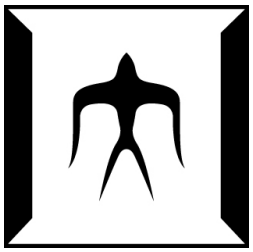
論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Efficient Voice Activity Detection and Speech Enhancement Algorithms based on Spectral Features
著者(和文)	MaYanna
Author(English)	Yanna Ma
出典(和文)	学位:博士（工学）, 学位授与機関:東京工業大学, 報告番号:甲第9632号, 授与年月日:2014年9月25日, 学位の種別:課程博士, 審査員:西原　明法,國枝　博昭,山田　功,府川　和彦,篠田　浩一,入野　俊夫
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第9632号, Conferred date:2014/9/25, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Efficient Voice Activity Detection and Speech Enhancement Algorithms based on Spectral Features



Department of Communications and Integrated Systems
Tokyo Institute of Technology
Yanna Ma

June 2014

Efficient Voice Activity Detection and Speech Enhancement Algorithms based on Spectral Features

by

Yanna Ma

A dissertation submitted in partial fulfillment of the
requirement for the degree of
Doctor of Philosophy

Advisor: Professor Akinori Nishihara
Department of Communications and Integrated Systems
Tokyo Institute of Technology

Abstract

EFFICIENT VOICE ACTIVITY DETECTION AND SPEECH ENHANCEMENT ALGORITHMS BASED ON SPECTRAL FEATURES

Yanna Ma

Department of Communication and Integrated Systems

Tokyo Institute of Technology

A novel and robust Voice Activity Detection (VAD) algorithm utilizing long-term spectral flatness measure (LSFM) and an efficient speech enhancement (SE) algorithm based on modified Wiener filtering method have been proposed in this thesis. The LSFM-based VAD improves speech detection robustness in various noisy environments by employing a low-variance spectrum estimate and an adaptive threshold. The discriminative power of the new LSFM feature is shown by conducting an analysis of the speech/non-speech LSFM distributions. Based on the analysis, we find that LSFM has the potential to be used as a robust feature for VAD. The proposed LSFM-based VAD algorithm was evaluated under twelve types of noises (eleven from NOISEX-92 and Speech shaped noise) and five types of signal-to-noise ratio (SNR) in core TIMIT TEST corpus. Comparisons with three modern standardized algorithms (ETSI AMR option 1&2 and ITU-T G.729 AnnexB) demonstrate that our proposed LSFM-based VAD scheme achieved best average accuracy rate. A long-term signal variability (LTSV)-based VAD scheme is also compared with our proposed method. The results show that our proposed algorithm outperforms it for most of the noises considered including difficult noises like Machine gun noise and Speech babble noise.

After the introduction of the LSFM-based VAD, we continue to show the proposed efficient SE algorithm. It utilized constraints to the Wiener gain function in which the wavelet thresholded multitaper spectrum was taken as the clean spectrum for the constraints. The proposed algorithm was evaluated under eight types of noises and

seven SNR levels in NOIZEUS database and was predicted by the composite measures and the SNR_{LOSS} measure to improve subjective quality and speech intelligibility in various noisy environments. Comparisons with two other algorithms (KLT and WT) demonstrate that in terms of signal distortion, overall quality and the SNR_{LOSS} measure, our proposed constrained SE algorithm outperforms the KLT and WT schemes for most conditions considered.

Publication List

Journal papers

- Yanna Ma and Akinori Nishihara. Efficient Voice Activity Detection Algorithm using Long-term Spectral Flatness Measure. EURASIP J. Audio Speech Music Process. 2013, Article 21 (July 2013), 18 pages. doi:10.1186/1687-4722-2013-21
- Yanna Ma and Akinori Nishihara. A Modified Wiener Filtering Method Combined with Wavelet Thresholding Multitaper Spectrum for Speech Enhancement. (Conditionally Accepted)

International Conference papers

- Yanna Ma and Akinori Nishihara, A Novel Voice Activity Detection Algorithm using Long-term Spectral Flatness Measure, 2013 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, Mar.2013, The Island of Hawaii, USA. (Student Paper Award)
- Yanna Ma, Hisayori Noda, Izumi Ito and Akinori Nishihara, Modified Delta Encoding and Its Applications to Speech Signal, IEEE Region 10 Conference TENCON, Nov. 2010, Fukuoka, Japan.

