of three-level wavelet decomposition.....	95
Fig. 4-4	Enhancement of formant information by using TEO.....	95
Fig. 4-5	Illustration of TEO applying into wavelet coefficients derived from each subband	96
Fig. 4-6	Examples of normalized SACF for voiced sound, unvoiced sound and white noise on each subband.....	96

Fig. 4-7 Examples of the SAE parameters without band-decomposition and derived from four subbands 97

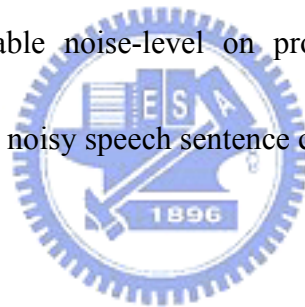
Fig. 4-8 Block diagram of proposed wavelet-based VAD..... 97

Fig. 4-9 Adaptive thresholding strategy for extracting the boundary of voice activity..... 98

Fig. 4-10 Comparisons among VAS, MD and proposed SAE feature parameters... 98

Fig. 4-11 Illustration of performance for Stegmann’s VAD [53] containing four energy-based parameter and VAD decision 99

Fig. 4-12 Effects of a variable noise-level on proposed SAE and Chen’s VAS parameters under a noisy speech sentence consisting continuous words . 99



CHAPTER 1

INTRODUCTION

1.1 Motivation

The purpose of voice activity detection (VAD) is the determination of the presence or absence of a voice component in a given signal, especially the determination of the beginnings and endings of voice segments, and it is also called speech detection, endpoints detection, and speech/non-speech segmentation. Such a technique is one of the essential parts for entire speech processing and required in many applications such as mobile telecommunication for discontinuous transmission (DTX) mode [1], noise reduction for speech enhancement [2], speech coding [3], [50], and speech recognition [4]. In GSM (Global System for Mobil communication cellular) system, a VAD scheme for DTX is required to determine whether a slice of speech waveform is voice or silence. After that, only the “voice” slices are transmitted by DTX to lengthen the battery life of the terminals by suppressing the overall transmitting data [5], [6].

Similarly, in speech recognition system a VAD strategy used as front-end process can improve the recognition ratio and reduce the computing power waste induced by incorrect speech detection even in the presence of noise. Besides, a good use of VAD can be used in speech coding system to control the average bit rate and the overall coding quality of speech in a variable bit rate [3].

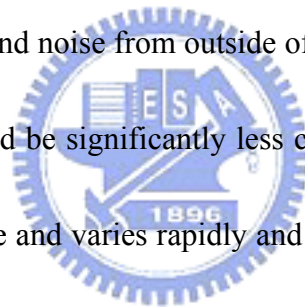
1.2 The Classification of Speech Signal

Speech sounds can be classified into three distinct classes, voice sounds, unvoiced sounds, and silences which are resulted from their mechanism of speech production [8]. Voiced sounds are generated by forcing air through the glottis or an opening between the vocal folds. The tension of the vocal cords is adjusted so that they vibrate in oscillatory fashion. Thus, voiced sounds show periodicity. All the vowels including the semivowels and the diphthongs are voiced sounds. Unvoiced sounds are produced by forming a constriction at some point in the vocal tract, and forcing air through the formed constriction at a high velocity to generate turbulence. Unlike voiced sounds, unvoiced sounds do not have any prominent periodic components. Silences are the parts of the speech sequence during which no sound is produced by the human speech production mechanism. In many ordinary environments, silences in the speech sequence are dominated by the background noise which varies dynamically in energy

level and spectral characteristics.

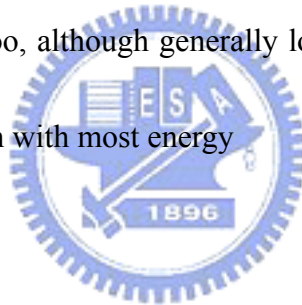
1.3 The Categories of Background Acoustical Noises

In general, the speech signal is corrupted by additive background acoustical noises that we can divide into two categories based on the environment. The first category is background noise from an office-type environment. The office-type environment contains some, but not a lot of, background noise. Office-type background noise varies slowly both in terms of volume and statistical characteristics. The second category is dynamic background noise from outside office environment. This category can be frequently occurred and be significantly less controlled; the background noise exhibits a wide dynamic range and varies rapidly and unpredictably. Besides, impulse noise may perturb the waveform of speech signals. The above-mentioned categories are illustrated as below:



- a. Office-type background noise
 - i. also called fixed-level noise
 - ii. short-term quasi-stationary noise
 - iii. strictly speaking, background noise is present throughout the entire signal, although it is mostly embedded in the voice signal
 - iv. associated with ambient background noise

- v. can be considered as low-energy, white noise
 - vi. is subject to a statistical characterization in two steps on the basis of its mean energy, the standard deviation of the energy, and mean zero-crossings.
- b. Dynamic background noise
- i. also called variable-level noise or non-stationary noise
 - ii. sporadic noise from a multitude of sources
 - iii. may appear at any moment in record
 - iv. duration is very variable
 - v. energy is variable too, although generally lower in comparison with that of the section of speech with most energy
- c. Impulse noise
- i. perturbations of the waveform, primarily in the last part of the speech signal, due to exhalation
 - ii. noises made by the tongue and/or lips in preparation for speech
 - iii. does not always appear
 - iv. energy is variable: may be confused with other sporadic noises, or even speech
 - v. can be dealt with by using time duration



So far, the existing VAD algorithms all assumed the additive noise as an office-type

background noise. However, the so-called non-stationary noises frequently exist in recordings. So, the detection of voice activity in dynamical noise environment is a challenging task.

1.4 Prior Art

The accurate determination of speech in the presence of background noise plays a significant role in many areas of speech processing. Over the past decade numerous studies have been proposed for the developments of the robust VAD algorithms operating in adverse acoustic noisy environments. In general, for the establishment of a complete VAD algorithm it mainly comprises two parts shown as bellows:

- *Parameter extraction*: Relevant parameters are extracted from the speech signal.

In order to allow a good detection of the speech signal regions, the chosen parameters have to show a discriminability between speech and non-speech segments.

- *Thresholding*: A threshold is applied to the extracted parameter in order to divide the speech signal between speech and non-speech segments. This threshold can be fixed or adaptive. In general, the adaptive threshold is used to adapt to variable noise level.

Most of the early algorithms are based on the feature parameters of short-time

energy [10], zero-crossing rate (ZCR) [12], [13], timing, pitch information [11] and the LPC coefficients [9]. Cepstral features [14], adaptive noise modeling of voice signals [15] and the periodicity measure [16] are some of the recent ideas in VAD designs. Unfortunately, these algorithms have some problems at low signal-to-noise ratios (SNRs), especially when the noise is non-stationary. For example, the energy-based estimate and ZCR are inadequate for distinguishing speech from noise at low SNRs. The performances of energy-based VADs degrade significantly when the speech level is smaller than the noise level. Moreover, the ZCR is highly sensitive to various types of noise. Additionally, the reliability of the LPC coefficients have been observed to depend strongly on the noise in adverse environment and not particularly suitable for nasal sound, fricative, etc. Although pitch information can help to detect speech, extracting the correct pitch in noisy environments is difficult. Some of the algorithms perform well in adverse environments; even so, the algorithms must be required more computing complexity. Various procedures for speech detection have been described in the literature so far. Sohn *et al.* [17] presented a VAD algorithm that uses a novel noise spectrum adaptation applying soft decision techniques. The decision rule drew from the generalized likelihood ratio test by assuming the noise statistics to be known a prior estimate. Cho *et al.* [18] presented improved version of the algorithm designed by Sohn. Specifically, Cho presented a

smoothed likelihood ratio test to reduce the detection errors. Furthermore, Beritelli *et al.* [19] developed a fuzzy VAD using a pattern matching block comprising a set of six fuzzy rules. Nemer *et al.* [20] designed a robust algorithm based on higher order statistics (HOS) in the residual domain of the linear prediction coding coefficients (LPC). The International Telecommunication Union-Telecommunications Sector (ITU-T) designed G.729B VAD [21] comprising a set of metrics including line spectral frequencies (LSFs), low band energy, ZCR and full-band energy.

Besides, Chen *et al.* [22] proposed speech detection algorithm using microphone array in a noisy environments. However, the assumption of stationary may result in estimation errors from the above method due to the non-stationary nature of speech signals, the limited length of observed data and the size of the microphone array. To determine the presence or absence of speech, the observed signal statistics in the current frame are compared with the estimated noise statistics according to some decision rules. In [23], using the assumption that the Discrete Fourier Transform (DFT) coefficients of speech and noise are asymptotically independent Gaussian random variable, a statistical model-based VAD has been successfully proposed. Recently, S. Gazor *et al.* [24] proposed a VAD that uses a Laplacian distribution model for speech and outperforms the previous VADs that use a Gaussian model. Those algorithms, however, require extensive training of a Hidden Markov Model with the set of speech

prototypes to be encountered. In general, different applications need different algorithms to meet their specific requirements in terms of computational accuracy, complexity, robustness or sensitivity, response time, etc.

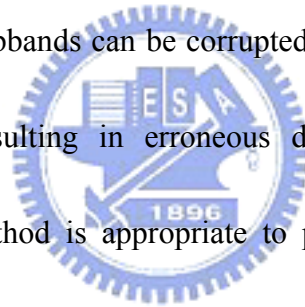
1.5 Objectives

Even if the noise is generated from a computer, an air conditioning system, or an automobile are not perfectly stationary. In practice, the background noise level usually varies with time. In this dissertation we mainly focus to propose a set of robust feature parameter for sufficiently characterizing speech signal even in variable noise level and/or in low SNRs. In addition to feature extraction, the adaptive threshold strategy is also considered as a VAD decision to adapt to the changes of the noise conditions. Enclosed herein we suggest a recursive-averaging based noise estimator containing the above-mentioned VAD to estimate noise spectrum continuously, quickly and accurately when existing a rapidly increasing noise level.

1.5.1 Frequency band analysis for voice activity detection using a measure of banded spectral entropy

The nature of banded lines on voice spectrogram generated from formant frequency band is a highly efficient, compact representation of the time-varying characteristics

of speech, especially for voiced sounds. Making good use of the representation, we can determine the detection of voice activity. In fact, the darkness of the corresponding spectrogram reacts where high powers concentrate. Through frequency band analysis, the input signal is firstly decomposed into 32 uniform subbands. The representation of peak and valley within subband power indicates the existence and absence of banded lines. Comparing with other feature parameter given from energy level, a measure of entropy defined in subband domain, regarded as *banded spectrum entropy* (BSE), is then presented to exploit the inherent nature. In practice, the BSE can be invalid when some subbands can be corrupted by noise. To avoid considering the harmful information resulting in erroneous detection of voice activity, an automatic band-selection method is appropriate to perform well in on-line, called *subband self-extraction* (SSE). In addition, *the ratio of low-band energy to full-band energy* (RLF) is presented to discriminating the unvoiced sound from background noises for compensating the limitation of entropy-based measure for modeling unvoiced sounds. Finally, the first approach for determining voice activity is regarded as entropy-based VAD algorithm.



1.5.2 A Noise Estimator with Rapid Adaptation in Variable-Level of Noisy Environments

Enclosed herein we propose a method for tracking the noise spectrum quickly, even when the noise levels suddenly increase. An explicit use of speech/silence detection is needed for estimating noise spectrum. So, the entropy-based VAD mentioned-above is used to continuously classify each frame of speech into the voice active/absent frames, and the noise spectrum estimate is updated using constant smoothing factor for voice absent frames and a time-frequency dependent smoothing factor for voice active frames. The time-frequency dependent smoothing factor is chosen as a Sigmoid function that changes with the voice present probabilities in frequency bins. The voice present probability is determined by computing the ratio of the noisy speech power spectrum to its local minimum. To speed up the minimum tracking, a fast method for tracking the minimum of the noisy speech power spectrum is presented. Additionally, to allow detection with entropy-based VAD under colored noise conditions, we propose to subtract the current spectrum by the estimated noise spectrum iteratively from the previous frame.

1.5.3 Voice activity detection based on wavelet analysis using a measure of auto-correlation function and Teager energy operator

To further exploit the transient components and non-stationary property for extracting speech signals, the alternative VAD approach is based on wavelet analysis.

In this method, the structure of three-layer wavelet decomposition is utilized to decompose speech signal into four non-uniform subbands here. In general, the well-known “Auto-Correlation Function (ACF)” is commonly used to detect periodicity of speech. Herein the ACF is defined in subband domain and then called as “subband auto-correlation function (SACF)”. In fact, the voiced sound has more significant periodicity than unvoiced sound and noise signal and the periodicity almost concentrates low frequency bands. So, we let the low frequency bands have high resolution to enhance the periodic property by decomposing only low band on each level. In addition, a nonlinear Teager energy operator (TEO) is then utilized into each subband signals. We show that the TEO can enhance the discrimination between speech and noise, and TEO is beneficial for the result of auto-correlation function (ACF) since it can provide a better representation of formant information.

Finally, to count accurately the intensity of periodicity on the envelope of the SACF on each subband, a Mean-Delta (MD) method [47] is utilized on each subband. So, the *sum of Mean-Delta values of Subband Auto-Correlation Function (MDSACF)* derived from the wavelet coefficients of three detailed scales and one appropriated scale is defined as a new robust feature parameter and called as “speech activity envelope (SAE)”. Comprising an adaptive threshold strategy, a robust wavelet-based VAD approach is further presented for extracting the boundary of voice activity

especially in variable noise level and low SNRs.

1.6 Organization of This Dissertation

The dissertation is organized as follows. In Chapter 2, we would illustrate the derivation of the two measures of BSE and RLF. In addition, a *subband self-extraction* (SSE) strategy is shown to automatically select useful information on some subband for compensate a complete measure of BSE. Afterwards, Chapter 3 would state a noise spectrum estimator with rapid adaptation algorithm for highly non-stationary noises. The above-mentioned VAD will be modified to employ into a noise spectrum estimator as an indicator of updating noise. Continuously, Chapter 4 presents the alternative VAD approach based on wavelet analysis for detecting voice activity. The Teager energy and mean-delta (MD) operators are respectively employed into auto-correlation function (ACF) of each subband to form a new SMDSACF (*sum of mean delta of subband auto-correlation function*) parameter. Ultimately, Chapter 5 summarizes our conclusion and gives some future developments.

CHAPTER 2

FREQUENCY BAND ANALYSIS FOR VOICE ACTIVITY DETECTION USING A MEASURE OF BANDED SPECTRAL ENTROPY



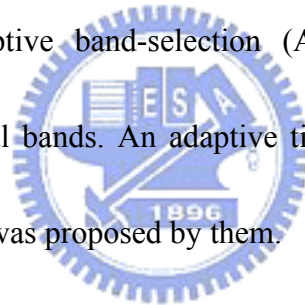
The formant frequency representation is a highly efficient, compact representation of the time-varying characteristics of speech, especially for voiced sounds. In addition, the magnitude arrangement of spectral response during formant frequency is relatively important for characterizing speech signals. The so-called *Banded Spectral Entropy* (BSE), which is a measure of spectral entropy defined in subband domain, is then presented for extracting voice activity. In fact, noises can focus on some subbands to contaminate the useful information that results in error decision of detecting voice activity with BSE. In order to compensate this finds, a method of automatically selecting subband is then presented to meet the requirement, which is regarding as *subband self-extraction* (SSE). Besides, the *ratio of low-band energy to full-band*

energy (RLF) is presented to discriminating the unvoiced sound from background noises since entropy-based measure can provide only for detection of unvoiced sounds.

2.1 Introduction

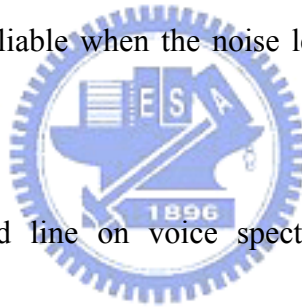
A feature parameter that can sufficiently characterize speech signals or be robust against the highly noisy environments is relatively required. So far, the current algorithms are based on short-time or spectral energy, zero-crossing rate (ZCR) and duration parameters [25]-[27]. All of these parameters, however, are rather sensitive to noise and cannot fully specify the characteristics of a speech signal. For example, the energy-based parameter and ZCR are not sufficient to distinguish a speech from a noise at low SNRs. In particular, the ZCR is very sensitive to various types of noise. Several other parameters have also been proposed, including linear prediction coefficients (LPCs), Cepstral coefficients and pitch [9], [11], [14]. Although these parameters are quite effective in expressing the characteristics of speech signals, the performance of VAD using such parameters remains poor in adverse environments. The reliability of the LPCs has been observed to depend strongly on the noise in adverse environment. Pitch information can help to detect speech; even so, extracting the correct pitch in noisy environments is difficult. Additionally, some algorithms

cannot be implemented for practical applications due to their high computational complexity, even though they perform well [28]. Among such approaches, however, Junqua *et al.* [29] proposed a time-frequency (TF) parameter to detect speech, which assumes that frequency information in the frequency ranges 250-3500 Hz is less contaminated by noise. The TF parameter is composed of both frequency energy in the fixed frequency bands and time energy. Based on the motivation that the frequency energies of various types of noise are concentrated in different frequency bands, Wu *et al.* [7] used the multi-band technique to analyze noisy speech signals, and then proposed an adaptive band-selection (ABS) method to cancel noise effectively by selecting useful bands. An adaptive time-frequency (ATF) parameter extended from TF parameter was proposed by them.



Although the ATF-based algorithm outperforms several algorithms commonly used for detecting voice activity in the presence of various types of noise, it cannot be reliably implemented in practical environments. It is found that the selection of useful bands depends on the information of an entire recorded signal. Additionally, the ATF parameter is also energy-based parameter and therefore less reliable in the presence of non-stationary noise or in a changing noise level. J. L. Shen *et al.* [30] firstly used the entropy-based parameter to detect speech signals. Their study indicated that the spectral entropy of a speech segment differed significantly from that of a noise

segment. In fact, the result of spectral entropy relies on the variance of spectral magnitude to distinguish a speech signal from a noise signal, but the variance of spectral magnitude depends strongly on the noisy environments. L. S. Huang [31] integrated both the time energy and spectral entropy to form a new feature parameter (EE-feature), since the spectral entropy failed under multi-talker babble and background music, but the energy performed well because of its additive property: the energy of the sum of speech plus noise always exceed the energy of noise. Although the EE-feature parameter proposed by L. S. Huang improved the endpoint detection under babble noise, it is unreliable when the noise level greatly exceeds the speech level.



The appearance of banded line on voice spectrogram resulted from formant frequency is a highly efficient, compact representation of the time-varying characteristics of speech, especially for voiced sounds. Since the locations of banded lines reveal that high powers concentrate on the some frequency bands, the band decomposition is used for locating formant frequency components by obtaining peaks while non-formant frequency components are characterized by obtaining valleys. It has been sufficient to display the location of power of formant frequency when the bandwidth of each subband is approximately 125 Hz [8]. A measure of entropy is defined in subband domain and is regarded as *banded spectrum entropy* (BSE)

parameter. In fact, the magnitude arrangement of spectral response during formant frequency is alternative relatively important factor for characterizing speech signals. So, a set of weighting factors among those subbands are also employed into BSE measure to discriminate the magnitude arrangement between speech signals and noise. According to the experimental result from Wu *et al.* [7], some subbands contaminated by noise can provide harmful information resulting in error decision of voice activity detection (VAD) with BSE. So, an automatic band-selection method derived from the refined version of the adaptive band selection (ABS) method proposed by Wu *et al.* is preferable to perform well in on-line, called *subband self-extraction* (SSE). In order to compensate the limitation of BSE measure for modeling unvoiced sounds, *the ratio of low-band energy to full-band energy* (RLF) is presented to discriminating the unvoiced sound from background noises.

This Chapter 2 is organized as the followings. Section 2.2 will introduce the theory of entropy. What is the motivation of using the entropy measure to describe the nature of banded lines on voice spectrogram? Additionally, the proposed feature parameters are stated, respectively. In Section 2.3, we derive the so-called *subband self-extraction* (SSE) method, which is extended from ABS and can adaptively select useful bands in on-line. And then, the procedure for implementing the proposed entropy-based VAD algorithm based on the measures of BSE and RLF and the strategy of SSE is outlined.

Section 2.4 discusses the performance of the proposed VAD algorithm under various noise conditions and compares its performance with that of ATF-based one. Finally, Section 2.5 summarizes the findings and discusses possible directions for future work.

2.2 The Robust Feature Parameters

This section introduces the theory of entropy and further shows the motivation of using the entropy for detecting speech. In addition, the robust feature parameters will be illustrated herein in detail.

2.2.1 Motivation



Fig. 2-1 displays that the waveform of a mixed signal comprising vehicle noise, multi-talker babble noise, factory noise, speech, and white noise and the corresponding spectrogram. Regarding to Fig. 2-1 (b), the voice-active spectrogram is dominated by the inherent nature of banded lines (or called *formant traces*). It is found that the nature is able to sufficiently discriminate speech signal from background noise. Fig. 2-2 displays the spectrograms of clean speech and noisy speech with four kinds of noise at 0 dB. In this figure, the nature of banded lines on voice spectrogram is seen to be existed against various types of additive noise. So, the formant frequency representation is a highly efficient, compact representation of the

time-varying characteristics of speech. The following statements will show how to use the representation of banded lines to detect voice activity by using a measure of entropy.

Entropy, firstly used in information theory by C. Shannon [32], is regarded as the amount of information that must be provided about a random signal x in order to specify it uniquely. It measures the degree of organization (uncertainty) of the signal and is defined by

$$H(x) = \sum_k P(x_k) \cdot \log[1/P(x_k)], \quad (2-1)$$

where $x = \{x_k\}_{0 \leq k \leq N-1}$ and $P(x_k)$ is the probability of x_k .

How to use the definition of entropy for characterizing speech signal? Regarding to Fig. 2-3, the waveform of a Mandarin digit “eight” uttered by native speaker and the corresponding spectrogram are shown in Fig. 2-3(a)-(b), respectively. Since the pitch varies continuously within a speech segment for speech production, the banded lines on voice-active spectrogram are also continuous. When such a clear set of banded lines exist in some frequency bands for a long enough time, the voice activity can be quite certainly presented [33]. Fig. 2-2(c)–(d) show the spectrum magnitude of voice activity obtained by the short-time Fourier transform (STFT) over a solid-line region in Fig. 2-3(a) and that of voice inactivity obtained segment by STFT over a dashed-line region in Fig. 2-3(a), respectively. Inspecting the difference between the

spectrum magnitude of voice activity and that of voice inactivity, we can regard that the amount of variance (uncertainty) from the spectral magnitude during voice activity indeed exceeds that during voice-inactivity. Consequently, a measure of spectral entropy can be used for discriminating the spectral difference between during voice activity and during voice inactivity even if noise level is greater than speech level.

2.2.2 A Measure of Conventional Spectral Entropy (SE)

J. L. Shen *et al.* [30] firstly used a measure of entropy for detecting speech segments under adverse conditions. The measure was defined in spectral domain and named as spectral entropy (SE). Their experimental results have been revealed that the result of SE during voice activity differs from that during voice-inactivity. The deviation for calculating SE parameter is described as follows.

The STFT of a given time frame $s(n, l)$ is accomplished by

$$X(\omega, l) = \sum_{n=0}^{N_w-1} W(n) \cdot s(n, l) \cdot \exp(-j2\pi\omega n / N_w), \quad 0 \leq \omega \leq N_w - 1, \quad (2-2)$$

where $X(\omega, l)$ represents the spectral magnitude of the ω^{th} frequency bin of the l^{th} frame. N_w is the total number of frequency bins in STFT for each frequency frame. $W(n)$ is a Hamming window. The spectral energy of each frame, $X_{\text{energy}}(\omega, l)$, is described as follows.

$$X_{\text{energy}}(\omega, l) = |X(\omega, l)|^2, \quad 0 \leq \omega \leq N_w/2 - 1. \quad (2-3)$$

Then, the probability associated with each spectral energy component, $P(i,l)$, can be estimated by normalizing:

$$P(k,l) = \frac{X_{energy}(k,l)}{\sum_{\omega=1}^{N_w/2-1} X_{energy}(\omega,l)}, \quad 0 \leq k \leq N_w/2-1. \quad (2-4)$$

Following normalization, the corresponding SE, $H(l)$, is defined as follows.

$$H(l) = \sum_{\omega=1}^{N_w/2-1} P(\omega,l) \cdot \log[1/P(\omega,l)]. \quad (2-5)$$

2.2.3 A Measure of The Proposed Banded Spectrum Entropy (BSE)

In fact, the SE measure performs well in white or quasi-white noise, but fails in colored noise. The magnitude associated with each frequency bin is easily contaminated by noise. This results in degradation of the efficiency of VAD performing in seriously low SNRs. So, frequency band analysis is employed into a measure of SE for improving the robustness.