Domestic Conference papers

- Yanna Ma, M. T. Akhtar and Akinori Nishihara, A Robust Voice Activity Detection based on Long-term Mean Power Spectrum, Signal Processing Symposium, 2011, Nov. 2011, Sapporo, Japan.

Acknowledgments

I would like to express my gratitude to all the people who helped me during the period I pursue my PhD degree at Tokyo Institute of Technology. Without them, this dissertation would not have existed.

Firstly, I would like to express the deepest appreciation to my supervisor, Professor Akinori Nishihara, for giving me this precious opportunity to study in his lab since 2009. The free atmosphere here guaranteed me to study the topics that I am interested in and motivated me to do research independently. I believe this research experience is very precious and will greatly benefit my career in the future. I'm very grateful for his many useful and enlightening suggestions and constant support during my five-year study in Tokyo Tech.

Secondly, I want to express my sincere thanks to the assistant professor Izumi Ito and special researcher Akhtar Muhammad Tahir. The discussions with them enable me to do research in a more enjoyable way. Their suggestions and comments always enlighten me to solve the problem in a different but effective way.

Then, I would like to thank the Japan government to grant me the MEXT (Ministry of Education, Culture, Sports, Science and Technology) scholarship. This financial support enables me to concentrate on my research without worrying about the living expense.

Also, I would like to thank all the student members in Nishihara laboratory for their useful research comments on weekly seminars and their hospitality in daily life. Among them I'd like to express my special thanks to my Chinese senior, Jianxia Cao, who always support and encourage me like an elder sister whenever I feel frustrated in my study or daily life.

Of course, I am also grateful to my parents Wenxing Ma and Xiangyun Yan who don't even have a college education but know that education is the best thing for their children. In such difficulties they managed to send me to a good university in China and encouraged me to pursue more advanced education in Japan. Without them I would never have come so far in pursuing my dream.

Finally and most important of all, I am greatly indebted to my husband Qirui Li, for his constant encouragement, understanding, support, and love for the past six years. Without his sacrifice this dissertation would never have been accomplished.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Research Significance and Problem Statement

The speech signal has been studied for various reasons and applications by many researchers for many years. Among them voice activity detection (VAD) is widely used within the field of speech communication for achieving high coding efficiency and low-bit rate transmission. VAD is the method to discriminate voice activity (i.e., speech presence) and silence (i.e., speech absence) from the input noisy speech. Researchers have proposed a variety of features exploiting different properties of speech and noise to achieve a better VAD accuracy. However, the VAD in low signal-to-noise ratios (SNR) and some specific noises such as babble noise and machine gun noise still remain challenging and require the design of further robust features and algorithms.

We are living in a noisy world where the noise generated from either natural sources or human activities can be found almost everywhere: car, train, street, restaurant, etc. Those noises are captured by the microphone during voice communication and adversely affect the quality of voice communication [1]. To address this problem, noise reduction, also called speech enhancement (SE), technology is developed to remove those noise components, and produce an enhanced speech signal that sounds more pleasant to human ears.

1.2 Motivations for the Present Works

1.2.1 Motivation for VAD

The primary motivation for VAD stems from the fact that a typical speech conversation is characterized by a speech to non-speech ratio of forty to sixty. Therefore, it can be used as an important preprocessing stage to identify and compress silence in communication systems. VAD's benefits in applications such as speech coding and voice over IP (VoIP), include decreasing the average bit rate, increasing the number of users and lowering the power consumption. Furthermore, VAD can be also used to improve recognition accuracy in an automatic speech recognition (ASR) systems.

Most parametric representations of speech for the VAD problem have used time-domain features (e.g., energy, zero crossing rate) or features derived from the spectral shape (e.g., cepstral features). In general, time-domain parameters exhibit dependencies on estimates of background noise and changes in thresholds. Spectral shape, on the other hand, tends to lose its effectiveness with an increase in noise level. This would lead to a poorer discrimination between the speech and non-speech regions. Furthermore, the majority of the VAD algorithms encounter problems in low SNR conditions, particularly when the noise is nonstationary. Therefore, a more robust representation needs to be explored.