Next, we decompose the input signals into 32 uniform subbands. The subband energy, $E_b(m,l)$, is given by

$$E_b(m,l) = \sum_{\omega=(m-1) \times \frac{N_w/2}{N_b}}^{(m-1) \times \frac{N_w/2}{N_b} + \left(\frac{N_w/2}{N_b} - 1\right)} X_{energy}(\omega,l), \quad 1 \leq m \leq N_b, \quad (2-6)$$

where N_b is the number of total decomposed subband on each frame. The limits of the summation denote the boundary of each subband. For example, if $m = 1^{th}$, the boundary of the first subband means that k is 0 to 3.

Consequently, we modify (2-4) as (2-7) shown as below:

$$P_b(m,l) = E_b(m,l) / \sum_{k=1}^{N_b} E_b(k,l), \quad 1 \leq m \leq N_b. \quad (2-7)$$

where $P_b(\omega,l)$ is the probability associated with band energy.

Then, the measure of banded spectral entropy (BSE), $H_b(l)$, is given by

$$H_b(l) = \sum_{m=1}^{N_b} P_b(m,l) \cdot \log[1/P_b(m,l)]. \quad (2-8)$$

In fact, the magnitude arrangement of spectral response during formant frequency is alternative relatively important factor for characterizing speech signals. So, the order of subband power must be considered for describing the magnitude arrangement.

The well-known measure of entropy, however, cannot indicate a distribution (spatial information) of the data sequence. Fig. 2-4 denotes the power distribution of all decomposed 32 subband during voice-active frames and voice-absent frames. It illustrates that the classical spectral entropy is not able to discriminate the difference between the duration of speech and the duration of non-speech for the distribution of subband energy. The distribution of subband energy during a speech segment and that during a non-speech segment are shown in Fig. 2-4(a) and Fig. 2-4(b). Measuring on the two kinds of distributions of subband energy by spectral entropy, we can get the same logarithmic BSE value ($H=21.9601$). Major cause is resulted form the invalid description in the nature of banded lines on voice-active spectrogram.

To solve this problem, a set of weighting factors $W(m,l)$ among those subbands

are employed into BSE measure to discriminate the magnitude arrangement between speech signals and noise.

$$W(m,l) = \frac{1}{3} \sum_{i=m-1}^{m+1} (e_i - \bar{\mu})^2, \quad \bar{\mu} = \frac{1}{3} \sum_{i=m-1}^{m+1} e_i, \quad (2-9)$$

$$\begin{aligned} e_{m-1} &= \frac{\min[P_b(l)]}{P_b(m-1,l)}, \\ e_m &= \frac{\min[P_b(l)]}{P_b(m,l)}, \\ e_{m+1} &= \frac{\min[P_b(l)]}{P_b(m+1,l)}, \end{aligned} \quad (2-10)$$

where e_m means the biased energy of subbands. $W(m,l)$ represents the variance amount among the three biased energy of subbands. $\min[P_b(l)]$ is the minimal spectral energy among all 32 subbands and regarded as a normalized factor. If $W(m,l)$ on the m^{th} subband is great, this implies that the banded line may be located around these subbands from the $(m-1)^{\text{th}}$ subband to the $(m+1)^{\text{th}}$ subband. Conversely, if the $W(m,l)$ value is low, it indicates that the banded line is not located on these subbands. By using the set of weighting factors $W(m,l)$, the magnitude arrangement of spectral response can be explicated the nature of banded lines on voice-active spectrogram and (2-8) is modified as (2-11) as follows:

$$H_b(l) = \sum_{m=1}^{N_b} W(m,l) \cdot P_b(m,l) \cdot \log[1/P_b(m,l)]. \quad (2-11)$$

Fig. 2-5 clearly indicates that the proposed BSE parameter, using the method of band decomposition and a set of weighting factors $W(m,l)$, more sufficiently characterizes the speech signals than other entropy-based parameter so that SE

parameter proposed J. L. Shen *et al.* [30].

In fact, the frequency energies of difference types of noise are concentrated on different frequency bands [7], as shown in Fig. 2-6. This observation demonstrates that the subbands with larger noisy energy more contaminate the useful frequency information than do the other bands. The bands with larger noisy energy can be regarded as harmful subbands afterwards and must be discarded accurately to yield more accurate frequency information. Although the BSE remains a good feature parameter, the detection sometimes fails at seriously low SNRs, especially when relatively harmful bands are involved. How to discard the harmful subbands or preserve the useful subbands becomes a serious task. The number of harmful bands (or useful subbands) is relatively related to the background noise level [7]. To easily estimate the background noise level, we extend the MiMSB parameter [34] to adaptively choose one subband with minimum energy for estimating the varying noise level roughly. Regardless of changing level of noise, a *normalized minimum band energy* (NMinBE) parameter is proposed to estimate the background noise level for precisely deciding the number of useful subbands. The NMinBE parameter is determined as follows:

$$\text{NMinBE}(l) = -\log[\min\{E_b(m, l)\} / \sum_{m=1}^{N_b} E_b(m, l)], \quad (2-12)$$

where the $\min\{\cdot\}$ operator selects the minimum band energy among all 32 subbands

energies for a given frame, and $\log[\cdot]$ is the logarithmic operation. The number of useful subbands, $N_{ub}(l)$, required to yield reliable information. Fig. 2-7 displays the relation between $N_{ub}(l)$ and $N_{MinBE}(l)$.

$$N_{ub}(l) = \begin{cases} 30 & N_{MinBE}(l) < 5; \\ [36.5 + \frac{N_{MinBE}(l)}{(25-5)} \times (4-30)] & 5 < N_{MinBE}(l) < 25; \\ 4 & \text{otherwise.} \end{cases} \quad (2-13)$$

It is observed that a large $N_{ub}(l)$ should be used at a low noise level (corresponding to a high SNR), and a small $N_{ub}(l)$ should be used at a high noise level (corresponding to a low SNR). According to (2-13), for the l^{th} frame the first $(32 - N_{ub}(l))$ frequency bands with larger energies are adaptively selected to remove noise component. Finally, the measure of BSE parameter with a strategy of extracting the useful subband is achieved by the follows:

$$H_b(l) = \sum_{m=1}^{N_{ub}(l)} W(m,l) \cdot P_b(m,l) \cdot \log[1/P_b(m,l)]. \quad (2-14)$$

The variable noise level and the varying statistics of noise, in general, results in a varying $N_{ub}(l)$ over an entire signal. According to the relationship between $N_{ub}(l)$ and noise level described in (2-13), we can evaluate the efficiency of a strategy of extracting the useful subband on a measure of BSE. Fig. 2-8(a) and 2-8(b) plots the waveform of someone's saying the Mandarin digit "eight" with increasing-level of factory noise and the corresponding spectrogram, respectively. The voice activity is not easily detected in an adverse environment due to a measure of BSE involving

some harmful subbands, as shown in Fig. 2-8(c). Regarding to the Fig. 2-8(d), it is shown that the proposed NMinBE parameter can reflect the variation of noise level. According to the relation in Fig. 2-7, the number of useful subbands can be determined as shown in Fig. 2-8(e). Fig. 2-8(f) displays that a measure of BSE with a manual extraction of useful subbands can greatly improve the performance of detecting voice activity, especially at variable level of noise. Consequently, how to automatically extract useful subbands with time is crucial and discussed in next section.

2.2.4 The Ratio of Low-band Energy to Full-band Energy (RLF)

Unlike voiced sounds, unvoiced sounds do not have any component of formant frequency. By measuring the energy level, the unvoiced sound is difficultly discriminated from background noise. The majority of unvoiced sounds, however, display string spectral concentration in higher frequency range. The background noise display uniform spectral distribution, It is possible to distinguish between speech-active and background noise by examining the distribution of energy along the frequencies. Low-band energy, $E_{low}(l)$, measured on the below 1000 Hz, is computed as follows:

$$E_{low}(l) = 10 \times \log \left(\sum_{\omega=0}^{\omega=\pi/4} \hat{X}_{energy}(\omega, l) \right), \quad (2-15)$$

where \hat{X}_{energy} is obtained by ignoring the harmful subbands.

Similarly, full-band energy, $E_{full}(l)$, measured on entire frequency bandwidth (0~4000 Hz), is given by

$$E_{full}(l) = 10 \times \log \left(\sum_{\omega=0}^{\omega=\pi} \hat{X}_{energy}(\omega, l) \right). \quad (2-16)$$

So, the ratio of low-band energy to full-band energy, $R_{lf}(l)$, is given by

$$R_{lf}(l) = \frac{E_{low}(l)}{E_{full}(l)}. \quad (2-17)$$

2.3 The Proposed Entropy-based VAD Algorithm

According to [26], the required characteristics of an ideal voice activity detector are reliability, robustness, accuracy, adaptation, simplicity and real-time processing.

Although some existing VAD algorithms are extremely accurate, they all depend on complicated computation and are not reliable in real applications. For example, Wang *et al.* [28] proposed a robust algorithm based on wavelet analysis, but it indeed performs in off-line. E. Nemer *et al.* [20] used higher-order-statistics (HOS) parameter to detect speech, but the calculation of this parameter required too much computing time. Wu *et al.* [7] suggested an adaptive band-selection (ABS) method, which can select useful bands automatically, as noise cancellation to perform ATF parameter well. However, the execution of ABS method depends on the obtainment of all information from the entire recorded signals. Although those algorithms are

inappropriate for practical implementation, some ideas related to those algorithms are adopted herein. The ABS method proposed by Wu *et al.* [7] is strong with respect to noise cancellation. The ABS was used to preserve the useful bands (or discard the harmful bands) for each frame, but the execution of band selection depends on entire recorded signal. The drawbacks of ABS are thus as the following:

-- Firstly, the decision of band selection is not immediately determined. Since the method is an off-line strategy, its decision must be determined by analyzing an entire recorded signal.

-- Secondly, the indexes associated with the harmful bands vary with time for entire recorded signals in practice. However, Wu *et al.* assumed that the indexes of harmful bands were fixed. This assumption does not hold.

So, how to detect whenever the index of harmful bands vary with time is relatively required. Regarding to the Fig. 2-8(f) again, we observe that the selected subbands are not contaminated by noise. The corresponding BSE value in voice absence is small and smoothly and slightly varies with time as comparing with Fig. 2-8(c). Conversely, regarding to Fig. 2-8(c), the selected bands are contaminated by background factory noise. However, the entropy value in voice absence is large and its variation is also violent. In the conclusion, it reveals that the determined entropy value is quite large and violently varying whenever the considered subbands include harmful subbands;

the determined entropy value is small and its variation is very smooth if the considered subbands do not include harmful subbands. This finding provides a hint about how to detect whenever the indexes of harmful bands vary with time.

2.3.1 An Adaptive Threshold Method

In order to extract the harmful subbands automatically, an adaptive thresholding not only provides a decision of VAD, but also determines what time is execution of selecting the useful subbands. Herein, an adaptive threshold with minimal processing delay [35] is employed to on-line work. The work of adaptive threshold method used for a VAD decision is illustrated as below. To adapt to time-varying characteristic of noise, the detection method must sets an adaptive threshold to classify voice-active frames or voice-absent frames. During a short period of initialization, the mean and variance of the logarithmic value of BSE measure is estimated over voice-absence after manual extracting useful subbands. The adaptive speech threshold, T_s , is initially computed from the local noise statistics shown as below

$$T_s = \mu + \alpha \cdot \sigma, \quad (2-18)$$

where μ and σ are the mean and variance of the logarithmic value of BSE respectively, and α is an adjustment coefficient determined in experiment.

Then, the threshold T_s is compared to the value of logarithmic BSE of the previous

frame. Whenever the difference surpasses a specified threshold, voice-active frame is detected. If a given frame is detected to fall in a non-speech period, the speech threshold T_s is updated.

During voice activity, we remain the speech threshold until voice absence. The mean and variance of the logarithmic value of BSE measure are updated as follows.

$$\mu_{new} = \beta \cdot \mu + (1 - \beta) \cdot H_b(l), \quad (2-19)$$

$$\sigma_{new} = \sqrt{|H_{b,mean}^2(l) - \mu_{new}^2|}, \quad (2-20)$$

$$H_{b,mean}^2(l) = \beta \cdot H_{b,mean}^2(l-1) + (1 - \beta) \cdot H_b^2(l), \quad (2-21)$$

$$H_{b,mean}^2(l-1) = \frac{1}{init_period} \sum_{k=1}^{k=init_period} H_{b,mean}^2(l-k), \quad (2-22)$$

where β is also experimentally determined.

Fig. 2-9 depicts that an adaptive threshold method is used as a VAD decision performing in on-line. In this figure, an utterance of the Mandarin word for “one” is together with the logarithmic value of BSE measure and adaptive speech threshold T_s .

The results indicate clearly that the speech threshold T_s is updated in voice absence and maintained in voice activity. In addition, it is found that the value of adaptive threshold is indeed adaptive to time-varying noisy environment.

2.3.2 The Strategy of Subband Self-Extraction (SSE)

What time is execution of selecting the useful subband? In this subsection, the

strategy of *subband self-extraction* (SSE) used for automatically extracting the useful subbands would be illustrated in detail. The execution of SSE depends on the fact that whenever the value of logarithmic BSE exceeds the adaptive threshold. If the value of logarithmic BSE exceeds the adaptive threshold, it means two kinds of probabilistic results. One is that the voice activity is detected. Other is error of VAD detection resulted from the consideration of harmful subbands when computing a measure of BSE.

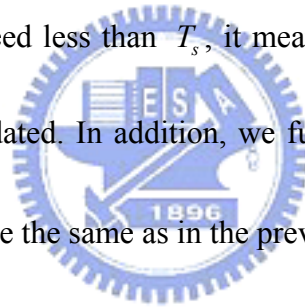
An indicator for *execution of band extraction* (EOBE) parameter is used to stand for binary decision of performing band selection on each frame. The EOBE is defined as follows.

$$EOBE = \begin{cases} \text{high} & H_b(l) > T_s; \\ \text{low} & \text{otherwise.} \end{cases} \quad (2-23)$$

If indicator of EOBE is low for a given frame, it implies that the considered subbands do not include any harmful subbands and the indexes of harmful subbands within noise power do not vary violently with time. In general, the indexes can be maintained from the previous frame to the current frame until EOBE is high so that the task of harmful subbands detection cannot be executed which results in significant decrease of computing complexity. Similarly, if indicator of EOBE is high for a given frame, it implies that the considered subbands maybe include some harmful subbands resulting in an abrupt change in contour of BSE. The indexes of harmful subbands

maybe vary violently with time so that the task of harmful bands detection is performed right now.

Fig 2-10 illustrates the SSE strategy in detail. When the current BSE value with the previous estimated useful subbands exceeds the adaptive threshold T_s , one of two possibilities obtains: one implies that the voice activity is detected, and another means that the BSE value in error due to the inclusion of harmful subbands so that the harmful bands must be detected again. If the re-detected BSE value remains greater than T_s , the voice-active frame is detected and T_s is not updated. Conversely, if the re-detected BSE value is indeed less than T_s , it mean that the voice-absent frame is detected and T_s must be updated. In addition, we further assume that the subbands where noise is concentrated are the same as in the previous frame.



2.3.3 The Block Diagram

The proposed entropy-based VAD algorithm is diagramed in Fig. 2-11 and each of VAD is illustrated in detail.

1) *Decomposition of 32-uniform subband:*

For the frequency band analysis, the input noisy speech is decomposed into 32 uniform subbands as preprocessing of developing a robust feature parameter.

2) *Computation of a set of weighting factors:*

When formulating a measure of BSE, we exploit the magnitude arrangement to discriminate the ambiguity between the short-time spectral magnitude during voice activity and that during voice-inactivity

3) *The strategy of subband self-extraction (SSE):*

In fact, some subbands can be contaminated by noise to result in the harmful information for detecting voice activity. So, automatically extracting useful subbands via the result of adaptive thresholding (AT) would avoid the occurrence of error BSE value.

4) *The computation of BSE measure:*

After the computation of a set of weighting factors and excitation of SSE strategy on each frame, a measure of entropy is defined in the selected subbands.



5) *The computation of RLF measure:*

By ignoring the estimated harmful subbands, the robust RLF measure can be reached. Regarding to Fig. 2-12, it illustrates more detail for the configuration of proposed entropy-based VAD.

6) *An adaptive thresholding (AT):*

Performing AT strategy to adapt the two thresholds to time-varying statistics of noise, the boundaries of voiced and unvoiced sounds are detected.

A voiced sound flag, f_v , is set by the following equation,

$$f_v = \begin{cases} 1, & \text{if } H(l) > T_s, \\ 0, & \text{otherwise.} \end{cases} \quad (2-24)$$

In addition, a unvoiced sound flag, f_{uv} , is set according to following equation

$$f_{uv} = \begin{cases} 0, & \text{if } T_{lf1} \leq R_{lf}(l) \leq T_{lf2} \\ 1, & \text{otherwise} \end{cases} \quad (2-25)$$

where T_{lf1} and T_{lf2} are the adaptive thresholds for unvoiced and voiced speech, respectively.

Fig. 2-12 shows the result of RLF measure tested in recorded speech sentence /start/.

It is can be seen that the existence of the unvoiced sound in phoneme /s/. Fig. 2-13

states the process of developing the two feature parameters more clearly.

7) *The VAD Decision:*

Using (2-24) and (2-25) the VAD result of the proposed method is now given by

$$VAD(l) = f_v(l) \cup f_{uv}(l) \quad (2-26)$$

The zero and non-zero of VAD result mean the voice-inactive frame and voice-active frame, respectively.

2.4 Evaluation

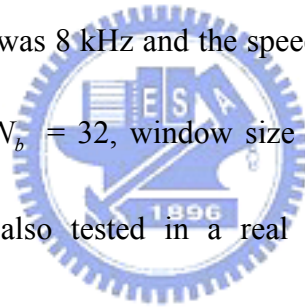
To evaluate the proposed VAD algorithm comprising the two measures of BSE and RLF, the recorded signal is tested in four kinds of noise including vehicle, multi-talker babble, factory, and white noises taken from the NOISEX-92 database [49] at various

SNRs. In addition, the probabilities of correct and false detection are used as objective measurement and defined as bellows:

--Probability of correctly detecting speech frames, P_{cs} : computed as the ratio of correct speech detections to the total number of hand-labeled speech frames.

--Probability of falsely detecting speech frames, P_{fs} : computed as the ratio of falsely classified speech frames to the total number of hand-labeled speech frames.

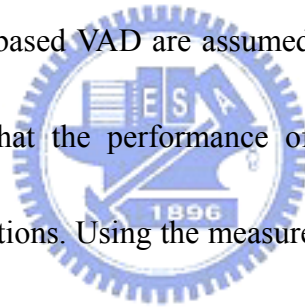
In the following experiment, the speech corpora are collected from MAT (Mandarin across Taiwan) database including a set of isolated utterances of the ten digits in Mandarin. The sampling rate was 8 kHz and the speech was stored as 16-bit integers. The decomposed band size $N_b = 32$, window size $N_w = 256$, and overlap size = $N_w/2$. The recordings are also tested in a real car environment with musical background noise.



2.4.1 Artificially Added Noise

The four noise signals were added to the recorded speech signals with different SNRs (the SNRs is herein defined as the ratio of the power of the entire recordings containing silence and voice parts to the power of additive noise) including -5 dB, 0dB, 10 dB, and 40 dB to generate noisy speech signals. Using these various types of noise and different SNRs, the P_{cs} and P_{fs} of proposed algorithm were compared

with those of the ATF-based algorithm. TABLE 2-I displays the comparisons between the presented entropy-based and ATF-based VADs. It is clearly found that the proposed entropy-based VAD algorithm is superior to the ATF-based algorithm, especially at low SNRs. At high SNRs the ATF-based algorithm performs as well as the RABS-BSE-based one; however, at low SNRs the ATF parameter related to pure energy-based feature is no longer valid. Although ABS associated with ATF-based algorithm claims that it can extract useful subbands, the claim cannot reflect the correct information within the selected subbands. In addition, the indexes of the harmful subbands from ATF-based VAD are assumed to be fixed with time, and this assumption is incorrect so that the performance of ATF-based VAD is seriously degraded under adverse conditions. Using the measure of spectrum entropy defined in subbands, the nature of banded lines on voice spectrogram can be exploited and further classified voice-active frame or voice-absent frame. For vehicle and white noises, whose frequencies are spread simply over the spectrum, the proposed algorithm is superior to the ATF-based one. Especially for white noise, it is proved that the VAD based on spectral entropy performs well in this case. Besides, the two types of noises have no any prominent banded lines on the corresponding voice spectrograms. The factory noise is also same as described above. The babble noise is pronounced by multi-talker, but the nature of banded lines of babble noise is weaker

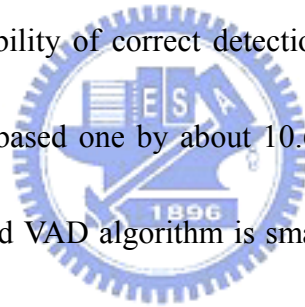


than that in speech signal. Our experimental result shows that the proposed entropy-based VAD algorithm still successfully outperforms the ATF-based algorithm under babble noise. For the average probability of correct detection of speech frames, P_{cs} , the proposed algorithm exceeds the ATF-based algorithm by around 8.73%. Similarly, for the average probability of false detection speech frames, P_{fs} , the proposed algorithm is less than the ATF-based algorithm by around 7.08%.

2.4.2 Recordings in a Car

To evaluate the effectiveness of the proposed feature parameter in a real environment, the musical noise is generated around the recordings in a real car. Fig. 2-14 shows the comparison between various types of feature parameter under musical background noise inside a car. Fig. 2-14(a) displays an utterance in Chinese, “Guo Li Chiao Tung Da Xue” made with musical background noise in a car. The corresponding spectrogram also shows in Fig. 2-14(b) and clearly displays that the banded nature appears only in voice-active spectrogram. Fig. 2-14(c) and 2-14(d) demonstrate that the short-time energy and ZCR both fail in a car environment. Fig. 2-14(e) shows that the ATF parameter outperforms the other two due to that the ATF parameter can extract useful frequency information by selecting proper subbands; however, it is still a purely energy-based parameter so that the ATF parameter fails in

a rapid increase of noise. Fig. 2-14(f) indicates that the BSE parameter is superior to the other parameters, especially in a rapid increase of noise since BSE parameter catches only the profiles of banded lines not their energy levels. In TABLE 2-II, the proposed entropy-based VAD is compared with the ATF-based VAD [7] testing in musical background noise within a car. The speech database used in the experiments contains ten isolated controlled-commands in Mandarin Chinese produced by 15 native speakers. The employed speech recognizer is based on a DTW-based recognizer. During the entire process, the car was moving and its radio was on. It is observed that the total probability of correct detection of the proposed algorithm is greater than that of the ATF-based one by about 10.6%, and the total probability of false detection of the proposed VAD algorithm is smaller than that of the ATF-based one by about 5.9%.



2.5 Discussion & Future Work

In Chapter 2, the objective mainly exploits the nature of the banded lines (or called formant traces) on voice-active spectrogram to regard as a robust feature. Via frequency band analysis, the inherent nature can be characterized by a measure of entropy defined subband domain. In fact, some subbands can be contaminated by noise to result in error of BSE value. Consequently, an on-line SSE strategy is used to

automatically extract the useful subbands. So, the BSE parameter with an SSE strategy can correctly detect the boundary of voice activity. In addition to the BSE parameter, the alternative parameter, RLF (ratio of low-band energy to full-band energy) is then presented to compensate the limitation of BSE for characterizing unvoiced speech. Compare the presented parameter with others, it certainly provides a reliable performance even in variable noise level. Experimental results show that the proposed VAD comprising the two BSE and RLF feature parameters and adaptive thresholding, which is used for a VAD decision and an indicator of executing subband extraction, is prior to other VAD, especially at low SNRs and at variable level of noise.



However, the subband extraction is either discarded or preserved. It is not good enough to help the result of VAD. Although some subbands are contaminated by noise, they still offer some useful information. In future work, each subband has respective weighting and the total weighting is 1. The weighting in harmful subband is lower than that in useful subband. Due to that the SSE strategy assume the larger subband power as harmful band, however this power is speech power (useful band) for a clean signal. So, this false subband discard may reduce the accurate rate of detection. In this case, the rate of correct detection can reduce to about 3%. Similarly, the rate of false detection can increase to about 1.7%. According this find, an improved subband

extraction for testing in a clean signal must be required.

In addition, the number of decomposed subbands is no longer fixed. The number of decomposed subband, in practice, varies with time. So, the rough solution is that we use a peak/valley detection on spectrum for each frame to locate the varying banded lines.