1.2.2 Motivation for SE

The objective of SE algorithms is to improve one or more perceptual aspects of the noisy speech by decreasing the background noise without affecting the intelligibility of the speech [2]. Much progress has been made in the development of SE algorithms capable of improving speech quality [3, 4] which was evaluated mainly by the objective performance criteria such as SNR [5]. However, SE algorithm that improves speech quality may not perform well in real-world listening situations where background noise level and characteristics are constantly changing [6]. The first intelligibility study done by Lim [7] in the late 1970s found no intelligibility improvement with the spectral subtraction algorithm for speech corrupted in white noise at -5 to 5 dB SNR. Thirty years later, study conducted by Hu and Loizou [2] found that none of the examined eight different algorithms improved speech intelligibility relative to unprocessed (corrupted) speech. Moreover, according to [2], the algorithms with the highest overall objective quality may not perform the best in terms of speech intelligibility (e.g.,the Log Minimum Mean Square Error (logMMSE) [8]). And the algorithm which performs the worst in terms of overall quality may perform well in terms of preserving speech intelligibility (e.g.,the generalized Karhunen-Loeve Transform (KLT) approach [9]). To our knowledge, very few speech enhancement algorithms [10, 11, 12] claimed to improve speech intelligibility by subjective tests for either normal-hearing listeners or hearing-impaired listeners. Hence, we focused in this thesis on improving performance on speech intelligibility of SE algorithm.

1.3 Objectives

1.3.1 Objective for VAD

The primary objective of this work is to design a new robust VAD scheme. The goals of this part are:

- To find a new robust feature to discriminate speech from noisy signals in

high efficiency and design the whole system for VAD.

- To analyze the robustness of the proposed feature.
- To test the new proposed system on different noises and compare the results with other advanced VAD schemes.

1.3.2 Objective for SE

As for SE algorithm, we intend to achieve the following goals:

- To understand the reasons why most existing SE algorithms can't improve the performance on speech intelligibility.
- To utilize the useful finding of [5] about the different perceptual effects of attenuation and amplification distortion on speech intelligibility.
- To evaluate the performance of the proposed SE algorithm on improving subjective quality and speech intelligibility through comparing with other effective SE algorithms.

1.4 Thesis Organization

This thesis is organized as follows. Chapter 2 gives a comprehensive review of the stat-of-the-art VAD and SE algorithms. The proposed VAD algorithm utilizing logn-term spectral flatness measure (LSFM) is presented in Chapter 3. In Chapter 4, the SE method imposing constraints to the Wiener gain function is described. Chapter 5 gives a summary of our contributions.

Chapter 2

Literature Review

Contents

2.1 Three Modern VAD Standards

For the performance comparison with the proposed method, three modern standardized VAD schemes are introduced here. They are ITU G.729 Annex B VAD, ETSI AMR VAD Option1 and Option2.

2.1.1 ITU G.729 AnnexB VAD

International Telecommunication Union (ITU) recommendation G.729 Annex B (G.729B) VAD [13] makes a VAD decision every 10 ms. The block diagram of the algorithm is shown in Figure 2.1. The sampling frequency for the test speech is 8 kHz.

G729B VAD uses a piecewise linear discriminant based on the full band energy E_f, the low band energy E_l, the zero-crossing rate ZC and line spectral frequencies LSF_i. LSF_i is derived from the set of linear prediction coefficients which is derived from the autocorrelation, the details are described in Section 3.2 of [14]. The long-term averages of the parameters during non-active voice segments follow the change nature of the background noise. A set of differential parameters is obtained at each frame. These are a difference measure between each parameter and its respective long-term average. These four differential parameters are defined as follows [15]:

- Spectral Distortion ΔLSF:

$$\Delta LSF = \sum_{i=1}^{p} (LSF_i - \overline{LSF_i})^2 \tag{2.1}$$

- Full band energy difference:

$$\Delta E_f = \overline{E_f} - E_f \tag{2.2}$$

- Low band energy difference:

$$\Delta E_l = \overline{E_l} - E_l \tag{2.3}$$

- Zero crossing difference:

$$\Delta ZC = \overline{ZC} - ZC \tag{2.4}$$

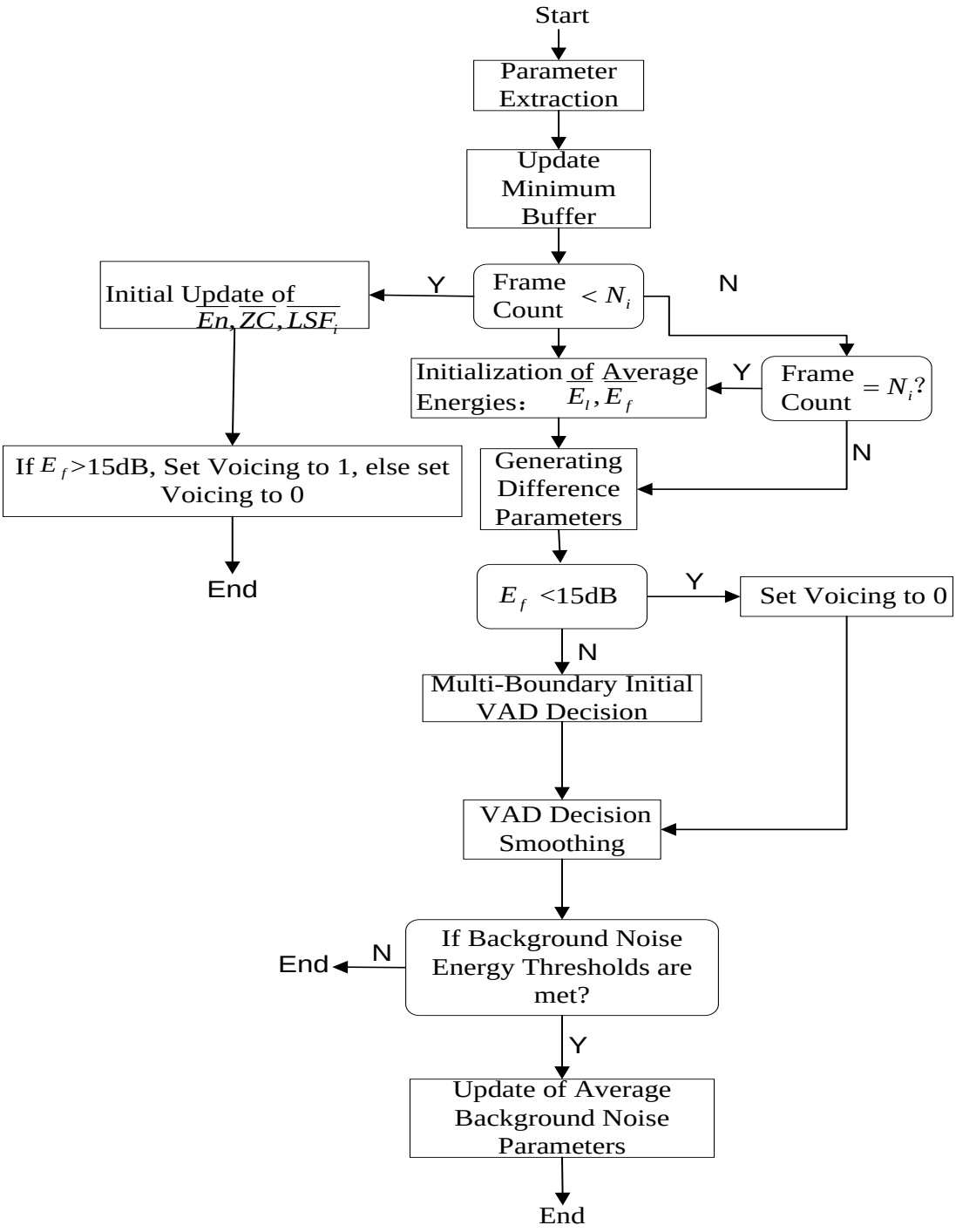


Figure 2.1: Block diagram of ITU G.729 Annex B VAD algorithm.

where LSF_i, E_f, E_l and ZC are line spectral frequencies (LSF), current frame full band energy, current frame low band energy and the zero crossing rate respectively.

Furthermore, $\overline{LSF_i}$, $\overline{E_f}$, $\overline{E_l}$ and $\overline{ZC}$ are the running averages of the corresponding parameters of the background noise. The first N_i frames are used to initialize the running averages of the background noise characteristics. The running averages have to be updated after the VAD decision when only the background noise is present .

The initial decision is made by using multi-boundary decision regions in the space of the four differential parameters. A final detection is obtained by smoothing the initial decision using energy consideration and neighboring past decisions.

2.1.2 ETSI AMR VADs

In this subsection, we will introduce European Telecommunications Standards Institute (ETSI) recommendations AMR VAD option1 (AMR1) and option2 (AMR2) [16]. They are spectral shape based VAD methods which use sub-band and channel energies, respectively, to make the VAD decisions in conjunction with an extensive hangover scheme. They make a decision every 20ms. The sampling frequency for the test speech of their system is 8 kHz. A brief description of the two AMR VAD options is provided herein.

2.1.2.1 ETSI AMR VAD Option 1

The block diagram of the AMR1 algorithm is shown in Figure 2.2. The VAD algorithm uses parameters of the speech encoder to compute the VAD decisions.

The input signal is divided into frequency bands using a 9-band filter bank. Cut-off frequencies for the filter bank are shown in Table 2.1. The input frames are divided into sub-bands and level of the signal in each band is calculated. The energy of the input speech signal from nine frequency sub-bands is computed. The bandwidths of these sub-bands are non-uniform in nature, with the lower

frequency sub-bands having smaller bandwidths. Open-loop lags, which are calculated by open-loop pitch analysis of the speech encoder, are input for the pitch detection function. A pitch flag is then computed for the indication of

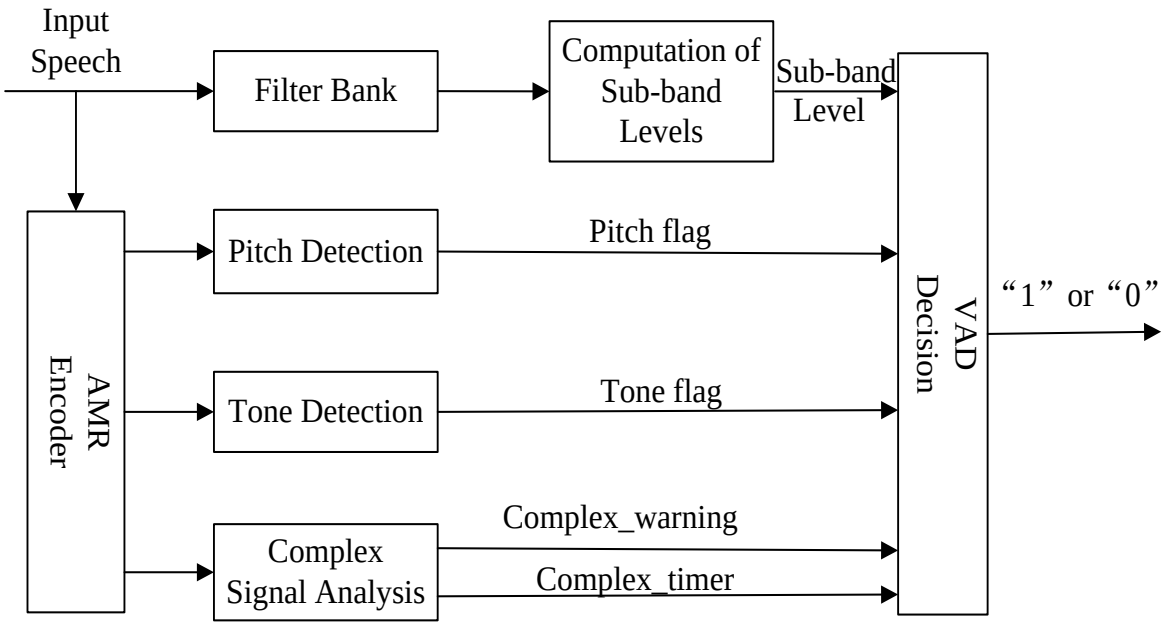


Figure 2.2: Block diagram of the AMR1 VAD algorithm.

the presence of pitch. The purpose of the pitch detection function is to detect vowel sounds and other periodic signals. Tone detection function calculates a tone flag which indicates the presence of an information tone, since the pitch detection function cannot always detect these signals. Tones are detected based on pitch gain of the open-loop pitch analysis. The pitch gain is estimated using autocorrelation values received from the pitch analysis.