TABLE 2-I
COMPARISON BETWEEN THE PROPOSED ENTROPY-BASED VAD AND
ATF-BASED VAD [7] UNDER VARIOUS NOISE CONDITIONS

Noise Conditions		Proposed entropy-based VAD		ATF-based VAD [7]	
Type	SNR(dB)	Probability of correct detection, P_{cs} (%)	Probability of false detection, P_{fs} (%)	Probability of correct detection, P_{cs} (%)	Probability of false detection, P_{fs} (%)
Vehicle Noise	40	98.4	1.2	96.5	1.8
	10	94.2	3.9	86.3	9.8
	0	91.3	4.6	85.1	14.3
	-5	89.3	6.5	81.2	17.2
Babble Noise	40	96.1	2.6	94.8	3.8
	10	89.2	5.8	80.4	10.8
	0	84.8	7.4	78.9	15.3
	-5	79.6	10.4	69.4	21.2
Factory Noise	40	97.9	2.1	95.5	3.2
	10	91.7	4.6	83.4	10.5
	0	86.3	7.8	79.3	15.1
	-5	84.1	9.1	68.1	18.4
White Noise	40	99.8	0.9	95.3	3.4
	10	96.6	1.9	82.6	9.5
	0	93.1	2.2	76.5	14.5
	-5	91.9	2.9	71.3	18.4
Average		91.52	4.62	82.79	11.7

TABLE 2-II
COMPARISON OF THE PROPOSED ENTROPY-BASED ALGORITHM AND ATF-BASED VAD
TESTING IN CAR WITH MUSICAL BACKGROUND NOISE

VAD types	Total probability of correct detection, P_{cs} (%)	Total probability of false detection, P_{fs} (%)
Proposed entropy-based	89.2	3.5
ATF-based VAD [7]	78.6	9.4

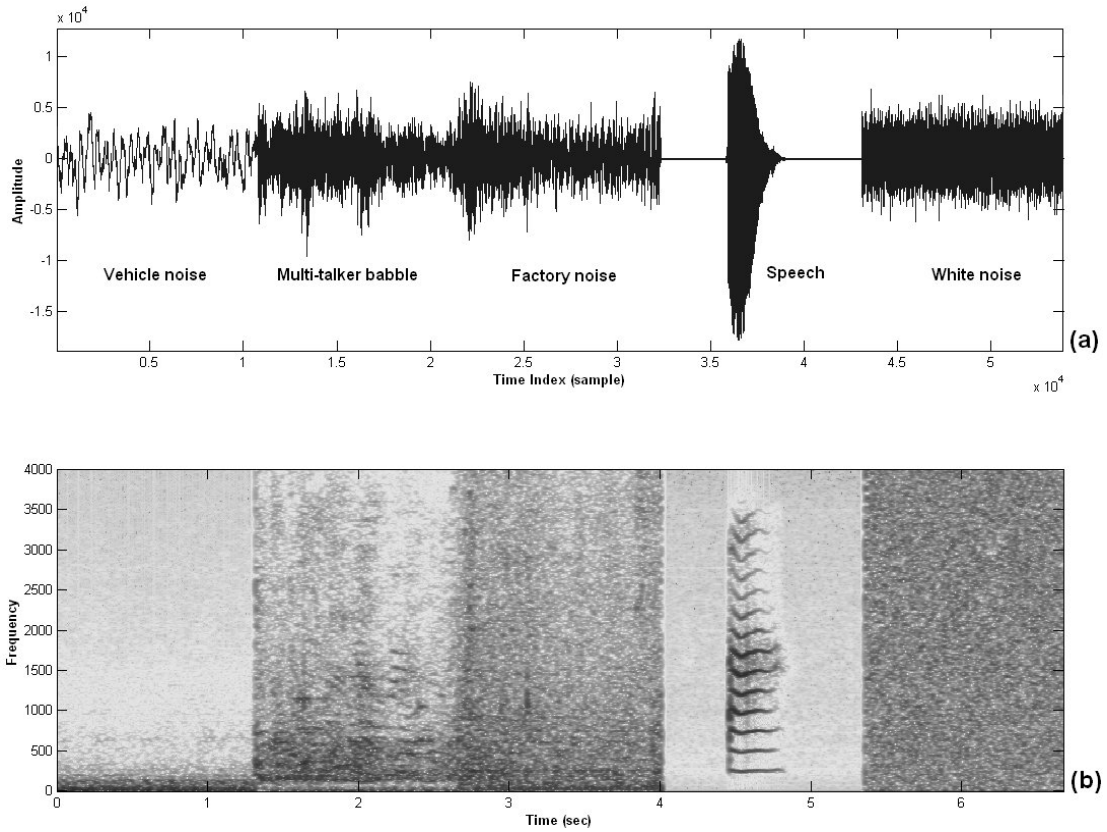


Fig. 2-1 Inherent characteristic of banded lines (formant traces) only appears on voice-active spectrogram: (a) Mixed signal waveform is composed of vehicle noise, multi-talker babble noise, factory noise, and speech signal and white noises in turn. (b) Spectrogram of the corresponding mixed signal.

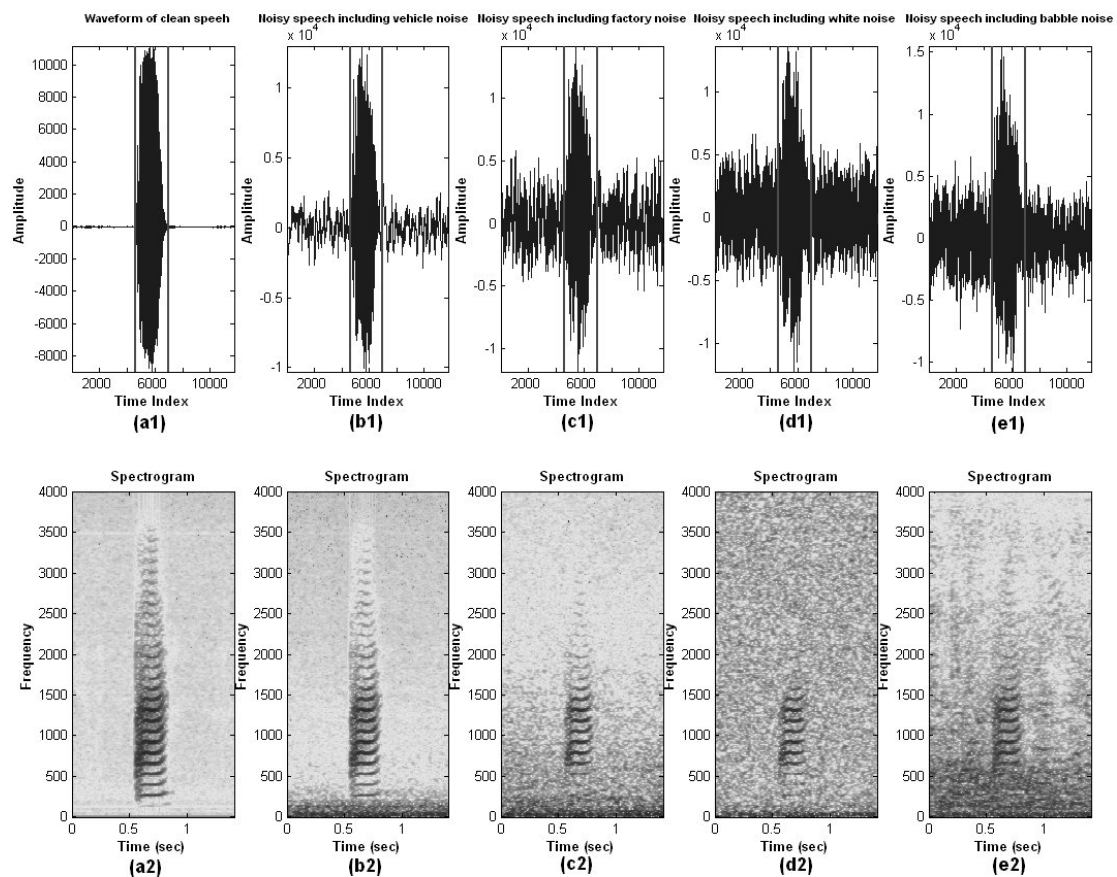


Fig. 2-2 Illustration of the banded lines existing in various types of noises.

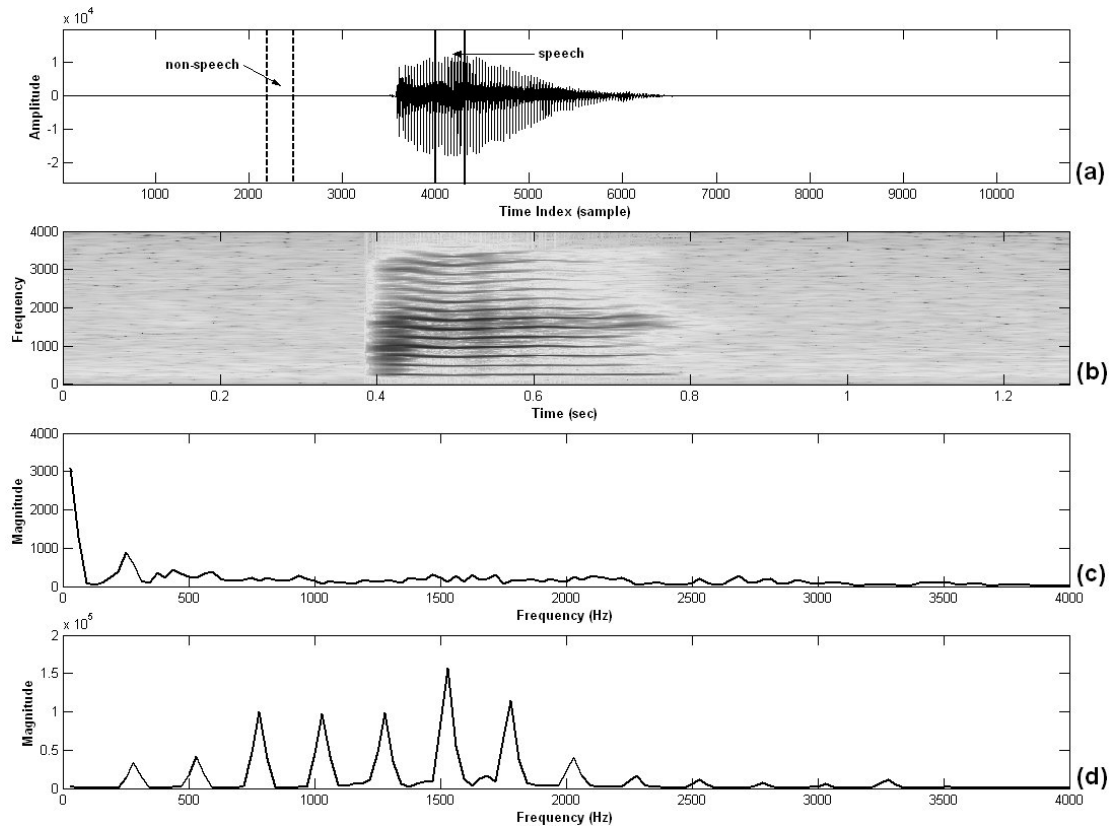


Fig. 2-3 Nature of banded lines on voice-active spectrogram: (a) A signal waveform of Mandarin digit “eight”. (b) The continuous, banded lines only appearing on the corresponding voice-active spectrogram. (c) Spectrum magnitude of voice activity. (d) Spectrum magnitude of voice-absent frame.

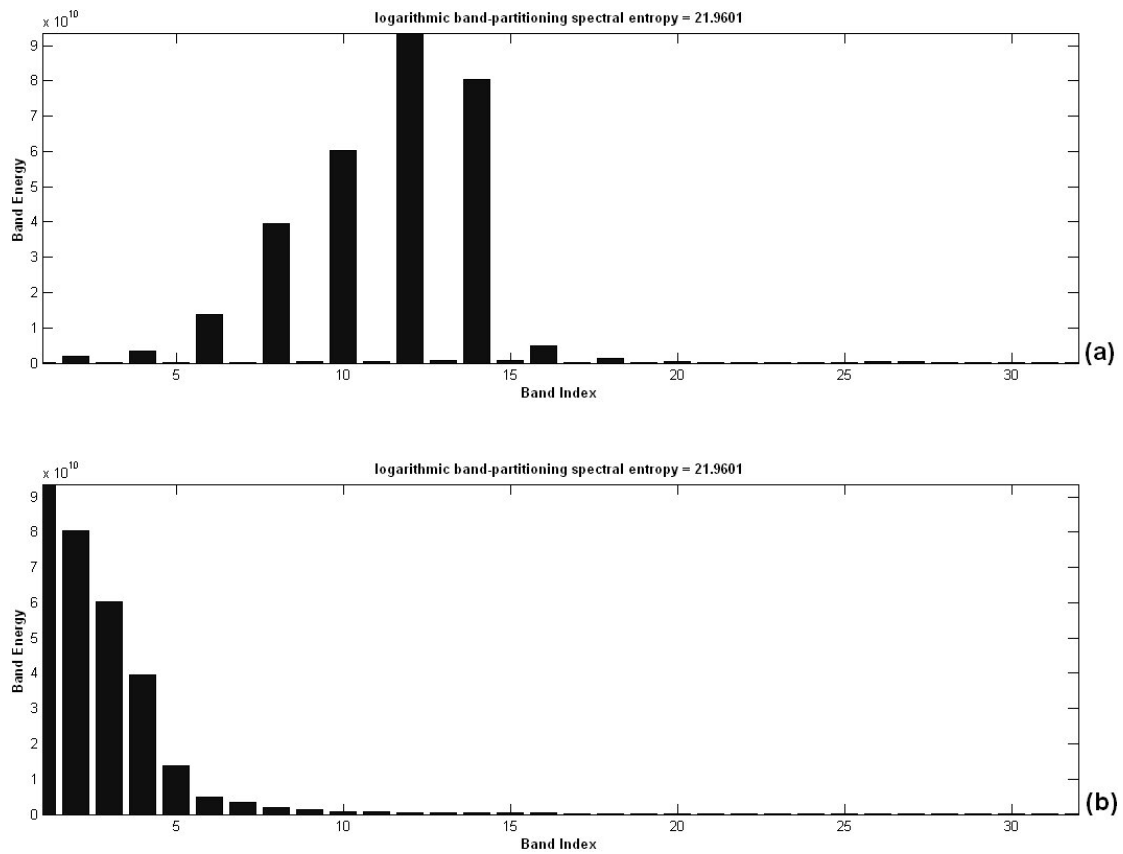


Fig. 2-4 Power distributions of all 32 uniform subbands with the same entropy (logarithmic BSE=21.9601): (a) During voice-active frame. (b) During voice-absent frame.

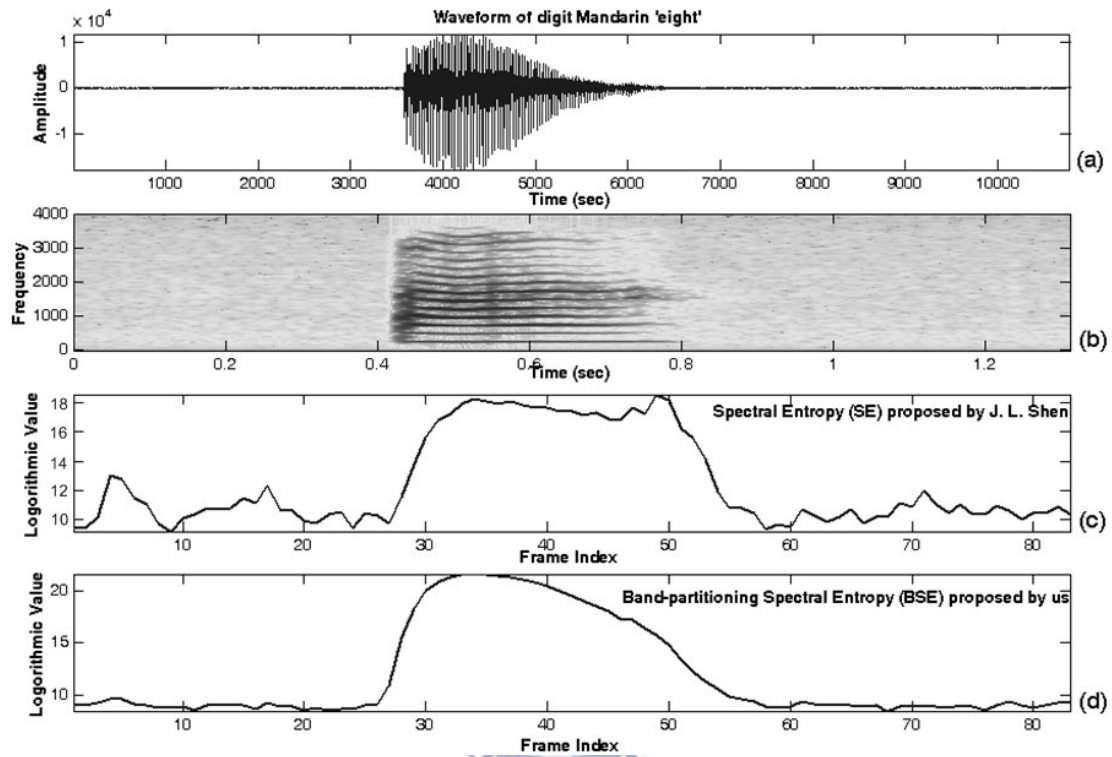


Fig. 2-5 Illustration of characterizing speech signals by using entropy-based feature parameter: (a) Waveform of a Mandarin digit “eight”. (b) The corresponding spectrogram. (c) Contour of SE proposed by J. L. Shen *et al.* [30]. (d) Contour of the proposed BSE.

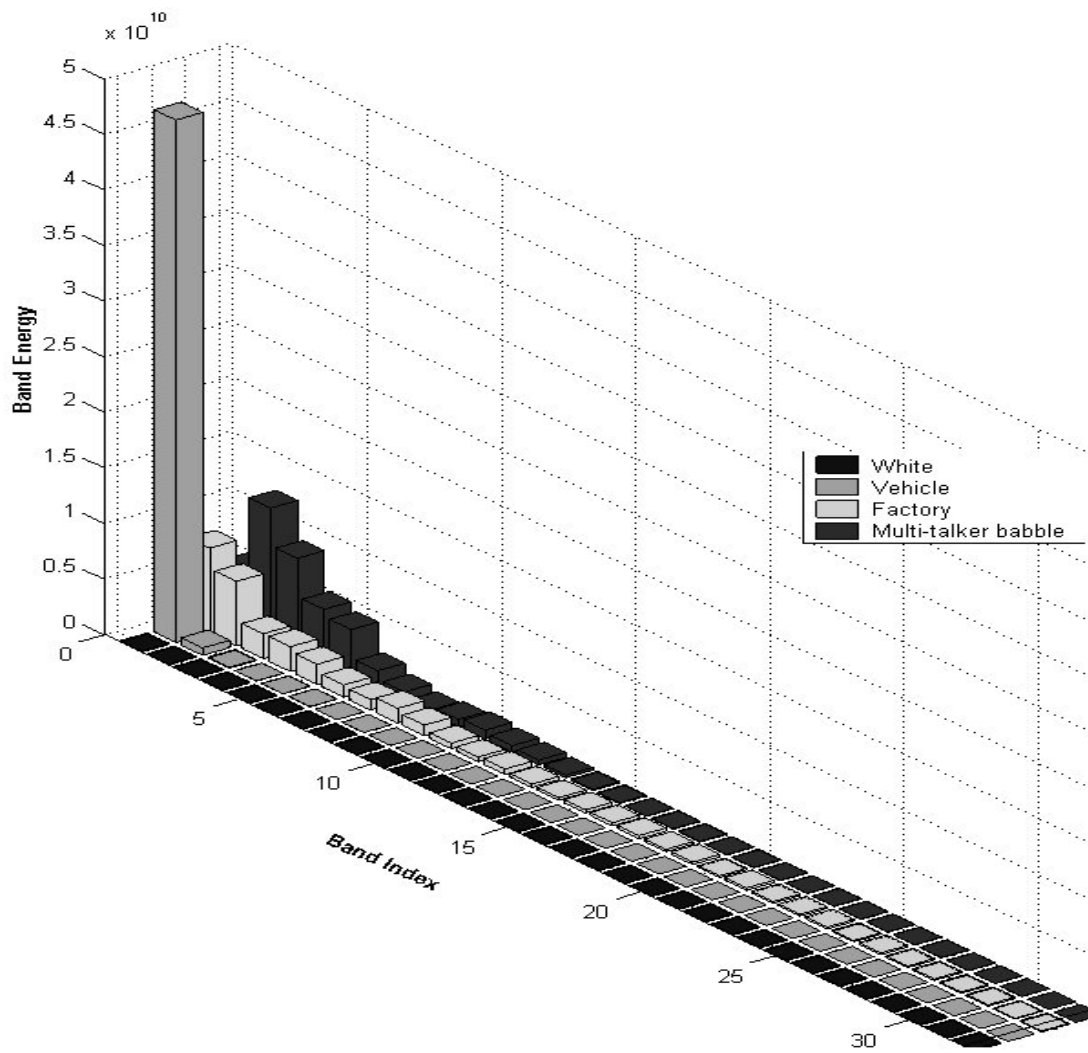


Fig. 2-6 Different types of noises focusing on different frequency subbands.

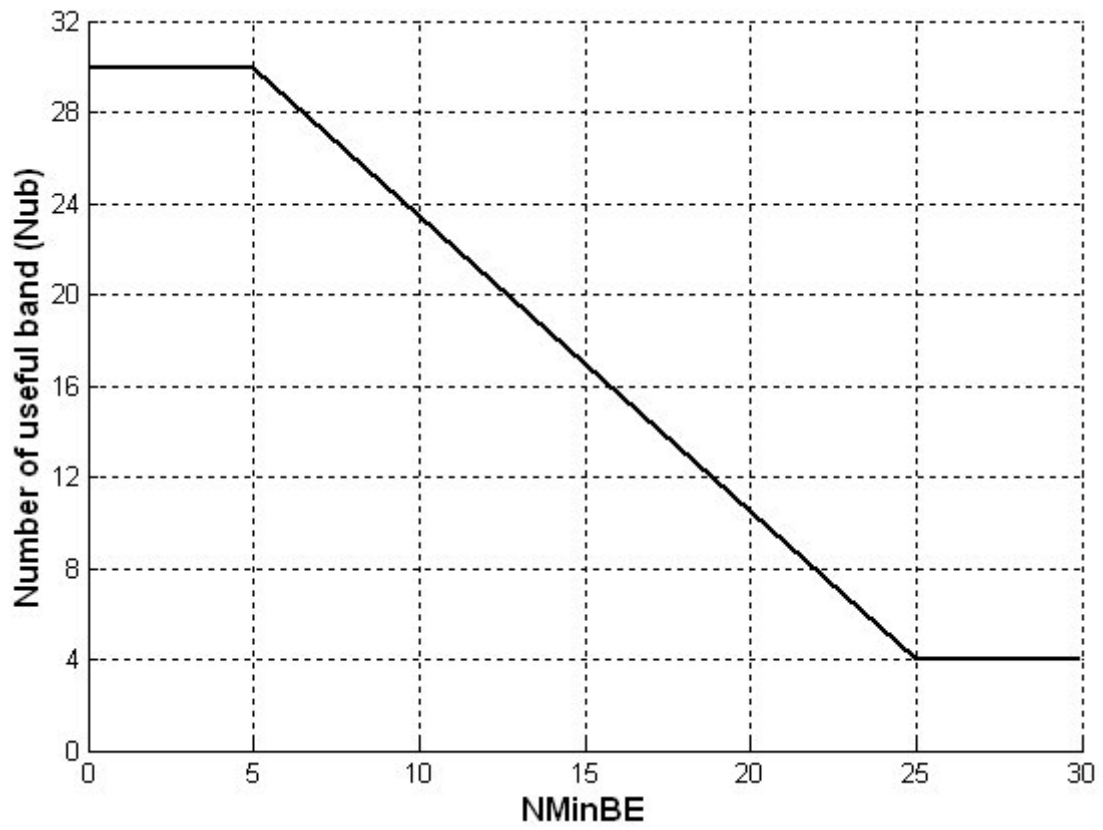


Fig. 2-7 Relation between the number of useful subbands and NMinBE parameter.

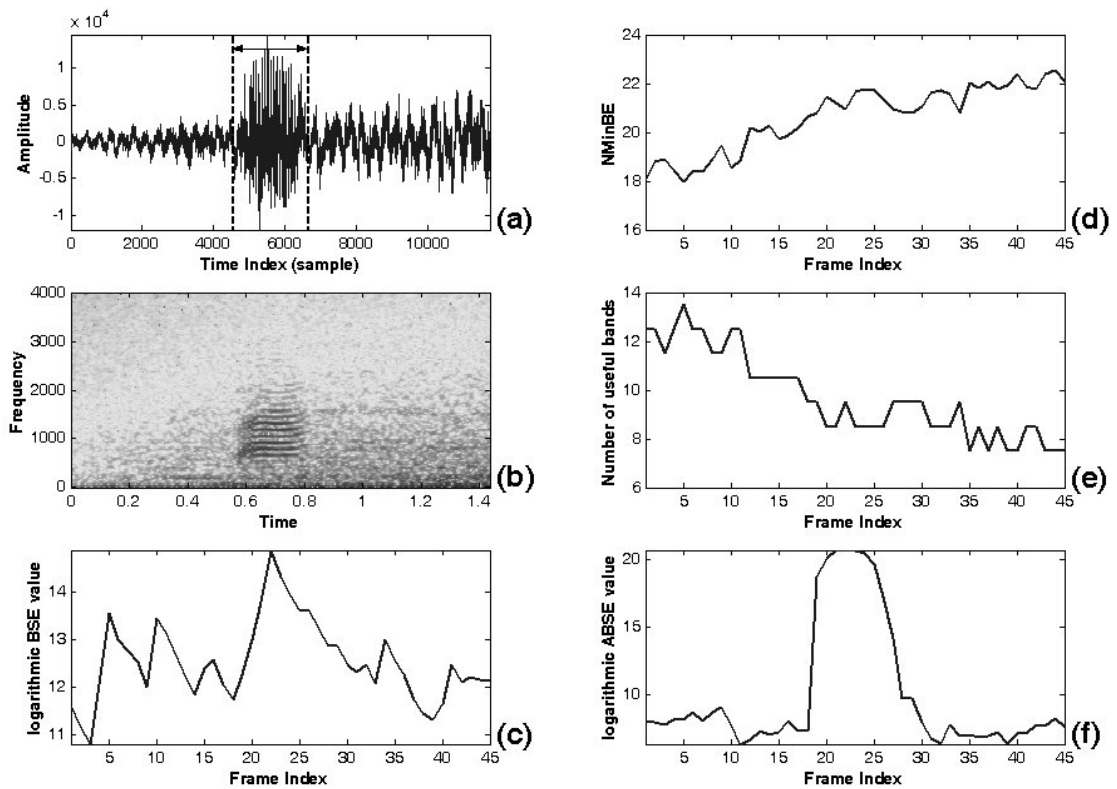


Fig. 2-8 Illustration of the efficiency of NMinBE parameter for applying in BSE parameter: (a) Waveform of the Mandarin digit “eight” at SNR -5 dB with increasing-level of factory noise. (b) The corresponding spectrogram. (c) The contour of BSE measure. (d) The NMinBE parameter. (e) The number of useful subbands varying with time. (f) The contour of BSE parameter obtained by manual selecting useful bands according to Fig. 2-7.

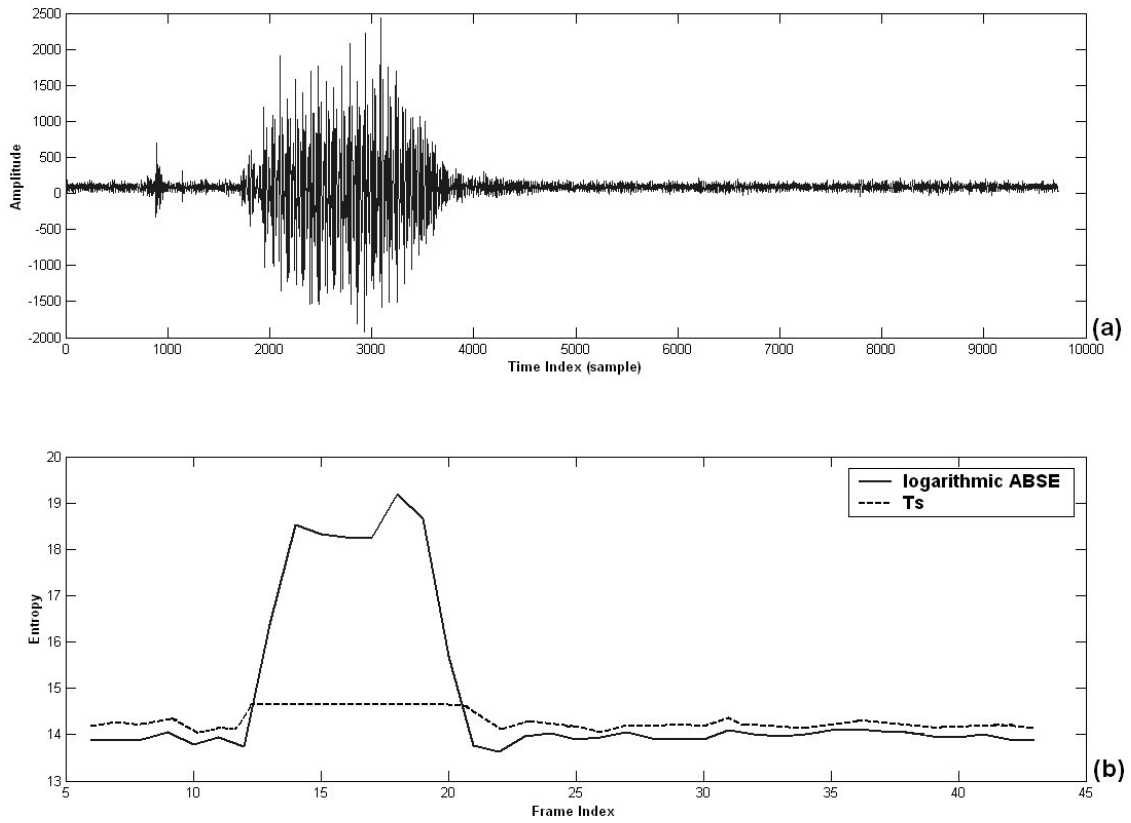


Fig. 2-9 An adaptive threshold method for VAD decision: (a) Waveform of an utterance of the digit “one”. (b) Detection of speech segments together with the logarithmic BSE value and speech threshold T_s .

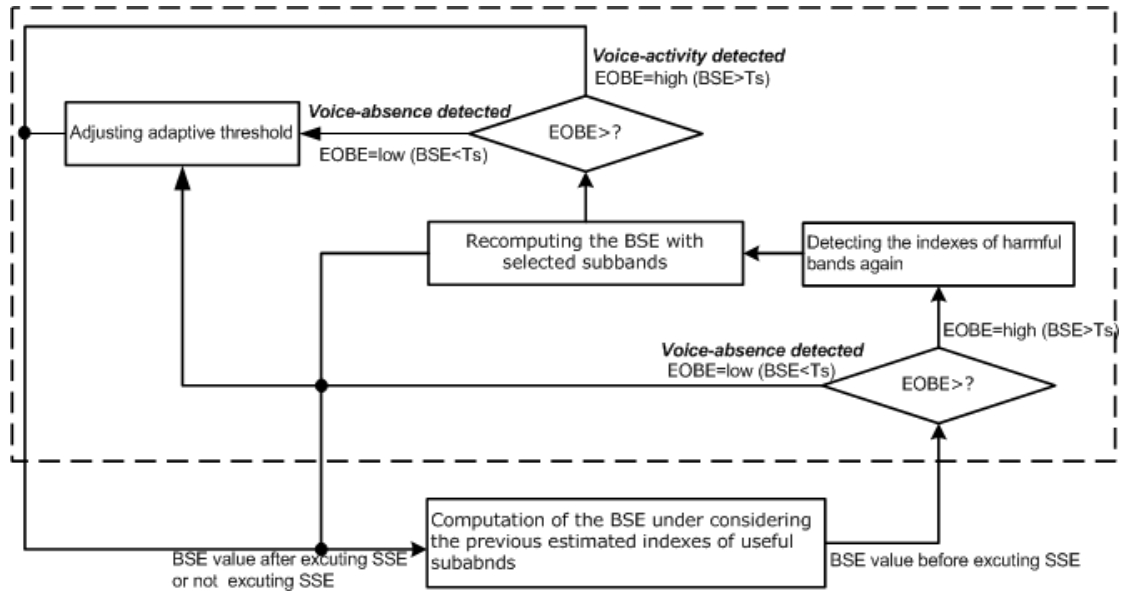


Fig. 2-10 Flowchart of SSE strategy for automatically extracting the useful subbands.

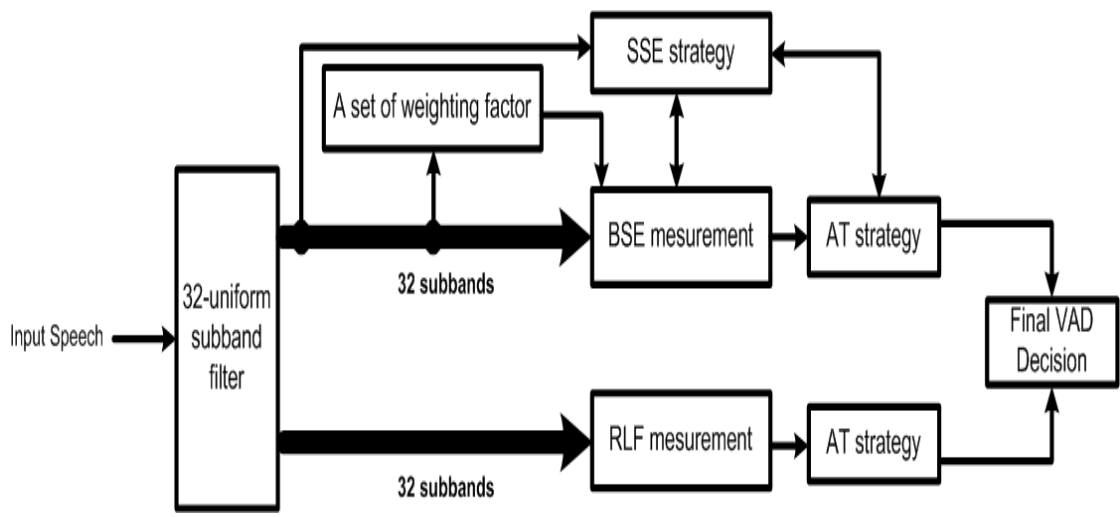


Fig. 2-11 Block diagram of the proposed entropy-based VAD algorithm.

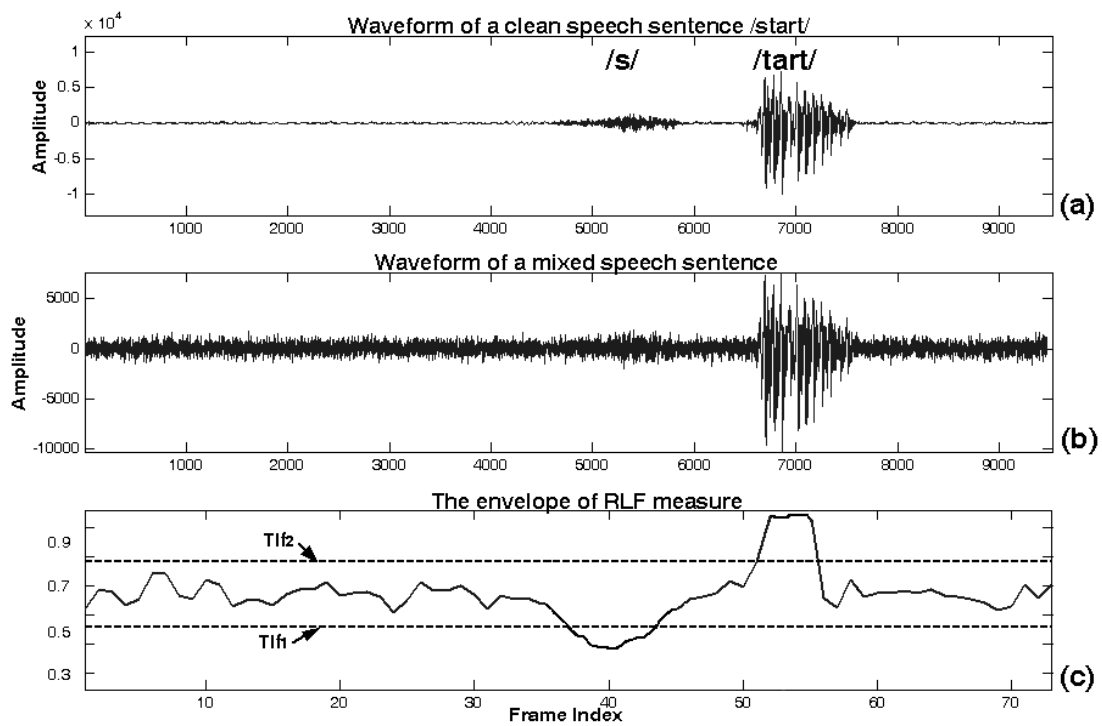


Fig. 2-12 Result of RLF measure tested in recorded speech sentence /start/. (a) Waveform of a noisy speech sentence. (b) The envelope of RLF measure.

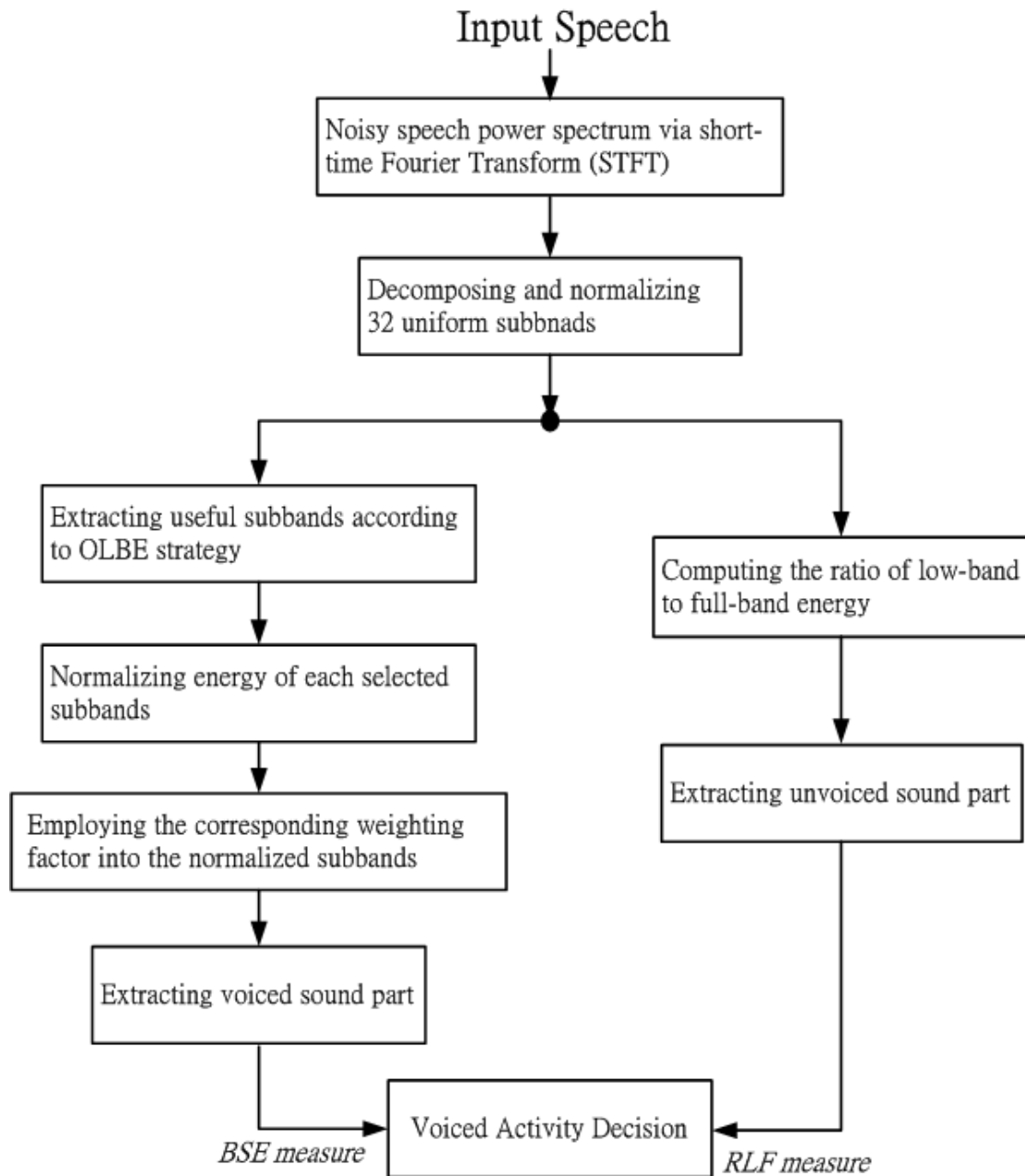


Fig. 2-13 Measurement of the two BSE and RLF parameters.

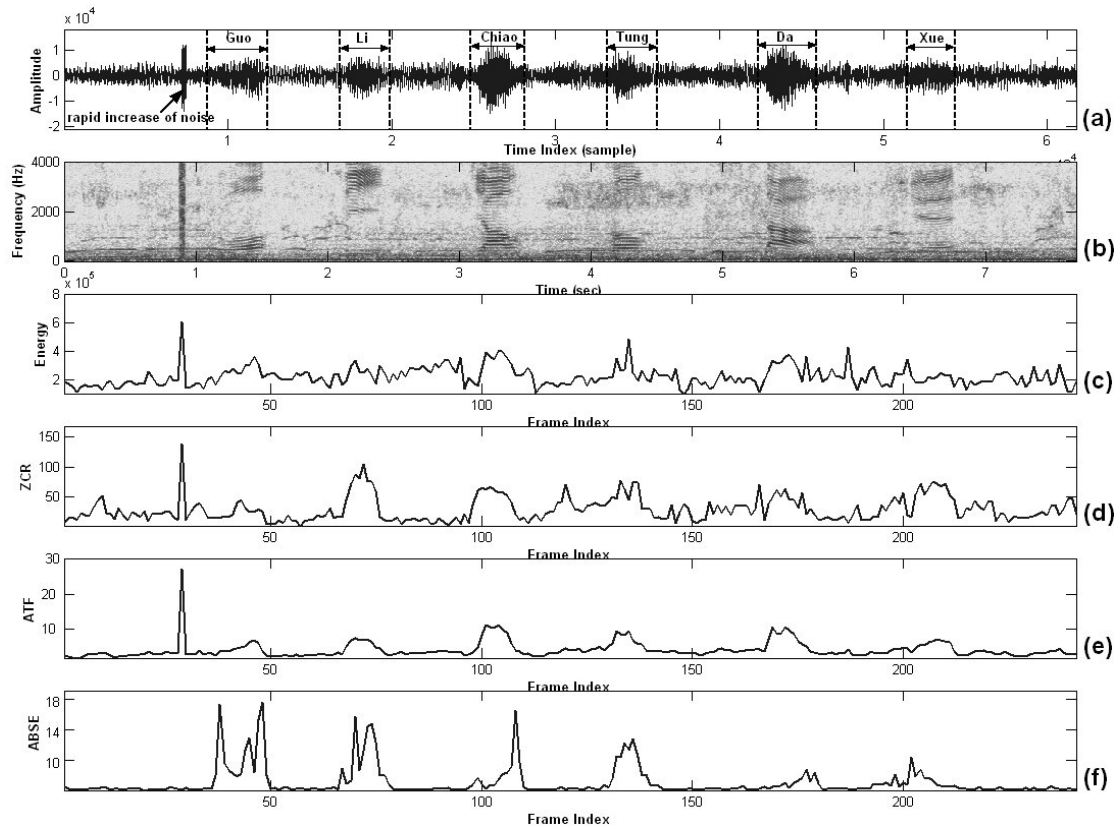


Fig. 2-14 Comparison between different feature parameters for VAD algorithm testing an utterance with musical background noise inside a car: (a) Waveform of an utterance in Chinese: “Guo Li Chiao Tung Da Xue (National Chiao Tung University)”. (b) The corresponding spectrogram (c) Contour of spectral energy. (d) Contour of ZCR. (e) Contour of ATF (f) Contour of BSE.

CHAPTER 3

A SINGLE CHANNEL NOISE SPECTRUM ESTIMATION WITH RAPID ADAPTATION IN VARIABLE-LEVEL OF NOISY ENVIRONMENTS



In this Chapter 3, a single channel noise estimation algorithm using only the power spectrum of noisy speech is presented. The proposed method can track the noise spectrum quickly, even when the noise levels suddenly increase. An explicit use of speech/silence detection is needed for estimating noise spectrum. So, the entropy-based VAD mentioned-above is used to continuously classify each frame of speech into the voice active/absent frames, and the noise spectrum estimate is updated using constant smoothing factor for voice absent frames and a time-frequency dependent smoothing factor for voice active frames. Time-frequency dependent smoothing factor is chosen as a Sigmoid function that changes with the voice-active

probabilities in frequency bins. And, voice-active probability is determined by computing the ratio of the noisy speech power spectrum to its local minimum. To speed up the minimum tracking, a fast method is presented for tracking the minimum of the noisy speech power spectrum. In addition, to allow detection with entropy-based VAD under colored noise conditions, we herein propose to subtract the current spectrum from the estimated noise spectrum of the previous frame.

3.1 Introduction

Most of the existing single channel noise estimations are slow in adapting to increasing levels of noise. This results in a perceptually annoying residual noise and speech distortion in aspect of speech enhancement. In general, noise estimation is usually done by explicit detection of speech detection. However, this can be very difficult in the case of varying background noise so that the background noise is assumed to be related stationary between speech pause.

Martin [36] proposed a method of noise spectrum estimation that is based on minimum statistics (MS). The noise spectrum is estimated by tracking the minimum of the noisy speech power spectrum over a particular window. Furthermore, Cohen *et al.* [37] introduced a minima-controlled recursive averaging (MCRA) method extended from the MS to estimate the noise power spectrum using a smoothing

parameter that is defined as the voice-active probability in the frequency bins. The VAD is carried out by comparing the probability with a specific threshold and then determines whether to update the estimate of the noise. However, its VAD decision clearly depends on energy level and performs poorly when noise level is higher than speech level. In addition, the noise estimation does not adapt quickly to a rapid change in noise level. Recently, Lin *et al.* [38] developed an adaptive noise estimation to easily implement. Its smoothing parameter can be chosen as a Sigmoid function changing with posteriori SNR. However, the stability of posteriori SNR is sensitive to a variable noise-level. So far, these kinds of algorithms contain no explicit VAD and their performances depend on the energy level. Accordingly, the noise estimation does not adapt quickly in situations involving rapid change of noise level. To overcome this problem, a noise spectrum estimation with rapid adaptation in variable-level of noisy environments is relatively required.

Enclosed herein we propose a method for tracking the noise spectrum quickly, even when the noise levels suddenly increase. An explicit use of speech/silence detection is needed for estimating noise spectrum. So, the entropy-based VAD mentioned-above is used to continuously classify each frame of speech into the voice active/absent frames, and the noise spectrum estimate is updated using constant smoothing factor for voice absent frames and a time-frequency dependent smoothing factor for voice active

frames. The time-frequency dependent smoothing factor is chosen as a Sigmoid function that changes with the voice-active probabilities in frequency bins. And, the voice-active probability is determined by computing the ratio of the noisy speech power spectrum to its local minimum. To speed up the minimum tracking, an efficient method extended from [52] for tracking the minimum of the noisy speech power spectrum is presented. Besides, to allow the decision of entropy-based VAD under colored noise conditions, we suggest that the current spectrum is subtracted from the estimated noise spectrum of the previous frame. After the subtraction, the resulting spectrum is similar to the white noise in voice-absence. In order to make sure of the BSE and RLF values in the next frame well, a subtractive-type method presented by Berouti *et al.* [39] is used herein to decrease significantly the annoying “musical noise” that is introduced by subtracting the estimated noise spectrum from the noisy speech spectrum.

This chapter is organized as follows. Section 3.2 details the configuration of the proposed noise estimation algorithm for quickly adapting variable noise level. In addition, the entropy-based VAD above mentioned is modified into the noise estimation algorithm. Section 3.3 evaluates the proposed noise estimation algorithm in variable noise level as comparing with others. Finally, Section 3.4 would summarize the conclusions.

3.2 Proposed Noise Estimation Algorithm

We use two different sceneries to update the noise estimate. The first strategy updates the noise power spectrum during voice-absent frames, and the second strategy updates the noise power spectrum during voice-active frames. In general, the noisy speech can be denoted in time domain as

$$y(n) = x(n) + u(n), \quad (3-1)$$

where $x(n)$ is the clean speech and $u(n)$ is additive noise.

First, the smoothing power spectrum of noisy speech, $P(\omega, l)$, can be given using the recursive averaging formula as follows:

$$P(\omega, l) = \eta \cdot P(\omega, l) + (1 - \eta) \cdot |Y(\omega, l)|^2, \quad (3-2)$$

where $|Y(\omega, l)|^2$ is the short-time power spectrum of noisy speech and η is a smoothing constant determined experimentally.

Fig. 3.1 displays the flow diagram of the proposed algorithm for updating noise estimate in detail.

3.2.1. A modified entropy-based VAD

A decision of voice-absent/active derived from entropy-based VAD in Chapter 2 is regarded as an indicator of updating noise spectrum illustrated in Fig. 3.2 modified

from Fig. 2-13. To allow detection with entropy-based VAD under colored noise conditions, we propose to subtract the current spectrum from the estimated noise spectrum of the previous frame. In order to make sure of BSE measured well in next frame, a subtractive-type algorithm presented by Berouti *et al.* [39] is used here to decrease significantly the annoying “musical noise” that is introduced by subtracting the estimated noise spectrum from the noisy speech spectrum.