Complex Signal Detection Function calculates the complex warning which indicates the presence of a correlated complex signal such as music. Correlated complex signals are detected based on analysis of the correlation vector available in the open-loop pitch analysis. An estimate of the background noise energy is needed for the VAD decision function. Intermediate VAD decision is calculated

Table 2.1: Cut-off frequencies for the filter bank

Band number	Frequencies
1	0-250 Hz
2	250-500 Hz
3	500-750 Hz
4	750-1000 Hz
5	1000-1500 Hz
6	1500-2000 Hz
7	2000-2500 Hz
8	2500-3000 Hz
9	3000-4000 Hz

based on the comparison of the background noise estimate and levels of the input frame. Finally, the VAD decision is smoothed by adding hangover scheme to the intermediate VAD decision.

2.1.2.2 ETSI AMR VAD Option 2

The block diagram of the AMR2 algorithm is shown in Figure 2.3. The input signal is pre-emphasized and windowed by a rectangular window. Frequency domain conversion is performed using the Discrete Fourier Transform (DFT). Channel energy and SNR are then estimated. The spectral deviation estimator is used as a safeguard against erroneous updates of the background noise estimate. If the spectral deviation of the input signal is too high, then the background noises estimate update may not be permitted.

AMR2 also estimates sub-band SNRs. However, the number of sub-bands is sixteen while in AMR1 it is nine. Also the nonlinear scale used in band grouping is different from that of AMR1. The estimation of the background noise energy in each sub-band is similar to that of AMR1, i.e., using a first order auto-regressive model. AMR2 also makes VAD decisions using an adaptive threshold. Non-stationary noise is handled by measuring the variance of the

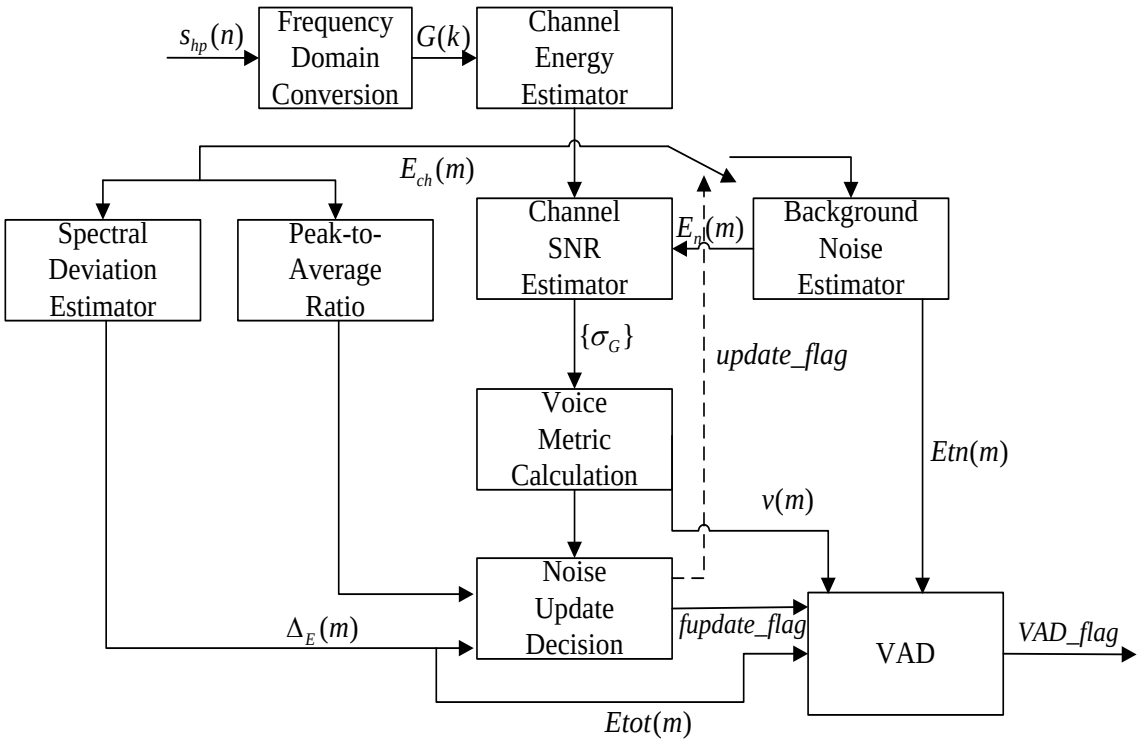


Figure 2.3: Block diagram of the AMR2 VAD algorithm.

instantaneous SNRs, estimated every frame. Finally, the decisions are smoothed by a hangover scheme. This scheme is based on peak-to-average SNR ratio.