3.2.2. Update of noise spectrum during voice absent frames

The one scenery is that the noise estimate is then updated with a constant smoothing factor if the frame is classified as voice-absent frame. This rule can be stated as follows

If $I(l) = 1$

$$\tilde{U}(\omega, l) = \alpha_c \cdot \tilde{U}(\omega, l-1) + (1 - \alpha_c) \cdot |Y(\omega, l)|^2 \quad (3-3)$$

end

where $I(l)$ denotes an indicator of updating noise spectrum. α_c is a constant smoothing factor. $\tilde{U}(\omega, l)$ is the estimated noise power.

3.2.3. Update of noise spectrum during voice active frames

The alternative scenery for updating noise spectrum in voice-active frames is depended on the voice-active probability in each frequency bin for each frame. Tracking the minimum of the noisy speech spectrum, the ratio of noisy speech power

to its local minimum is regarded as a voice-active probability in frequency bins. Then, the probability is used to determine a time-frequency dependent smoothing factor $\alpha_f(\omega, l)$ according to a Sigmoid function as shown in next section. Eventually, the estimated noise power is updated during voice-active frames.

$$\tilde{U}(\omega, l) = \alpha_f(\omega, l) \cdot \tilde{U}(\omega, l-1) + (1 - \alpha_f(\omega, l)) \cdot |Y(\omega, l)|^2. \quad (3-4)$$

3.2.3.1. Tracking the local minimum

To maintain the ability of quickly adapting to increasing noise level, we avoid the constraint of any window length to track the local minimum. The minimum of the noisy speech by continuously averaging the past spectral values [52] is achieved shown as below.



$$\begin{aligned} &\text{If } P_{\min}(\omega, l-1) < P(\omega, l), \\ &\text{then } P_{\min}(\omega, l) = \gamma \cdot P_{\min}(\omega, l-1) + \frac{1-\gamma}{1-\beta} \cdot (P(\omega, l) - \beta \cdot P(\omega, l-1)), \\ &\text{else } P_{\min}(\omega, l) = P(\omega, l), \end{aligned} \quad (3-5)$$

where $P_{\min}(\omega, l)$ represents the local minimum of the noisy speech power spectrum. γ and β are constants determined experimentally. The look-ahead factor β controls the adaptation time of the local minimum.

3.2.3.2. The calculation of time-frequency dependent smoothing factor

After the tracking of local minimum, the time-frequency smoothing factor

$\alpha_{yf}(\omega, l)$ is then chosen as a Sigmoid function changing with voice-active probability in frequency bins. First, the voice-active probability in each frequency bin is determined using the ratio of noisy speech power spectrum and its local minimum [37] defined as

$$P_r(\omega, l) = \frac{P(\omega, l)}{P_{\min}(\omega, l)}, \quad (3-6)$$

where $P_r(\omega, l)$ is the ratio of noisy speech power spectrum and its local minimum.

Referring to Fig. 3.3, it is clearly displayed that the time-frequency dependent smoothing factor changes with the voice-active probability on each frequency bin.

The time-frequency dependent smoothing factor is given by

$$\alpha_{yf}(\omega, l) = \frac{1}{1 + e^{-k(P_r(\omega, l) - T(\omega))}}, \quad (3-7)$$

where $T(\omega)$ is the frequency dependent threshold whose optimal value is determined experimentally. k is a curve ratio generally defined in experiment.

3.3 Experimental Results

To evaluate the availability of the proposed noise approach, we test the tracking capability of the noise estimator and measure the segmental relative estimation error given types and levels of noise. Four additive noise types are discussed in Chapter and are taken from the Noisex92 database. They are White noise, Vehicle noise, Factory

noise and Babble noise in turn. The speech database is sampled at 8000 Hz and linearly quantized at 16 bits per sample.

Fig. 3-4 shows a comparison between MCRA [37] and our proposed noise estimation for the case where there was a sudden increase in noise power level. Fig. 3-4(a) depicts a noisy speech sentence /May I Help You?/ spoken by a native man. The additive Factory noise suddenly increases in 23000th sample (or 2.875 sec). The corresponding spectrogram shows in Fig. 3-4(b) is found that the inherent nature of banded lines is robust against the change of noise-level. Fig. 3-4(c) plots the tracking capability between MCRA [37] and our proposed noise estimation. From the figure it can be seen that the proposed noise estimation for estimating noise (for single frequency bin $\omega = 60$) can track fast changes in the noise-level as comparing MCRA. In addition, the noise estimated by the proposed method (thick-line) is closer to the ideal spectrum (dotted-line) than that estimated by the MCRA (broken-line). An objective measure of segmental relative estimation error (SegErr) is defined by

$$SegErr = \frac{1}{N_l} \sum_{l=0}^{N_l-1} \frac{\sum_{\omega} [\tilde{U}(\omega, l) - U(\omega, l)]^2}{\sum_{\omega} U^2(\omega, l)}, \quad (3-8)$$

where N_l denotes the number of frames in recorded signal. \tilde{U} and U are estimated and real noise powers.

Table 3-I shows the outcomes of the SegErr measured by the proposed estimation method for four noise types with the SNRs range [-5 to 40dB]. The proposed

RMCRA approach is superior to the MCRA method over all SNRs. The S/N ratio in each subband is depicted in Fig. 3-5. Comparison between clean speech and mixed speech by adding a factory noise, it can be seen that the S/N ratios in lower frequency band decrease greatly than that in other bands because factory noise mainly focuses in lower frequency band. The distinct difference between ideal S/N and estimated S/N is also usually occurred in lower subbands required more adaptive time.

3.4 Discussion

Compare to the MCRA, the proposed noise estimation is superior because this noise estimation algorithm contains an explicit VAD scheme discussed in Chapter 2. This chapter makes use of the employed VAD's benefit of being robust against variable-level in noise to update noise spectrum. In addition, two sceneries are respectively used to accurately update noise power when being voice-active or voice-absent frames. Unlike MCRA, the adaptation of this time-frequency dependent smoothing parameter is not depended on a specific time window so that it can adapt quickly. Experimental results illustrate that the proposed noise estimation adapts to noise level faster than the MCRA.

Table 3-1 SEGERR FOR FOUR NOISE TYPES AND LEVELS

Input SegSNR [dB]	White Noise		Vehicle Noise		Factory Noise		Babble Noise	
	Propose method	MCRA	Propose method	MCRA	Propose method	MCRA	Propose method	MCRA
<i>40</i>	0.062	0.098	0.085	0.138	0.089	0.139	0.102	0.158
<i>10</i>	0.061	0.084	0.078	0.129	0.086	0.132	0.098	0.146
<i>0</i>	0.058	0.081	0.069	0.115	0.076	0.118	0.087	0.127
<i>-5</i>	0.055	0.078	0.065	0.107	0.072	0.109	0.085	0.114

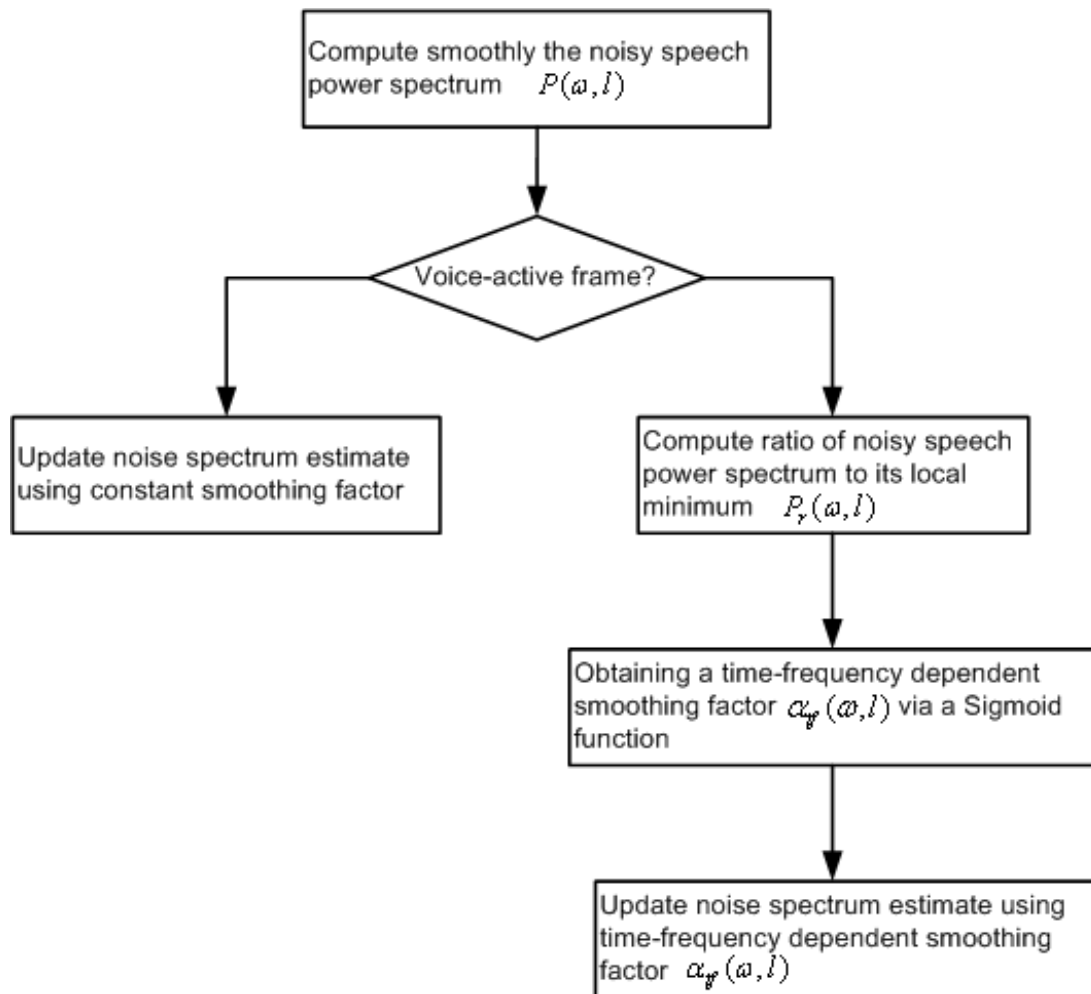


Fig. 3.1 Flow diagram of the proposed algorithm for updating noise estimate.

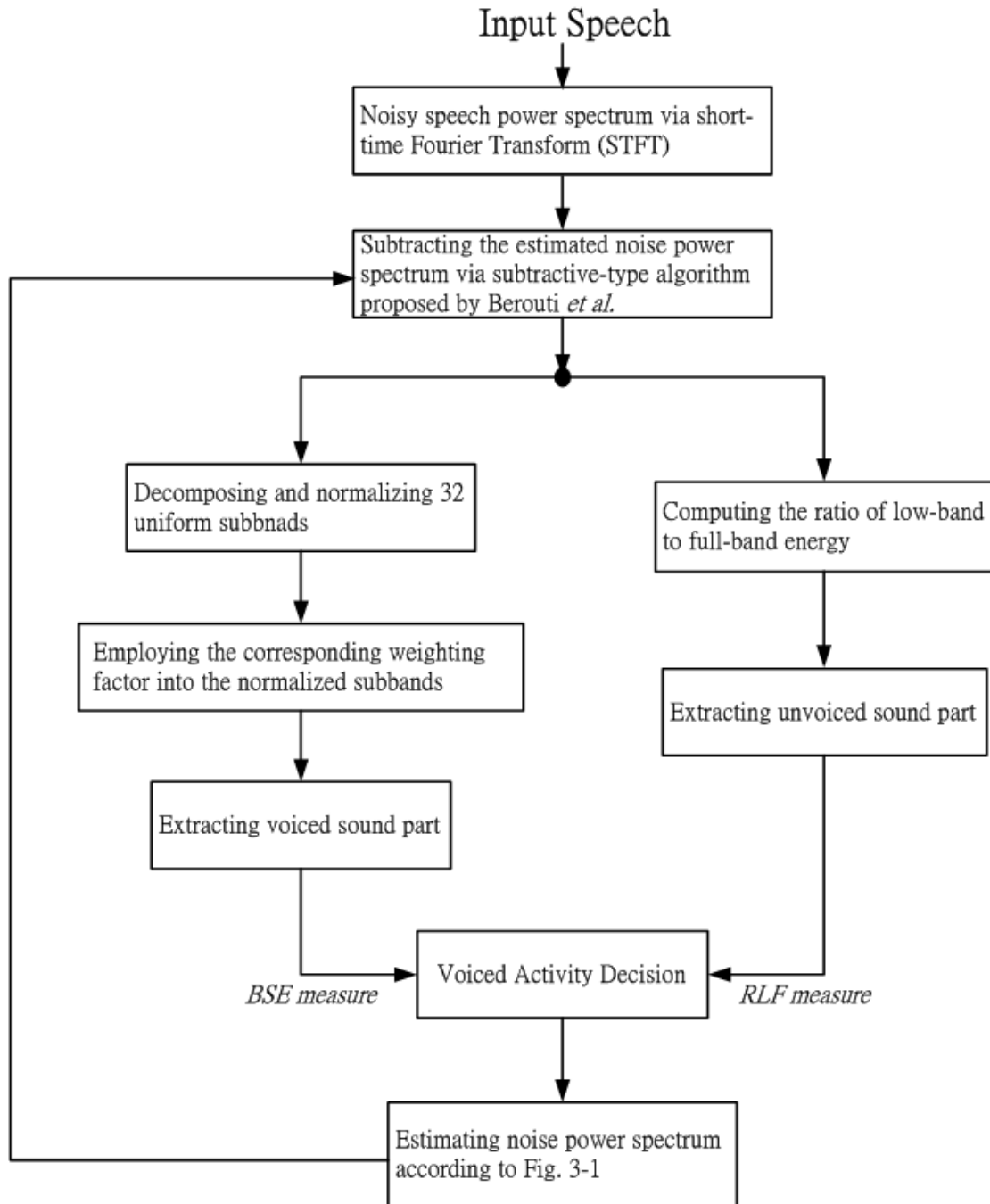


Fig. 3-2 Flow diagram of a modified entropy-based VAD.

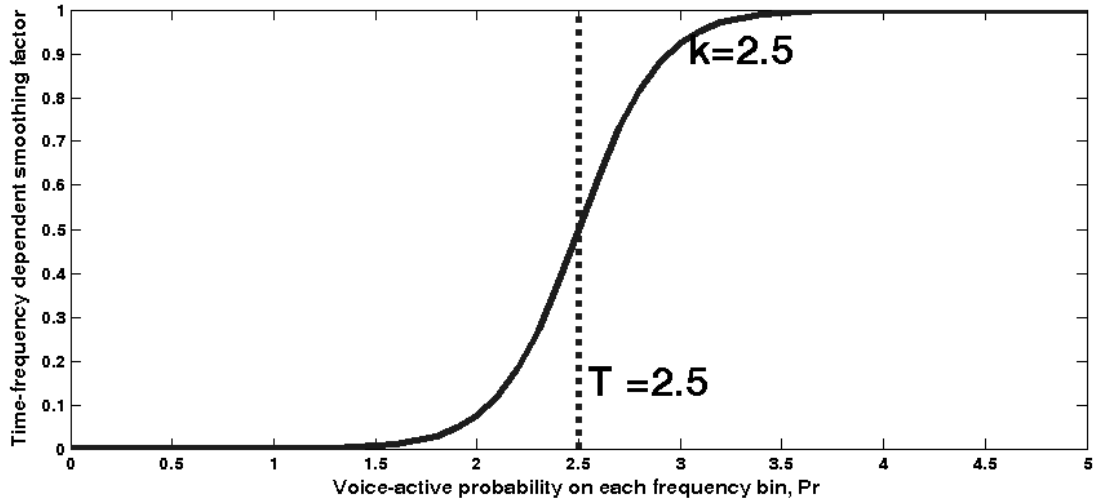


Fig. 3.3 Sigmoid function for computing a time-frequency dependent smoothing factor.

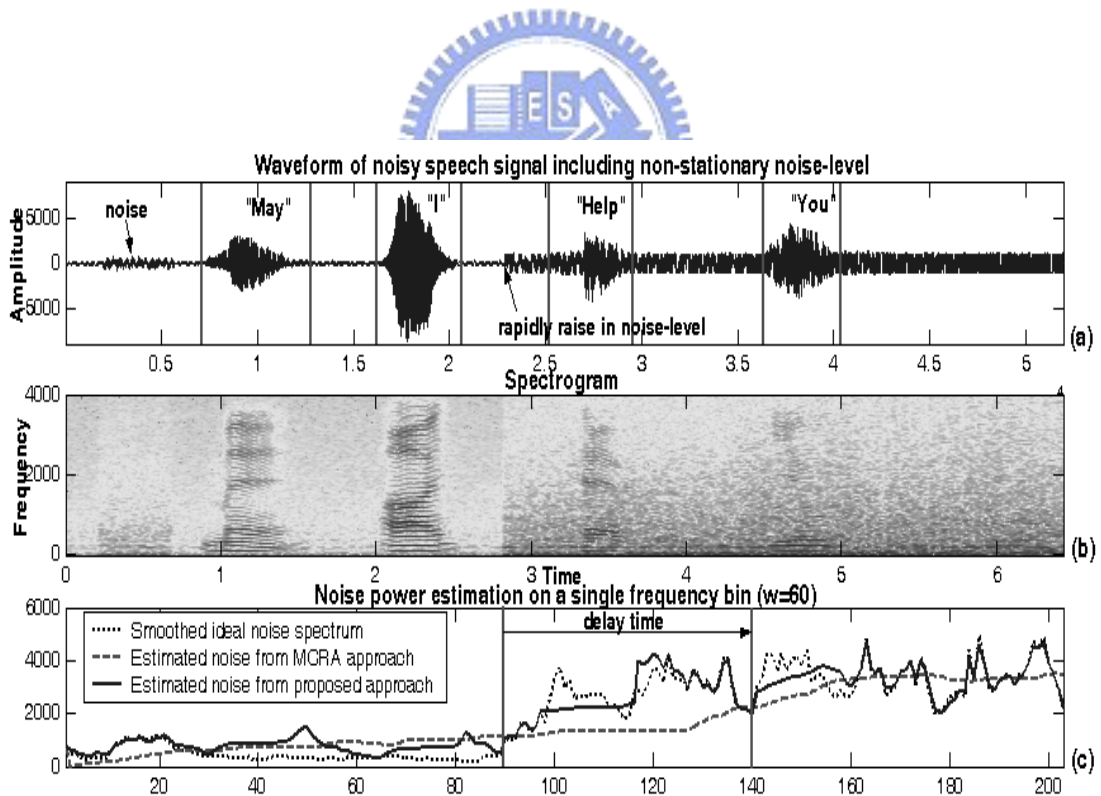


Fig. 3-4 Tracking capability of the noise estimator tested in noisy speech with a sudden increase in the level of factory noise.

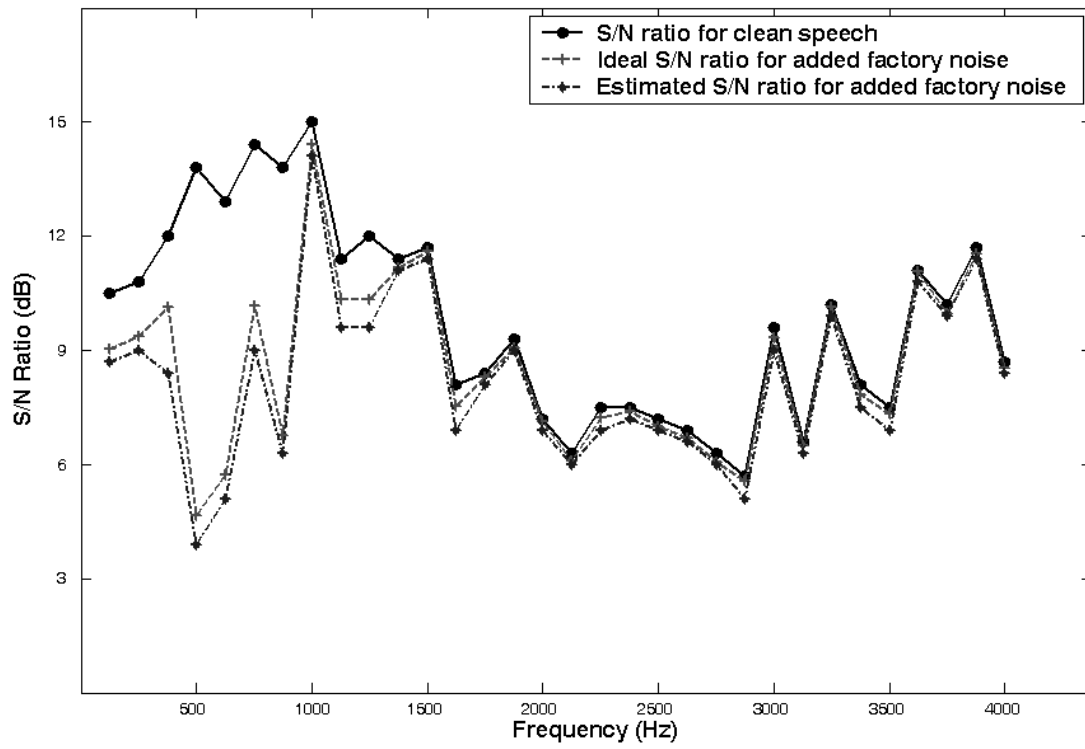


Fig. 3-5 The S/N ratio in each frequency subband (32 equally frequency bands):

(o)= clean speech, (+)= added factory noise, (*)= estimated noise.

CHAPTER 4

VOICE ACTIVITY DETECTION BASED ON WAVELET ANALYSIS USING A MEASURE OF AUTO-CORRELATION FUNCTION AND TEAGER ENERGY OPERATOR



According to the multi-resolution analysis (MRA) property of discrete wavelet transform (DWT), the dissimilarity among of voiced, unvoiced speech sounds and background noises from wavelet coefficients can be discriminated sufficiently [53]. In Chapter 4, the well-known auto-correlation function is defined in decomposed subband and regarded as *subband auto-correlation function* (SACF) here. A nonlinear Teager energy operator (TEO) is then utilized into each subband signal to decrease the influence of noise on subbands significantly before computing SACF. Besides, the other advantage of TEO is suitable for the result of SACF. To obtain the amount of periodicity, a Mean-Delta (MD) operator is then applied into each of SACF. Summing

up the all MDSACF's values for each frame, a robust wavelet-based feature parameter, *speech activity envelop* (SAE) parameter, is then proposed. Finally, we adopt an adaptive thresholding as a VAD decision. Compared to other two wavelet-based approaches respectively proposed by S. H. Chen *et al.* [28],[40] and Stegmann *et al.* [53], experimental results show that the proposed wavelet-based VAD is prior to other two wavelet-based methods and can work in a dynamically varying background noise. In addition, the proposed wavelet-based VAD is indeed an efficient and simple approach for a determination of existence of voice activity..

4.1 Introduction



So far, those common feature parameters used for the existed VAD are based on averages over windows of fixed length or derived from an analysis based on a uniform time-frequency resolution. For example, the conventional Fourier transform (FT) works well in wide sense stationary signals but fails in non-stationary since it achieves only uniform-resolution analysis. The discrete wavelet transform (DWT) [41], in contrast, calculates the decomposition into the time-frequency domain which leads to low-frequency but high-temporal resolution at high frequency bands and low-temporal but high-frequency resolution at low frequency bands. It is well known that speech signals contain many transient components and non-stationary property.

Making use of the multi-resolution analysis (MRA) property of the DWT, the classification of speech into voiced, unvoiced or transient components can be distinctly discriminated [53]. S. H. Chen *et al.* [28], [40] has been proposed a wavelet-based VAD, but their approach must requires too computation and memory requirements to meet on-line issue.

In general, the well-known “Auto-Correlation Function (ACF)” is commonly used to detect periodicity of speech. Herein the ACF is defined in subband domain and then regarded as *subband auto-correlation function* (SACF). In fact, the voiced sound has more significant periodicity than unvoiced sound and noise signal and the periodicity almost concentrates low frequency bands. So, we let the low frequency bands have high resolution to enhance the periodic property by decomposing only low band on each layer. The structure of three-layer wavelet decomposition is herein utilized to decompose speech signal into four non-uniform subbands. Although wavelet transform provides the property of MRA, how to resolve the problem about noise suppression on wavelet coefficients is mostly crucial part. Generally speaking, the noise suppressions are almost based on frequency domain. Those approaches, however, almost waste too much computing power to on-line work. We try to eliminate the noise components from the wavelet coefficients on each subband and to consider the appropriate way in terms of computing complexity. Teager energy

operator (TEO) is a powerful nonlinear operator and has been successfully used in various speech processing applications [42-45]. It is experimentally observed that the TEO which is derived from non-uniform subbands in Mel-scale can suppress the car engine noise and is easily implemented through time domain [46]. Based on the finding, we handle the process of TEO upon each subband of wavelet coefficients. In addition, the work of TEO also results in a better representation of formants so that the discriminability between speech and noise can be enhanced significantly. To accurately count the intensity of periodicity from the envelope of the SACF on each subband, the Mean-Delta (MD) method [47] is then utilized upon each subband. In [47], the MD-based feature parameter has been presented for the robust development of VAD, but is not performed well in the non-stationary noise shown in the followings. So, the sum of Mean-Delta values of Subband Auto-Correlation Function (MDSACF) derived from the wavelet coefficients of three detailed scales and one appropriated scale is defined as a robust feature parameter and regarded as *speech activity envelope* (SAE). Experimental results show that the proposed VAD has overall better performance than other two wavelet-based VAD algorithms when encountering variable-level of background noise. In addition, the proposed VAD is a simple, efficient, and robust algorithm working in a dynamically varying background noise.

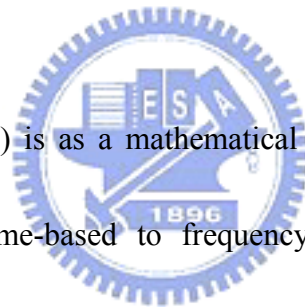
Chapter 4 is organized as follows. Section 4.2 describes the concept of discrete

wavelet transform (DWT) and shows the used structure of three-layer wavelet decomposition. Section 4.3 introduces the derivation of Teager energy operator (TEO) and displays the efficiency of subband noise suppression. Section 4.4 describes the proposed feature parameter, and the block diagram of proposed wavelet-based VAD algorithm is outlined in Section 4.5. Section 4.6 evaluates the performance of the algorithm and compare to other two wavelet-based VAD algorithms. Finally, Section 4.7 discusses the conclusions of experimental results.

4.2 Wavelet Transform

The Fourier transform (FT) is as a mathematical technique for transforming our view of the signal from time-based to frequency-based. It is found that time information is lost while in transforming to the frequency domain. However, most interesting signals contain numerous non-stationary or transitory characteristics. These characteristics are often the most important part of the signal. So, the FT is appropriate for wide sense stationary signals, but is not suitable for non-stationary signals.

The wavelet transform (WT), in contrast, is based on a time-frequency signal analysis. The wavelet analysis represents a windowing technique with variable-sized regions. It allows the use of long time intervals where we want more precise

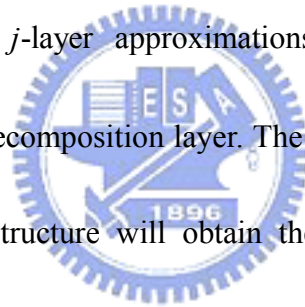


low-frequency information, and shorter regions where we want high-frequency information. It is well known that speech signals contain many transient components and non-stationary property. Making use of the MRA property of the WT, better time-resolution is needed a high frequency range to detect the rapid changing transient component of the signal, while better frequency resolution is needed at low frequency range to track the slowly time-varying formants more precisely. Through the MRA analysis, the classification of speech into voiced, unvoiced or transient components can be further exploited.

In fact, the WT is a time-scale transformation. The wavelet implied in the WT is a small wave from which many waves derived from the signal analyzed by translation and scaling. The multi-resolution capability relies on being able to dilate and translate the wavelet continuously. If we choose scales and positions based on powers of two -- so-called dyadic scales and positions -- then our analysis will be much more efficient and just as accurate. We obtain such an analysis from the DWT. An efficient way to implement this DWT using filter banks was developed in 1988 by Mallat [48]. In Mallat's algorithm, the approximations and details of input signal are accomplished by using quadrature mirror filters (QMF). Fig. 4-1 shows the basic DWT structure using filter banks for decomposing signal. Regarding to the figure, the approximated signal $L1$ and the detailed signal $H1$ are generated by respectively using the

high-pass filter and low-pass filter by the 18-tap Daubechies wavelet shown in Fig. 4-2, and the symbol $\downarrow 2$ is denoted as operator of downsampling by 2.

In order to extract speech activity accurately, the periodicity detection by using ACF is commonly used for voice activity detection. However, different frequency band provides various intensities of periodicity. In general, the strong intensity of periodicity is mainly focused on the low frequency band. Based on the finding, the input speech signal is divided into a multi-band form by using DWT. Fig. 4-3 displays the utilized structure of three-layer wavelet decomposition leading into four non-uniform subbands. The j -layer approximations L_j and details H_j of input signal are resulted from the decomposition layer. The later experiment will prove that the wavelet decomposition structure will obtain the more significant intensity of periodicity in sub-band domain than that in full-band domain.



4.3 Teager Energy Operator (TEO)

The Teager energy operator (TEO) is a powerful nonlinear operator, and can track the modulation energy and identify the instantaneous amplitude and frequency [42-45].

In continuous-time, the TEO is defined as

$$\Psi_c[s(t)] = [\dot{s}(t)]^2 - s(t)\ddot{s}(t), \quad (4-1)$$

where $s(t)$ is a continuous-time signal and $\dot{s} = ds/dt$. In discrete-time, the TEO can

be approximate by

$$\Psi_d[s(n)] = s(n)^2 - s(n+1)s(n-1), \quad (4-2)$$

where $s(n)$ is a discrete-time signal.

Let us consider a speech signal $s(n)$ degraded by uncorrelated additive noise $u(n)$.

The resulting signal is then

$$y(n) = s(n) + u(n). \quad (4-3)$$

The Teager energy of the noisy speech signal $\Psi_d[y(n)]$ is given by

$$\Psi_d[y(n)] = \Psi_d[s(n)] + \Psi_d[u(n)] + 2\tilde{\Psi}_d[s(n), u(n)], \quad (4-4)$$

where $\Psi_d[s(n)]$ and $\Psi_d[u(n)]$ are the Teager energy of the speech signal and the additive noise, respectively. $\tilde{\Psi}_d[s(n), u(n)]$ is the cross- Ψ_d energy of $s(n)$ and $v(n)$, such that

$$\tilde{\Psi}_d[s(n), u(n)] = s(n) \bullet u(n) - 0.5s(n-1) \bullet u(n+1) - 0.5s(n+1) \bullet u(n-1), \quad (4-5)$$

where the symbol \bullet is inner product. Since $s(n)$ and $u(n)$ are zero mean and independent, the $\Psi_d[s(n), u(n)]$ equals to zero. Thus, the derived formulation is approximated as below:

$$E\{\Psi_d[y(n)]\} = E\{\Psi_d[s(n)]\} + E\{\Psi_d[u(n)]\}. \quad (4-6)$$

According to [46], if $u(n)$ is the car engine noise with lowpass in nature the expected value of Teager energy of noise can be neglected due to first three autocorrelation values close to each other.

$$E\{\Psi_d[u(n)]\} = R_u(0) - R_u(2) \approx 0$$

So, the Teager energy of the observed signals almost equal speech signals.

$$E\{\Psi_d[y(n)]\} \approx E\{\Psi_d[s(n)]\}. \quad (4-7)$$

Although the TEO is valid only for noise with lowpass in nature to suppress noise, Fig. 4-4 displays that for the subband range D1~A3 the WT coefficient after TEO can discriminate speech from additive noise even if the additive noise has no lowpass nature. Inspecting the corresponding spectrograms, the voice-active durations are dominated by the some distinct harmonic frequency bands. These results are better for the periodicity detection by using subband ACF (SACF). Fig. 4-5 shows that the TEO is useful for SACF to discriminate the difference between speech and noise in detail.



4.4 The Robust Feature Extraction Derived From MDSACF

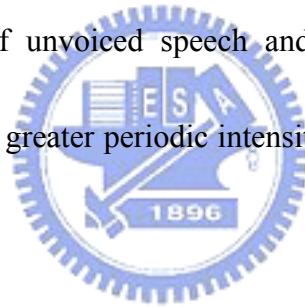
The well-known definition of the term “Auto-Correlation Function (ACF)” is usually used for measuring the self-periodic intensity of subband signal sequences shown as below:

$$R(k) = \sum_{n=0}^{p-k} s(n)s(n+k), \quad k = 0, 1, \dots, p, \quad (4-8)$$

where p is the length of ACF. k denotes as the shift of sample.

In general, the ACF is commonly used method to estimate pitch which is based on detecting the highest value of the ACF in the region of interest. In order to increase

the efficiency of ACF about making use of periodicity detection for detecting speech, the ACF is defined in subband domain, which called as “subband auto-correlation function (SACF)”, and sufficiently discriminates the dissimilarity among of voiced, unvoiced speech sounds and background noises from wavelet coefficients. In general, the voiced or vowel speech sound has more significant value of periodicity than unvoiced sound and noise signal, and the periodicity almost concentrates low frequency bands. Fig. 4-6 displays that the *normalized* SACFs ($R(0)=1$) of each subband, respectively. It is observed that the SACF of voiced speech has more obviously peaks than that of unvoiced speech and white noise. In addition, for unvoiced speech the ACF has greater periodic intensity than white noise especially in the approximation $A3$.



The intensity of periodicity on each subband is evaluated by utilizing a Mean-Delta (MD) method [47] over the envelope of each SACF. First, a measure which similar to delta cepstrum evaluation is mimicked to estimate the periodic intensity of SACF, namely “Delta Subband Auto-Correlation Function (DSACF)”, shown below:

$$\dot{R}_M(k) = \frac{\sum_{m=-M}^M m \left(\frac{R(k+m)}{R(0)} \right)}{\sum_{m=-M}^M m^2}, \quad (4-9)$$

where \dot{R}_M is DSACF over an M -sample neighborhood.

It is observed that the DSACF measure is almost like the local variation over the

SACF. Second, averaging the delta of SACF over a M -sample neighborhood \bar{R}_M , a mean of the absolute values of the DSACF (MDSACF) is given by

$$\bar{R}_M = \frac{1}{N} \sum_{k=0}^{N-1} |\dot{R}_M(k)|. \quad (4-10)$$

Observing the above formulations, the Mean-Delta method can be used to value the number and amplitude of peak-to-valley from the envelope of SACF. To develop a robust feature parameter (SAE), we sum up the four values of MDSACFs derived from the wavelet coefficients of three detailed scales and one appropriated scale.

Fig. 4-7 displays that the MRA property is important to the development of SAE feature parameter. The proposed SAE feature parameter is respectively developed with/without band-decomposition. In Fig. 4-7(b), the SAE without band-decomposition only provides obscure periodicity and confuses the word boundaries. Fig. 4-7(c)~Fig. 4-7(f) respectively show each value of MDSACF from D1 subband to A3 subband. It implies that the value of MDSACF can provide the corresponding periodic intensity for each subband. Summing up the four values of MDSACFs, we can form a robust SAE parameter. In Fig. 4-7(g), the SAE with band-decomposition can point out the word boundaries accurately from its envelope.

4.5 Proposed Voice Activity Detection (VAD) Algorithm

In this section, the proposed VAD algorithm based on DWT and TEO is presented.

Fig. 4-8 displays the block diagram of the proposed wavelet-based VAD algorithm in detail. For a given layer j , the wavelet transform decomposed the noisy speech signal into $j+1$ subbands corresponding to wavelet coefficients sets $w_{k,n}^j$. In this case, three-layer wavelet decomposition is used to decompose noisy speech signal into four non-uniform subbands including three detailed scales and one appropriated scale. Let layer $j = 3$,

$$w_{k,m}^3 = DWT\{s(n), 3\}, \quad n = 1 \dots N, \quad k = 1 \dots 4, \quad (4-11)$$

where $w_{k,m}^3$ defines the m^{th} coefficient of the k^{th} subband. N denotes as window length. The decomposed length of each subband is $N/2^k$ in turn.

For each subband signal, the TEO processing [46] is then used to suppress the noise component, and also enhance the periodicity detection. In TEO processing,

$$t_{k,m}^3 = \psi_d[w_{k,m}^3], \quad k = 1 \dots 4. \quad (4-12)$$

Next, the SACF measures the ACF defined in subband domain, and it can sufficiently discriminate the dissimilarity among of voiced, unvoiced speech sounds and background noises from wavelet coefficients. The SACF derived from the Teager energy of noisy speech is given by

$$R_{k,m}^3 = R[t_{k,m}^3], \quad k = 1 \dots 4. \quad (4-13)$$

where $R[\cdot]$ means the auto-correlation operator.

To count the intensity of periodicity from the envelope of the SACF accurately, the

Mean-Delta (MD) method [47] is utilized on each subband.

The DSACF is given by

$$\dot{R}_{k,m}^3 = \Delta[R_{k,m}^3], \quad k = 1 \dots 4. \quad (4-14)$$

where $\Delta[\cdot]$ denotes the operator of delta.

Then, the MDSACF is obtained by

$$\bar{R}_k^3 = E[\dot{R}_{k,m}^3]. \quad (4-15)$$

where $E[\cdot]$ denotes the operator of mean.

Finally, we sum up the values of MDSACFs derived from the wavelet coefficients of three detailed scales and one appropriated scale and denote as SAE feature

parameter given by

$$SAE = \sum_{k=1}^4 \bar{R}_k^3. \quad (4-16)$$



In order to point out the boundary of voice activity carefully, the VAD decision is usually performed by adaptive thresholding. To estimate the time-varying noise characteristics accurately, in this section an adaptive threshold value is derived from the statistics of SAE feature parameter during noise-only frame, and the VAD decision recursively updates the threshold using the mean and variance of the values of SAE parameters.

The above statement is same as section 2.3.1 and herein repeated. We compute the initial noise mean and variance with the first five frames, assuming that the first five

frames contain noise only. We then compute the thresholds for the speech and noise by the following:

$$T_s = \mu_n + \alpha_s \cdot \sigma_n \quad (4-17)$$

$$T_n = \mu_n + \beta_n \cdot \sigma_n \quad (4-18)$$

where T_s and T_n indicate the speech threshold and the noise threshold, respectively.

Similarly, μ_n and σ_n represent the mean and the variance among the values of SAE parameters, respectively. $\alpha_s = 5$ and $\beta_n = -1$ are the adjustment constants

which are used to determine the threshold, and are experimentally selected.

The VAD decision rule are defined as

$$\begin{aligned} &\text{if } (SAE(t) > T_s) \quad VAD(t)=1 \\ &\text{else if } (SAE(t) < T_n) \quad VAD(t)=0; \\ &\text{else } VAD(t)=VAD(t-1). \end{aligned} \quad (4-19)$$

If the detection result shows a noise period, the mean and variance among the values

of the SAE are updated by using the following:

$$\mu_n(t) = \gamma \cdot \mu_n(t-1) + (1-\gamma) \cdot SAE(t) \quad (4-20)$$

$$\sigma_n(t) = \sqrt{[SAE_{buffer}^2]_{mean} - [\mu_n(t)]^2} \quad (4-21)$$

$$[SAE_{buffer}^2]_{mean}(t) = \gamma \cdot [SAE_{buffer}^2]_{mean}(t-1) + (1-\gamma) \cdot SAE(t)^2 \quad (4-22)$$

where $\gamma = 0.95$ is chosen by experiment. $[SAE_{buffer}^2]_{mean}(t-1)$ is a mean among the

buffer of SAE value during noise-only frame. We then update the thresholds using the

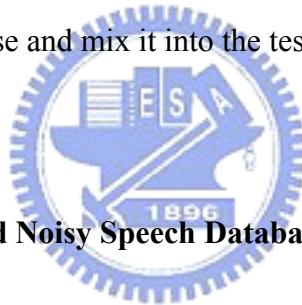
updated mean and variance of the values of SAE parameters. Fig. 4-9 displays the

VAD decision based on the adaptive threshold strategy. It is clearly found that the

boundary of voice activity is accurately extracted. The two thresholds are updated during voice-inactivity while the two thresholds are both kept during voice activity.

4.6 EXPERIMENTAL RESULTS

We evaluate the performance of proposed wavelet-based VAD and compare to other two wavelet-based approaches [28], [53] in two experiments. In our first experiment, the results of speech activity detection are tested in three kinds of background noise under various values of the SNR. In the second experiment, we adjust the variable noise-level of background noise and mix it into the testing speech signal.



4.6.1. Test Environment and Noisy Speech Database

The test environment is similar to Chapter 2, the speech corpora are collected from MAT (Mandarin across Taiwan) database including a set of isolated utterances of the ten digits in Mandarin. To vary the testing conditions, noise is added to the clean speech signal to create noisy signals at specific SNR of 40, 10, 0db and -5 dB (the SNRs is herein defined as the ratio of the power of the entire recordings containing silence and voice parts to the power of additive noise). Four noise types, including babble noise, white noise, vehicle noise and factory noise, are taken from the Noisex-92 database in turn [49]. The proposed wavelet-based VAD algorithm is based

on frame-by-frame basis, frame length $N_w = 256$ samples (or 32ms), overlap = $N_w/2$ (16ms).

4.6.2. Evaluation in Stationary Noise

In this experiment we only consider stationary noise environment. The proposed wavelet-based VAD is tested under three types of noise sources and three specific SNR values mentioned above. To evaluate the performance of the proposed VAD, two objective measures mentioned in Chapter 2 are applied here. *The probability of correctly detecting speech frames* P_{cs} is defined as the ratio of the correct speech decisions to the hand-labeled speech frames, while *the probability of falsely detecting speech frames* P_{fs} is defined as the ratio of the false speech decisions to the hand-labeled speech frames.

Table 4-I shows the comparison among the proposed wavelet-based VAD, other two wavelet-based VAD respectively proposed by Chen *et al.* [28, 40] and J. Stegmann [53]. The results from all the cases involving various noise types and SNR levels are averaged and summarized in the bottom row of this table. From Table I, the proposed wavelet-based VAD and Chen's VAD algorithms are all superior to Stegmann's VAD over all SNRs under various types of noise. In terms of the average correct and false speech detection probability, the proposed wavelet-based VAD is comparable to

Chen's VAD algorithm. Both the algorithms are based on the DWT and TEO processing. However, Chen *et al.* decomposed the input speech signal into 17 critical-subbands by using perceptual wavelet packet transform (PWPT). To obtain a robust feature parameter, called as "VAS" parameter, each critical subband after their processing is synthesized individually while other 16 subband signals are set to zero values. Next, the VAS parameter is developed by merging the values of 17 synthesized bands. Compare to the analysis/synthesis of wavelet from S. H. Chen *et al.*, we only consider analysis of wavelet. The structure of three-level decomposition leads into four non-uniform bands as front-end processing. For the development of feature parameter, we do not again waste extra computing power to synthesize each band. Besides, Chen's VAD algorithm must be performed in entire speech signal. The algorithm is not appropriate for real-time issue since it does not work on frame-based processing. Conversely, in our method the decisions of voice activity can be accomplished by frame-by-frame processing. Table 4-II indicates that the subjective evaluations of listening test for the extracted voice activity. The clean speech is obtained by a subtractive-type algorithm presented by Berouti *et al.* [39] to avoid the annoying "musical noise" introduced by subtracting the estimated noise spectrum from the observed speech signals. A five-scale absolute opinion from 1 (poor) to 5 (excellent) was adopted to subjectively evaluate the six listeners. Table 4-III displays

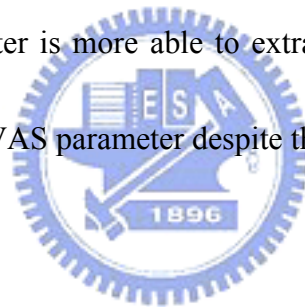


the computing time for the listed VAD algorithms running Matlab programming in Celeron 2.0G CPU for processing 118 frames of an entire recording. It is found that the computing time of Chen's VAD is nearly four times greater than that of other two VADs. Besides, the computing time of Chen's VAD is closely relative to the entire length of recording.

4.6.3. Evaluation in Non-stationary Noise

To evaluate the performance of the presented VAD in highly non-stationary whose statistical properties change over time, we add the decreasing and increasing level of background noise on a clean speech sentence in English and the SNR is set 0 dB. Comparisons among proposed wavelet-based VAD, other wavelet-based VAD proposed by S. H. Chen *et al.* [28, 40] and MD-based VAD proposed by A. Ouzounov [47] are exhibited in Fig. 4-10. From the figure, the mixed noisy sentence "May I help you?" is shown in Fig. 4-10(a). The increasing noise-level and decreasing noise-level are added into the front and the back of clean speech signal. Additionally, an abrupt change of noise is also added in the middle of clean sentence. The three envelopes of VAS, MD and SAE feature parameters are showed in Fig. 4-10(b)~Fig. 4-10(d), respectively. It is found that the performance of Chen's VAD algorithm seems not good in this case. The envelope of VAS parameter closely depends on the variable

level of noise. Similarly, the envelope of MD parameter fails in variable level of noise. Conversely, the envelope of proposed SAE parameter is insensitive to variable-level of noise. So, the proposed wavelet-based VAD algorithm is performed well in non-stationary noise. Fig. 4-11 illustrates the performance of Stegmann's VAD [53] for four energy-based parameters and VAD decision. It is found that the associated energy-based parameters are close to the variation of noise and these parameters cause Stegmann's VAD to fails in a variable-level of noise. Fig. 4-12 illustrates the performance of the proposed VAD for an utterance produced continuously. Similarly, the envelope of SAE parameter is more able to extract the exact boundary of voice activity than the envelope of VAS parameter despite the variable noise level.



4.7 Discussion & Future Work

In discussion, a new feature parameter is based on the sum of the values of MDSACFs derived from the wavelet coefficients of three detailed scales and one appropriated scale to develop a robust VAD algorithm. The robust VAD is a multi-band based algorithm using DWT and TEO processing. By means of the MRA property of DWT, the ACF defined in subband domain sufficiently discriminates the dissimilarity among of voiced, unvoiced speech sounds and background noises from

wavelet coefficients. For the problem about noise suppression on wavelet coefficients, a nonlinear TEO is then utilized into each subband signals. Additionally, we have shown that the Teager energy set of wavelet coefficients is beneficial for the result of ACF since TEO can provide a better representation of formants resulting distinct periodicity. Experimental results have shown that the proposed feature parameter can point out the boundary of speech activity and its envelope is insensitive to variable noise-level environment. Compare to other two VAD approaches, the three major advantages of proposed VAD are shown as below:

4.7.1. Comparison

➤ *Suitable for on-line work:*



Stegmann's VAD [53] and the proposed MDSSACF-based VAD algorithm are both performed on-line. Compare with Chen's wavelet-based VAD, the proposed MDSSACF-based VAD algorithm can be performed on frame-by-frame basis. An on-line adaptive thresholding strategy is just used for deciding the result of VAD immediately, so our method more meets the on-line issue than Chen's wavelet-based VAD. In fact, the processing of the Chen's VAD must require much memory space to store entire input speech.

➤ ***Efficient and simple implementation:***

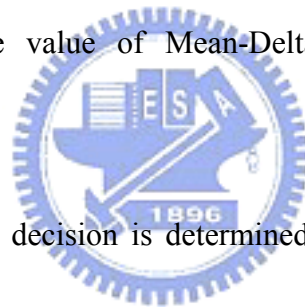
In the architecture of proposed wavelet-based VAD, we can only consider the decomposition of wavelet and do not again waste extra computing power to synthesize each subband for wavelet reconstruction. In fact, input speech is decomposed into four non-uniform subbands resulted from three-layer wavelet decomposition. In order to obtain the VAS feature parameter, we only sum up four values of MDSACFs on each subband. In addition, the determination about MDSACF also needs only the little computing complexity. Similar to proposed wavelet-based VAD, Stegmann's VAD also only considers the wavelet decomposition and further uses the logical combination among a set of detections: silence detection, stationary detection and background-noise detection. It seems to be easily implemented for extracting voice active frames.

Conversely, in Chen's wavelet-based VAD algorithm [18],[40], the input speech signal is decomposed into 17 critical-subbands by using perceptual wavelet packet transform (PWPT). Each critical subband is then synthesized individually while other 16 subband signals are set to zero values. Next, the VAS parameter is developed by summing up the values of 17 synthesized bands. So, Chen's wavelet-based VAD needs extra computing time to achieve the wavelet reconstruction and its computing complexity is more than other two wavelet-based methods: Stegmann's VAD and

proposed VAD.

➤ ***Reliable performance in variable noise-level environments:***

Of all the listed VADs in this paper, the proposed wavelet-based is prior to other two VADs in variable-level of background noise. From experimental results, the proposed SAE feature parameter can point out the boundaries of the noisy sentence in variable noise-level conditions. It is found that the SAE parameter relies on the periodicity of SACF defined in each subband. Through the subband TEO processing and normalization of SACF, the value of Mean-Delta on SACF is insensitive to variable-level of noise.



Conversely, in [53] the VAD decision is determined by a set of four energy-based parameters is derived from the WT coefficients and compared to fixed thresholds resulting in four binary flags that indicates silence, stationary and background noise in two different frequency regions, respectively. Finally, these flags are combined to get the VAD decision. It is found that the energy-based parameter is relatively sensitive to variable-level on noise and the fixed thresholds are also relatively appropriate for this case. In addition, the VAS parameter derived from Chen's VAD is observed that its envelope is closely depended on the variable noise level even though the de-noising processing is utilized into 17 critical subbands.

4.7.2. Future Work

The well-known speech enhancement based on wavelet thresholding is recently used for obtaining desired non-annoying speech signals. In this method, the subband threshold is usually assumed uniform or non-uniform for each subband. In fact, the value of subband threshold must be adapted to time-varying noise, especially in classification of voice activity/absence. During voice-absent frames, we can further adjust the value of subband threshold to obtain silence-like signals. Conversely, during voice-active frames, the adjustment of subband threshold is assumed to be related to masking properties of the human auditory system.

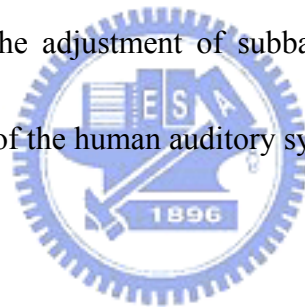


TABLE 4-I
COMPARISON BETWEEN THE THREE VAD ALGORITHMS TESTING IN VARIOUS
NOISE CONDITIONS

Noise Conditions		P_{cs} (%)			P_{fs} (%)		
Type	SNR(dB)	Proposed VAD	Chen's VAD [28],[40]	Stegmann's VAD [54]	Proposed VAD	Chen's VAD [28],[40]	Stegmann's VAD [54]
Vehicle Noise	40	99.6	97.7	93.3	1.8	2.9	5.4
	10	97.8	96.1	85.7	2.5	3.3	9.1
	0	94.9	93.2	76.5	2.6	6.9	14.3
	-5	91.9	88.2	71.9	2.9	9.9	20.9
Babble Noise	40	97.9	95.8	93.4	3.2	4.8	7.9
	10	92.4	89.4	82.5	5.6	8.3	14.5
	0	87.4	85.6	75.9	7.3	14.9	20.1
	-5	82.2	81.8	69.1	10.3	16.8	29.8
Factory Noise	40	98.1	97.4	92.6	3.4	5.3	6.3
	10	93.2	94.1	80.8	4.2	7.2	12.8
	0	88.4	88.3	76.1	6.8	12.7	18.4
	-5	85.8	85.6	70.7	9.4	15.4	26.2
White Noise	40	97.5	97.6	95.4	1.2	1.9	4.3
	10	93.3	98.1	90.3	1.7	2.4	7.9
	0	90.4	91.9	80.2	2.1	3.1	10.6
	-5	88.4	85.4	77.4	3.1	3.9	14.8
Average		92.45	91.64	81.99	4.26	7.48	13.96

TABLE 4-II
SUBJECTIVE EVALUATION OF LISTENING TEST

VAD types	Avg. of score
Proposed VAD	4.6
Chen's VAD	4.1
Stegmann's VAD	3.2

TABLE 4-III
ILLUSTRATION OF EFFICIENCY FOR THE FOUR VAD

VAD types	Computing time (sec)
Proposed VAD	0.089
Chen's VAD	0.436
Stegmann's VAD	0.097

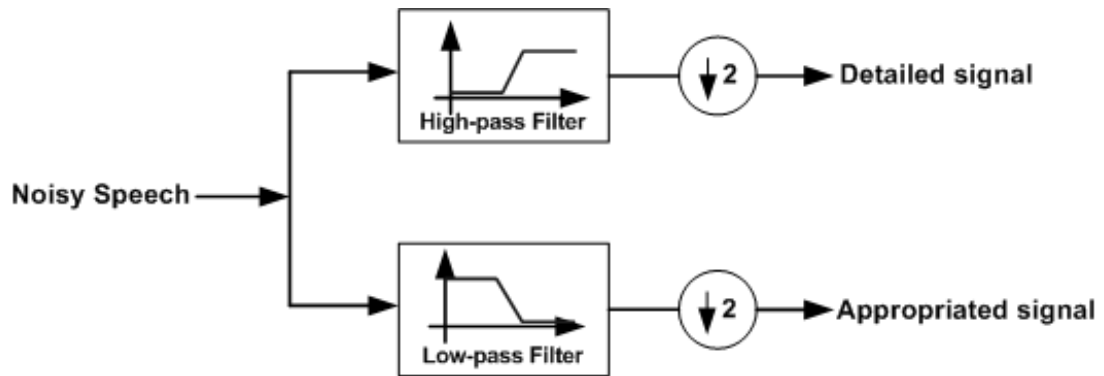


Fig. 4-1 Discrete wavelet transform (DWT) using filter banks.

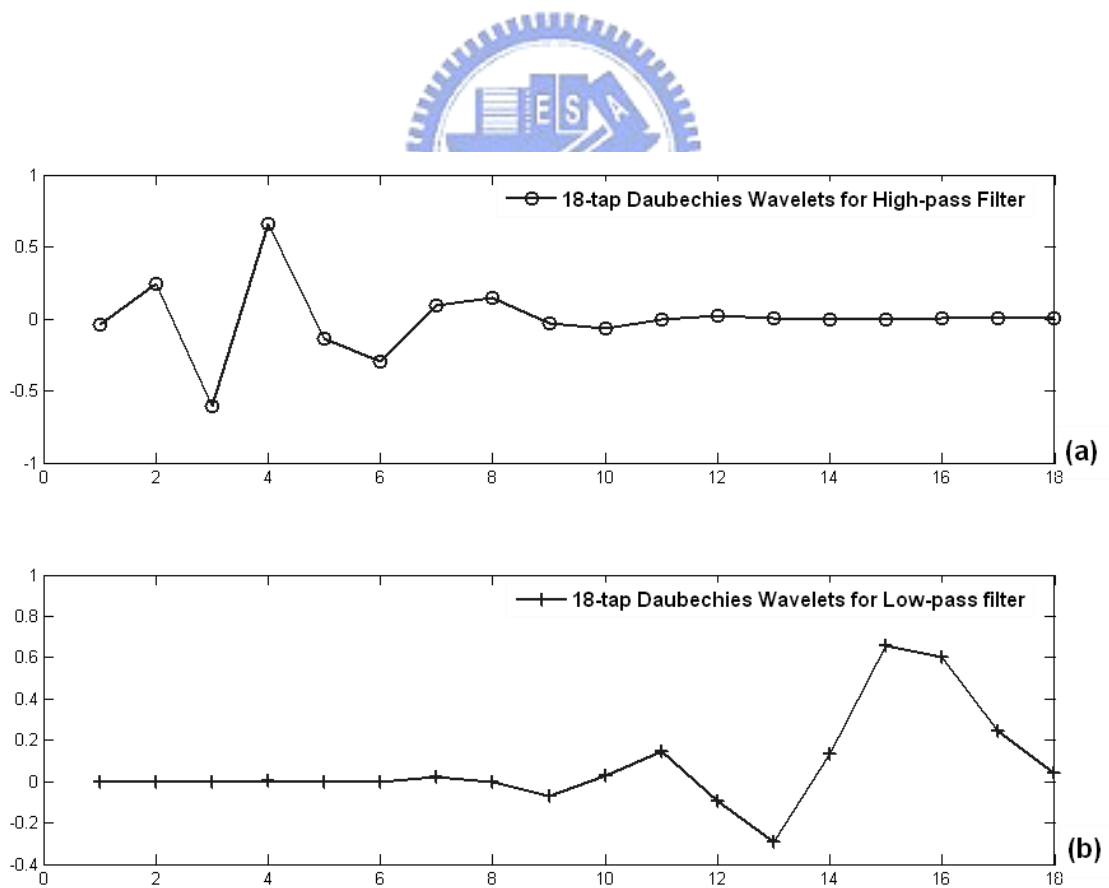


Fig. 4-2 18-tap Daubechies Wavelets.

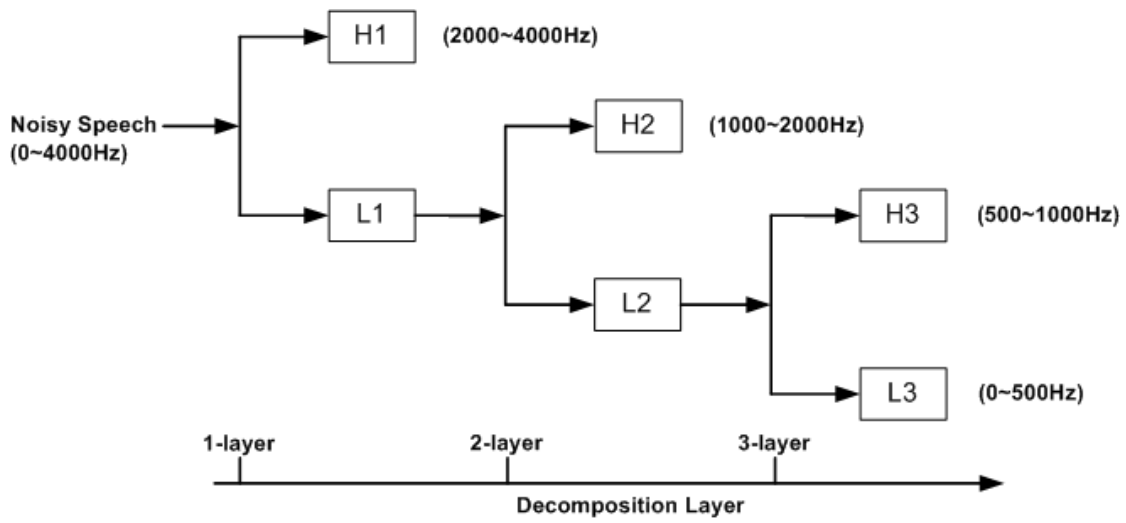


Fig. 4-3 Structure of three-layer wavelet decomposition.

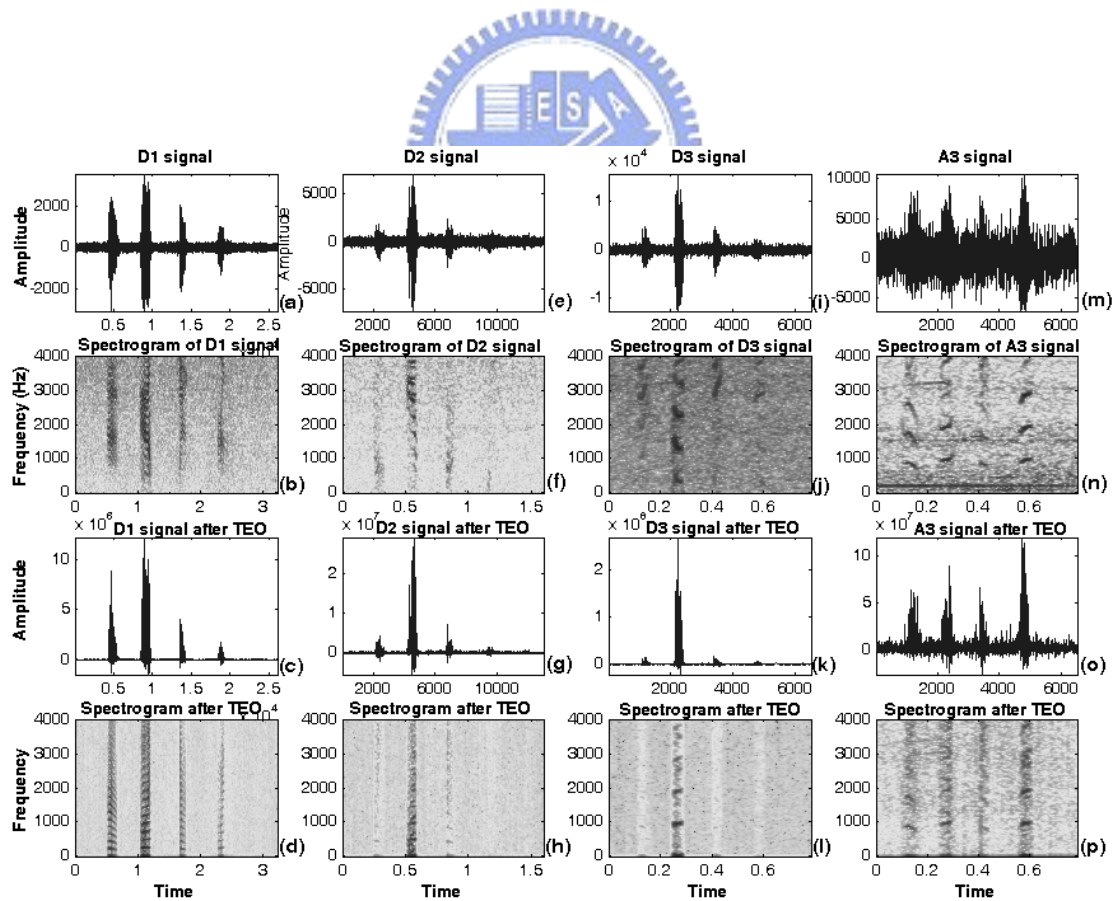


Fig. 4-4 Illustration of efficiency of TEO for four subbands.

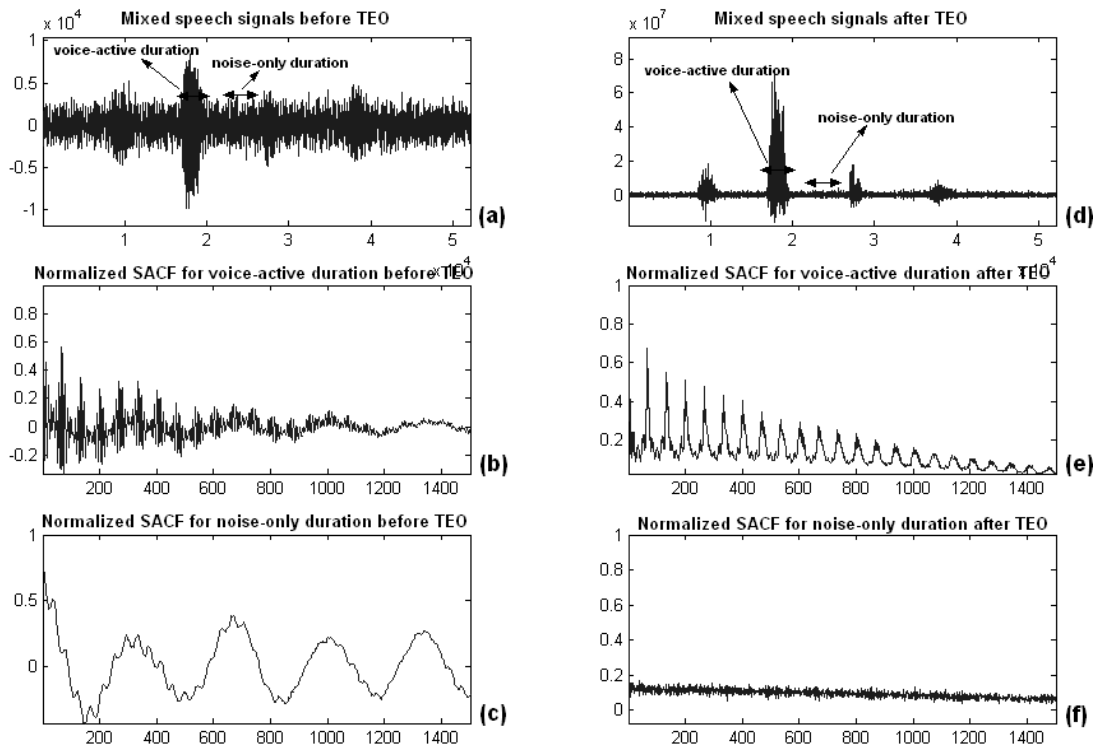


Fig. 4-5 Illustration of TEO for enhancing the discrimination between speech and noise.

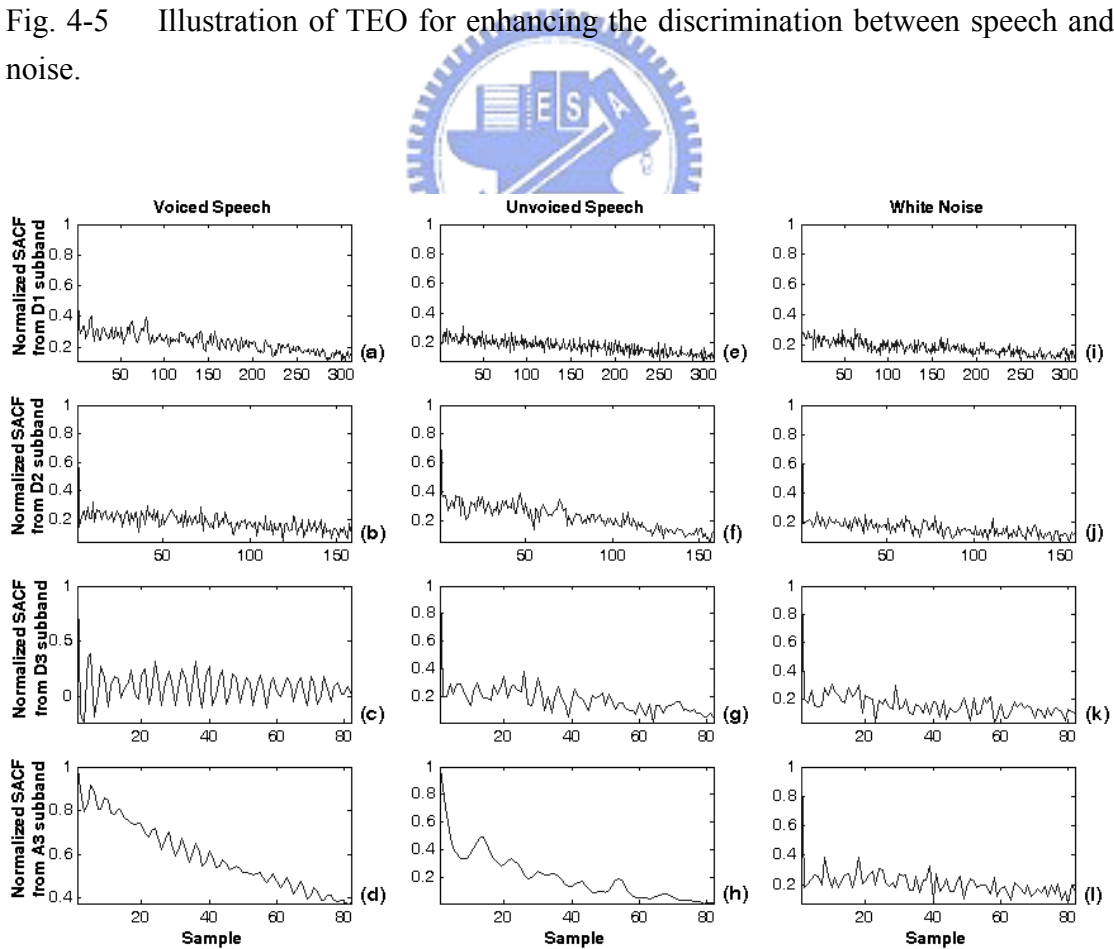


Fig. 4-6 Examples of normalized SACF for voiced sound, unvoiced sound and white noise on each subband.

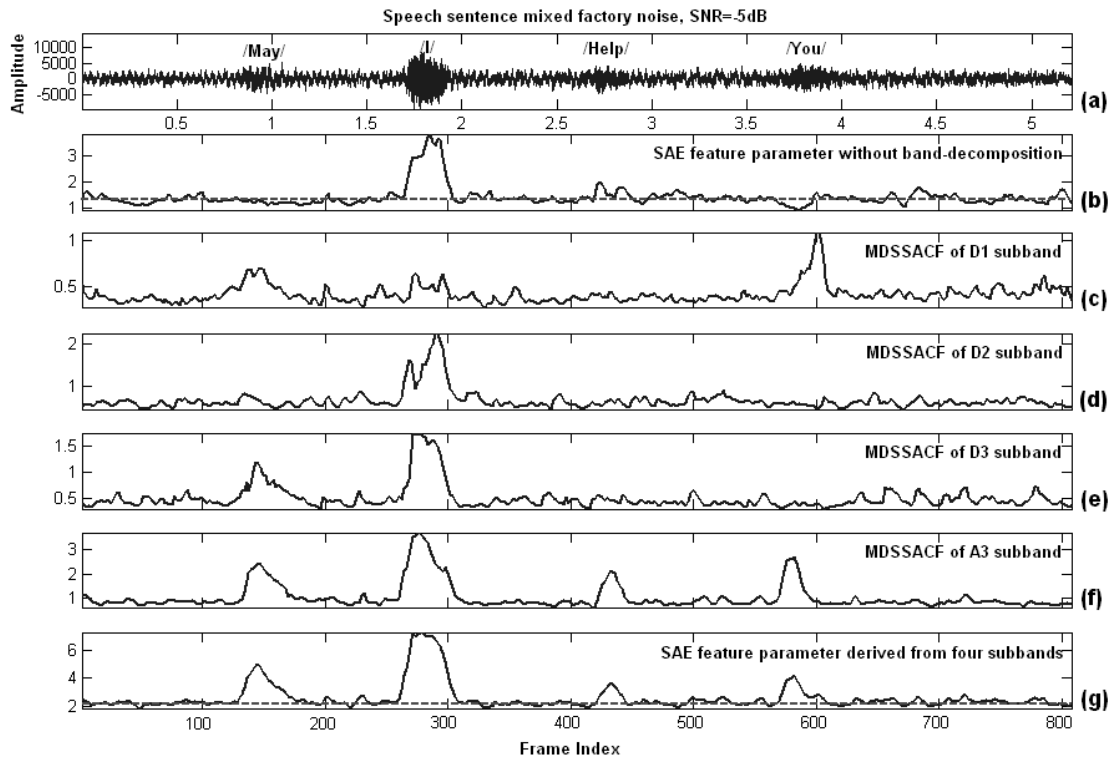


Fig. 4-7 Examples of the SAE parameters without band-decomposition and derived from four subbands.

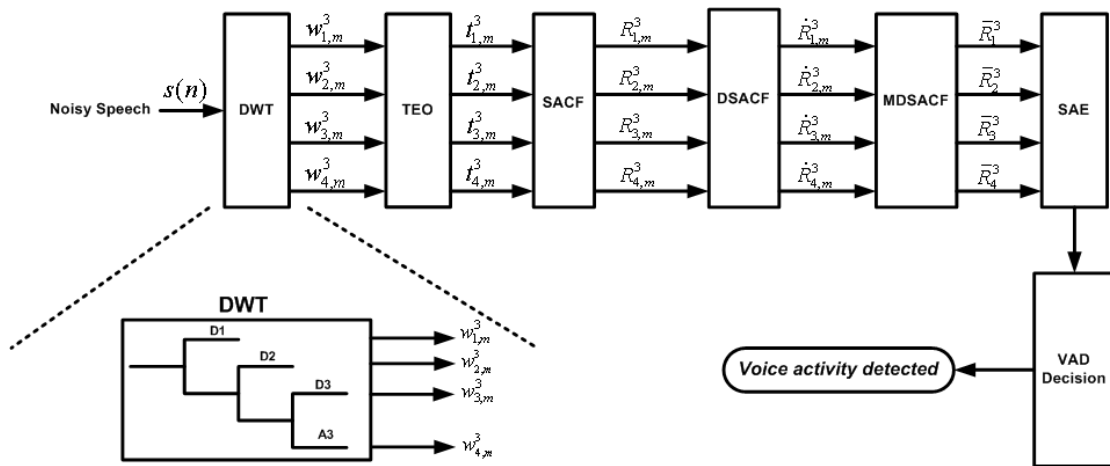


Fig. 4-8 Block diagram of proposed wavelet-based VAD.

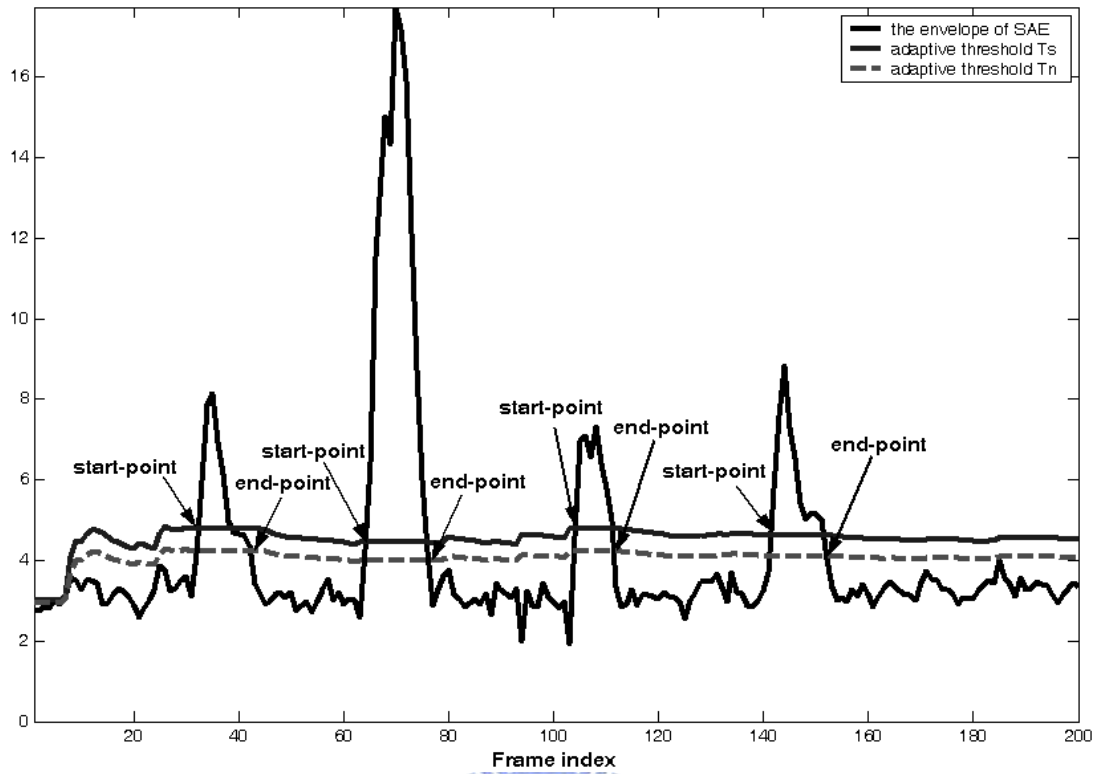


Fig. 4-9 Adaptive thresholding strategy for extracting the boundary of voice activity.

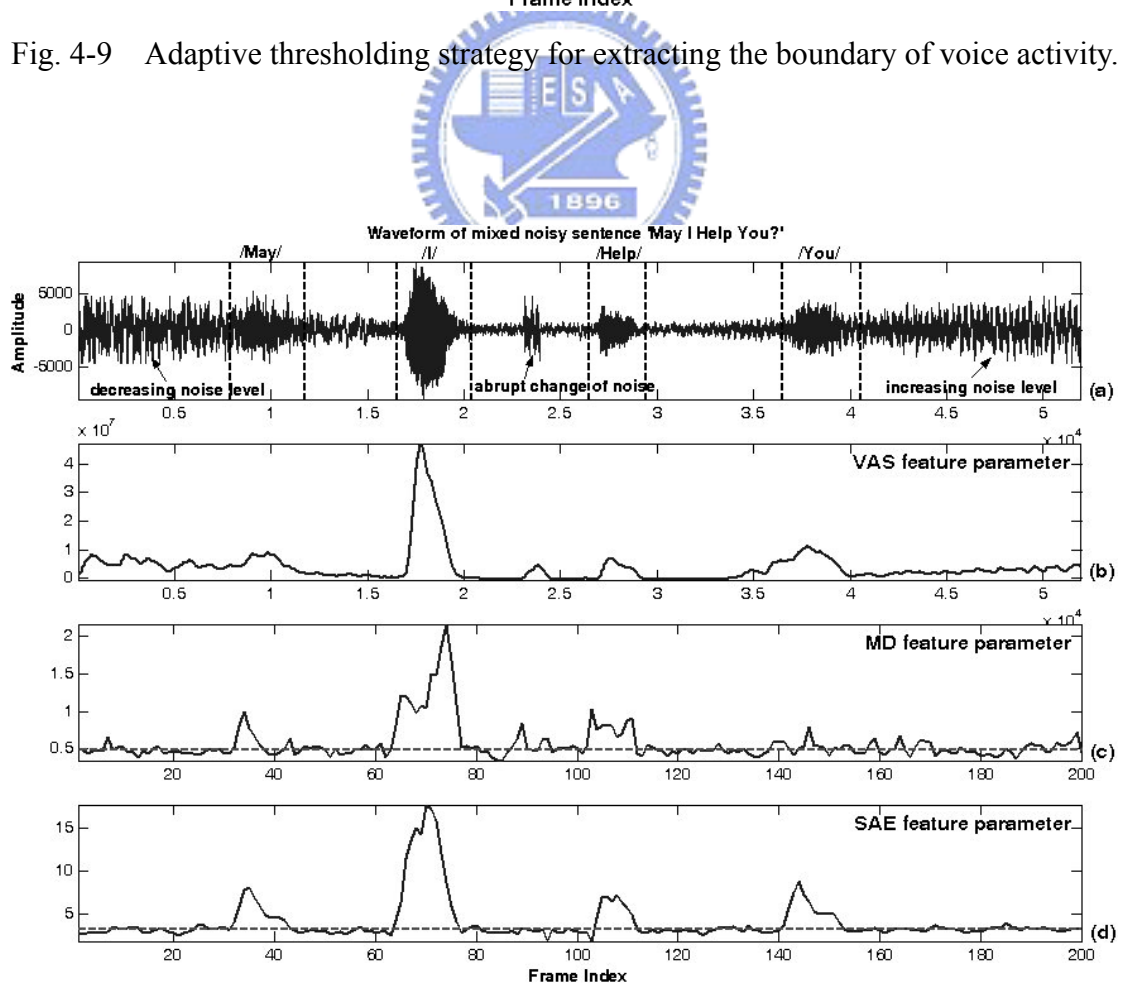


Fig. 4-10 Comparisons among VAS, MD and proposed SAE feature parameters.

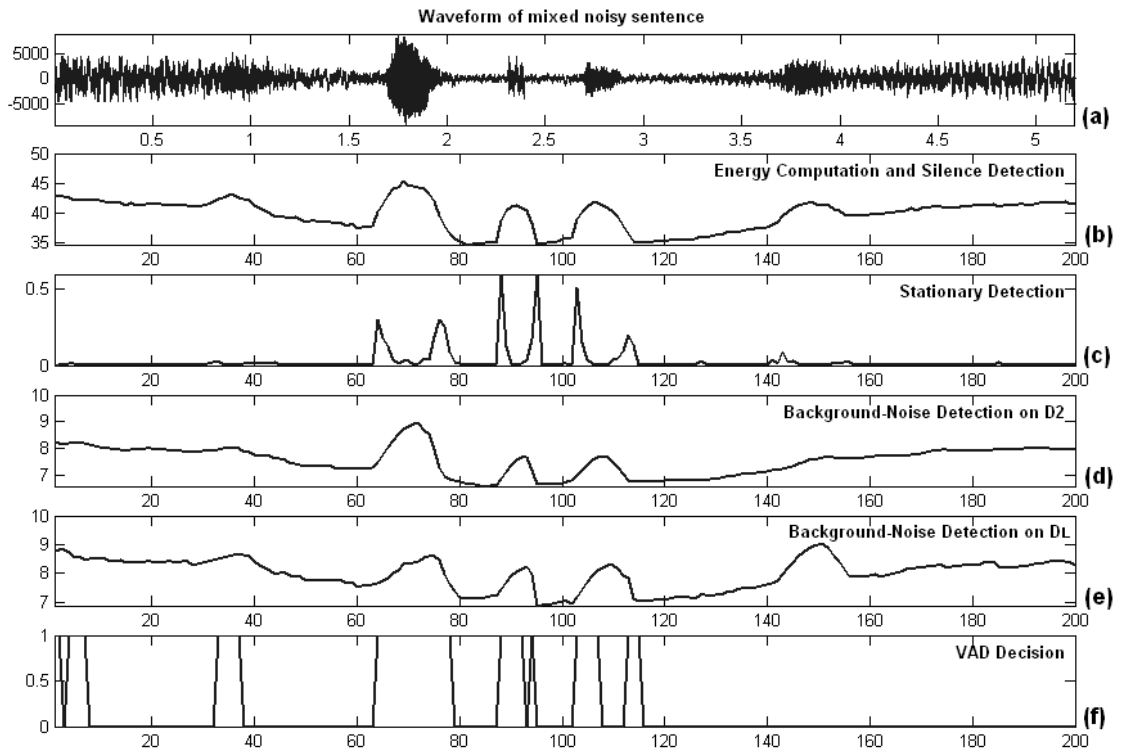


Fig. 4-11 Illustration of performance for Stegmann's VAD [53] containing four energy-based parameter and VAD decision.

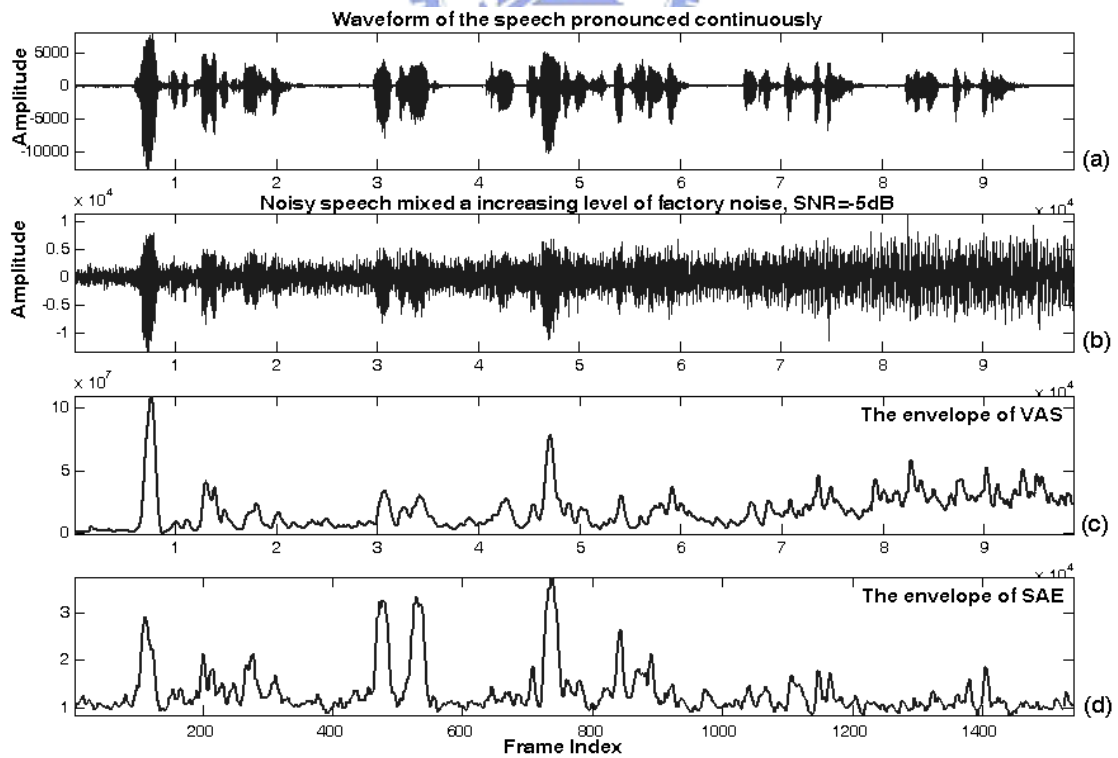


Fig. 4-12 Effects of a variable noise-level on proposed SAE and Chen's VAS parameters under a noisy speech sentence consisting continuous words.

CHAPTER 5

CONCLUSIONS

In this dissertation, we attempt to generate a robust feature set for classifying voice activity/absence frames. In Chapter 2 and Chapter 4, the banded lines existing only on voice-active spectrogram and periodicity measured in subband domain are used for characterizing speech signal, respectively. Experimental results show that the proposed two feature parameters are indeed robust against various types of noises. Consequently, the two types of voice activity detection (VAD) algorithms are built in turn. In addition, the two VAD algorithms can provide on-line work being suitable for real-world. In Chapter 3, an extended application of VAD strategy is outlined to form a fast adaptive algorithm for estimating noise power in non-stationary environments.

Based on the above experimental results, the comparisons between the proposed two VAD algorithms can be summarized as below:

1. The proposed two VAD algorithms not only perform well at low SNRs but also are insensitive to changes in the noise level, even if noise rapidly changes in level (or called dynamic background noise)
2. Due to the robustness of feature extraction for the two VADs, we can almost use fixed thresholds as VAD decisions so that the computing complexity can be significantly decreased.
3. The proposed two methods provide easy and interesting ideas for implementing a robust VAD.
4. Finally, compare to frequency band analysis, the second presented VAD via wavelet analysis is better than the first presented VAD based on entropy measurement shown in TABLE 5-I. Especially for low SNRs, the wavelet-based VAD is prior to the entropy-based VAD in term of P_{cs} . Similarly, in term of the average false speech detection probability, the wavelet-based VAD is also prior to the entropy-based VAD, especially in noise source with low frequency power so that vehicle noise since the noises with low frequency power is decreased using TEO applied into each wavelet coefficient.
5. In addition, for the entropy-based VAD the number of decomposed subband is assumed fixed and subband bandwidth is uniform, it does not meet the time-varying banded lines on voice-active spectrogram. Conversely, wavelet

analysis is more suitable for speech processing than frequency analysis.

TABLE 5-I
COMPARISON BETWEEN PROPOSED ENTROPY-BASED VAD AND
WAVELET-BASED VAD

Noise Conditions		P_{cs} (%)		P_{fs} (%)	
Type	SNR(dB)	Proposed entropy-based VAD	Proposed wavelet-based VAD	Proposed entropy-based VAD	Proposed entropy-based VAD
Vehicle Noise	40	98.4	99.6	1.2	1.8
	10	94.2	97.8	3.9	2.5
	0	91.3	94.9	4.6	2.6
	-5	89.3	91.9	6.5	2.9
Babble Noise	40	96.1	97.9	2.6	3.2
	10	89.2	92.4	5.8	5.6
	0	84.8	87.4	7.4	7.3
	-5	79.6	82.2	10.4	10.3
Factory Noise	40	97.9	98.1	2.1	3.4
	10	91.7	93.2	4.6	4.2
	0	86.3	88.4	7.8	6.8
	-5	84.1	85.8	9.1	9.4
White Noise	40	99.8	97.5	0.9	1.2
	10	96.6	93.3	1.9	1.7
	0	93.1	90.4	2.2	2.1
	-5	91.9	88.4	2.9	3.1
Average		91.52	92.45	4.62	4.26

BIBLIOGRAPHY

- [1] R. V. Cox and P. Kroon, "Low bit-rate speech coders for multimedia communications," *IEEE Commu. Mag.*, vol. 34, pp. 34-41, Dec. 1996.
- [2] K. Itoh and M. Mizushima, "Environmental noise reduction based on speech/non-speech identification for hearing aids," *Proc. ICASSP'97*, pp. 419-422, April 1997.
- [3] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan European digital cellular mobile telephone service," *Proc. ICASSP'89*, pp. 369-372, May 1989.
- [4] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [5] European Telecommunication Standard GSM 06.31, "European digital cellular telecommunications system (Phase 2); Discontinuous Transmission (DTX) for full rate speech traffic channel," Sep. 1994.
- [6] European Telecommunication Standard GSM 06.32, "European digital cellular telecommunications system (Phase 2); Voice Activity Detection (VAD)," Sep. 1994.
- [7] G. D. Wu, and C. T. Lin, "Word boundary detection with mel-scale frequency

- bank in noise environment,” *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 541-554, 2000.
- [8] B. F. Wu and K. C. Wang, “A Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 762-775, Sep. 2005.
- [9] L. R. Rabiner and M. R. Sambur, “Voiced-unvoiced-silence detection using the Itakura LPC distance measure,” *Proc. ICASSP’77*, pp. 323-326, May 1977.
- [10] L. G. Wilpon, L. R. Rabiner, and T. Martin, “An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints,” *AT&T Bell Labs. Tech. J.*, vol. 63, pp. 479-498, Mar. 1984.
- [11] R. Chengalvarayan, “Robust energy normalization using speech/non-speech discriminator for German connected digit recognition,” *Proc. Eurospeech’99*, Budapest, Hungary, pp. 61-64, Sep. 1999.
- [12] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” *Bell Sys. Tech. J.*, vol. 54, pp. 297-315, Feb. 1975.
- [13] J. C. Junqua, B. Reaves, and B. Mak, “A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize,” *Proc.*

Eurospeech, pp. 1371-1374, 1991.

- [14] J. A. Haign and J. S. Mason, "Robust voice activity detection using cepstral features," *Proc. IEEE TEN-CON*, pp. 321-324, 1993, China.
- [15] N. B. Yoma, F. McInnes, and M. Jack, "Robust speech pulse-detection using adaptive noise modeling," *Electron. Lett.*, vol. 32, July 1996.
- [16] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Elect. Eng.*, vol. 139, pp. 377-380, Aug. 1992.
- [17] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," *Proc. IEEE ICASSP*, vol. 1, pp. 365-368, 1998.
- [18] Y. D. Cho and A. Kondo, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Lett.*, vol. 8, pp. 276-278, Jan. 2001.
- [19] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE J. Select. Areas Comm.*, vol. 16, pp. 1818-1829, Dec., 1998.
- [20] E. Nemer, R. Goubran and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 217-231, 2001.
- [21] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit,

“ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications,” *IEEE Communications Magazine*, vol. 35, pp. 64-73, Sep. 1997.

[22] J. F. Chen and W. Ser, “Speech detection using microphone array,” *Electronics Letters*, vol. 36, issue 2, pp. 181–182, Jan. 2000.

[23] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, Jan. 1999.

[24] Y. D. Cho and A. Kondoz, “Analysis and improvement of a statistical model-based voice activity detector,” *IEEE Signal Process. Lett.*, vol. 8, no. 10, Oct. 2001.

[25] L. Lamel, L. Labiner, A. Rosenberg, and J. Wilpon, “An improved endpoint detect for isolated word recognition,” *IEEE ASSP Magazine*, vol. 29, pp. 777-785, 1981.

[26] M. H. Savoji, “A robust algorithm for accurate endpoint of speech,” *Speech Communication*, vol. 8, pp. 45-60, 1989.

[27] H. Ney, “A optimization algorithm for determining the endpoints of isolated utterances,” *Proc. ICASSP’81*, pp. 720-723, 1981.

[28] J. F. Wang and S. H. Chen, “A voice activity detection algorithm based on perceptual wavelet packet transform and teager energy operator,” *ISCSLP’02*,

pp. 177-180, Aug. 2002.

- [29] J. C. Junqua, B. Mak, and B. Revaes, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 406-412, July 1994.
- [30] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," *Proc. ICSLP-98*, 1998.
- [31] L. S. Sheng and C. H. Yang, "A novel approach to robust speech endpoint detection in car environments," *Proc. ICASSP*, pp. 1751-1754, 2000.
- [32] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, pp. 1751-1754, 2000.
- [33] S. Basu, B. Clarkson, and A. Pentland, "Smart headphones: enhancing auditory awareness through robust speech detection and source localization," *Proc. ICASSP*, pp. 3361-3364, 2001.
- [34] C. T. Lin, J. Y. Lin and G. D. Wu, "A robust word boundary detection algorithm for variable noise-level environment in cars," *IEEE Trans. Intelligent Transportation System*, vol. 3, pp. 89-101, March 2002.
- [35] S. V. Gerven, and F. Xie, "A comparative study of speech detection methods," *Proc. Eurospeech*, vol. 3, pp. 1095-1098, 1997.

- [36] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, July 2001.
- [37] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, Jan. 2002.
- [38] L. Lin, W. H. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," *Electronics Letters*, vol. 39, no. 9, pp. 754-755, May 2003.
- [39] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. of ICASSP'79*, pp. 208-211, April 1979.
- [40] S. H. Chen and J. F. Wang, "A Wavelet-based Voice Activity Detection Algorithm in Noisy Environments," *2002 IEEE International Conference on Electronics, Circuits and Systems (ICECS2002)*, pp. 995-998, 2002.
- [41] G. Strang, T. Nquyen, "Wavelet and Filter Banks," *Wellesley-Cambridge Press*, 1996.
- [42] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," *Proc. ICASSP'90*, pp. 381-384, 1990.
- [43] A. C. Bovik, P. Maragos, and T. Quatieri, "AM-FM energy detection and

separation in noise using multiband energy operators,” *IEEE Trans. Signal Processing*, vol. 41, pp. 3245-3265, 1993.

[44] P. Maragos, J. F. Kaiser, and T. Quatieri, “Energy separation in signal modulations with application to speech analysis,” *IEEE Trans. Signal Processing*, vol. 41, pp. 3025-3051, 1993.

[45] P. Maragos, T. Quatieri, and J. F. Kaiser, “On amplitude and frequency demodulation using energy operators,” *IEEE Trans. Signal Processing*, vol. 41, pp. 1532-1550, 1993.

[46] F. Jabloun, A. E. Cetin, and E. Erzin, “Teager energy based feature parameters for speech recognition in car noise,” *IEEE Signal Processing Lett.*, vol. 6, pp. 259-261, 1999.



[47] A. Ouzounov, “A Robust Feature for Speech Detection,” *Cybernetics and Information Technologies*, vol. 4, no 2, pp. 3-14, 2004.

[48] S. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation,” *IEEE Pattern Anal. and Machine Intell.*, vol. 11, no. 7, pp. 674-693, 1989.

[49] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Commun.*, vol. 12, pp. 247-251,

1993.

[50] A. M. Kondo, *Digital Speech Coding for Low Bit Rate Communications Systems*, John Wiley & Sons Ltd. 1994.

[51] J. W. Kim, M. S. Seo, B. S. Yoon, S. I. Choi and Y. G. You, "A voice activity detection algorithm for wireless communication systems with dynamically varying background noise," *IEICE Trans. Commu.*, vol. E83-B, no. 2, Feb. 2000.

[52] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," *Proc. EUROSPEECH*, pp. 1513-1516, 1995.

[53] J. Stegmann and G. Schroder, "Robust voice-activity detection based on the wavelet transform," *Speech Coding for Telecommunications Proceeding*, pp. 99-100, 1997.

[54] J. Stegmann and G. Schroder, "Robust voice-activity detection based on the wavelet transform," *IEEE Workshop on Speech Coding for Telecommunications Proceeding*, pp. 99 - 100, Sep. 1997.

VITA

博士候選人簡歷

姓名：王坤卿

性別：男

生日：民國 65 年 2 月 24 日

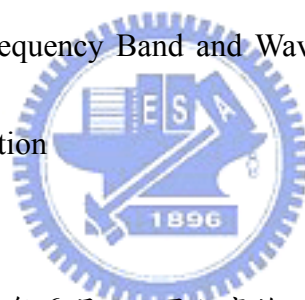
籍貫：台灣省嘉義縣

論文題目：

中文：以頻帶及小波分析為基礎的強健性語音偵測系統之研究

英文：A Study of Frequency Band and Wavelet Analysis for Robust Voice

Activity Detection



學歷：

1. 民國 80 年 9 月~85 年 6 月 國立高雄工商專電機工程科
2. 民國 85 年 9 月~87 年 6 月 私立南台技術學院電機工程系
3. 民國 87 年 9 月~89 年 6 月 私立逢甲大學電機研究所
4. 民國 89 年 9 月~迄今 國立交通大學電機與控制研究所博士班

經歷：

1. 民國 92 年 9 月~94 年 1 月 私立明新科技大學兼任講師
2. 民國 92 年 9 月~94 年 1 月 私立大華技術學院兼任講師
3. IEEE Student Member (2002-2005)
4. ISCA Student Member (2005-2006)

榮譽：

1. 民國 93 年 旺宏金砂獎第三屆旺宏金砂二獎

PUBLICATION LIST

博士候選人著作目錄

姓名：王坤卿 (Kun-Ching Wang)

Journal

- [1] Bing-Fei Wu and Kun-Ching Wang, “A Robust Entropy-Based Speech Detection in High Noisy Environments,” *GESTS International Transactions on Speech Science and Engineering*, vol.2, no.1, pp.79-90, Feb. 2005.
- [2] Bing-Fei Wu and Kun-Ching Wang, “A Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 762-775, Sep. 2005.
- [3] Bing-Fei Wu and Kun-Ching Wang, “Noise Spectrum Estimation with Entropy-based VAD in Non-stationary Environments,” *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*. (Accepted, Aug. 2005)
- [4] Bing-Fei Wu and Kun-Ching Wang, “Speech Activity Detection Based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator,” *International Journal of Computational Linguistics and Chinese Language Processing*. (Accepted, Sep. 2005)
- [5] Bing-Fei Wu and Kun-Ching Wang, “Wavelet Based Voice Activity Detection Algorithm in Non-stationary Noise,” *IEEE Transactions on Speech and Audio Processing*. (Revised, Sep. 2005)

- [6] Bing-Fei Wu and Kun-Ching Wang, “An Efficient Voice Activity Detection in Noisy Environments,” *WSEAS Transactions on Acoustics and Music*. (Submitted, June 2005)

Reference

- [1] Bing-Fei Wu and Kun-Ching Wang, “An Adaptive Band-Partitioning Spectral Entropy Based Speech Detection in Realistic Noisy Environments,” *INTERSPEECH 2004 ICSLP*, vol. 2, pp. 957~960, Oct. 4~8, 2004, Jeju Island, Korea.
- [2] Bing-Fei Wu, Kun-Ching Wang and Lung-Yi Kuo, “A Noise Estimator with Rapid Adaptation in Variable-Level Noisy Environments,” *ROCLING XVI*, pp. 33~38, Sep. 2~3, 2004, Taipei, Taiwan.