2.1.3 Statistical model-based VAD

In 1998, Sohn and Sung [17] proposed an robust VAD algorithm that uses a novel noise spectrum adaption employing soft decision techniques. The decision rule was derived from the generalized likelihood ratio test by assuming that the noise statistics are known a priori. An enhanced version [18] of the original VAD for the application to variable-rate speech coding was developed by employing the decision-directed parameter estimation method for the likelihood ratio test. And an effective hang-over scheme which considers the previous observations of

a first-order Markov process modeling speech occurrences was also proposed in [18]. The algorithm outperformed the G.729B VAD in terms of speech detection and false-alarm probabilities in various environmental conditions.

2.1.4 Long-term Signal Variability (LTSV) Scheme

For a better evaluation of our proposed VAD algorithm, except the comparison with three modern standards, we also compared it with a long-term signal variability (LTSV) based VAD algorithm proposed by P. K. Ghosh et al [19]. Figure 2.4 illustrates the block diagram of the LTSV based VAD algorithm which calculates the sample variance of the entropy measure on the normalized short-time spectrum. The details are as follows:

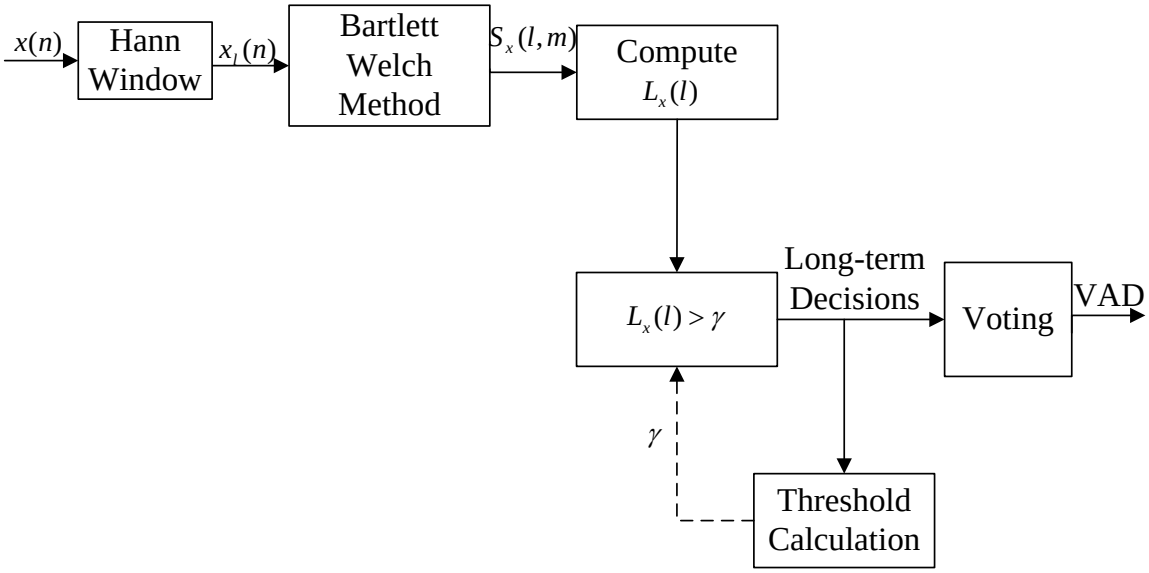


Figure 2.4: Block diagram of the LTSV based VAD algorithm.

First, to alleviate the discontinuities and reduce the spread of the spectral energy into the side lobes of the spectrum, the input signal $x(n)$ is Hann windowed. Then, the windowed signal $x_l(n)$ is used to estimate the power spectrum by Bartlett-Welch method. The entropy measure on the normalized short-time

spectrum is then computed. The LTSV measure at the lth window, $L_x(l)$, is the sample variance of the entropy measure which describes the degree of non-stationarity of the signal. Because a fixed threshold does not work for all noises, an adaptive threshold γ was designed to meet this requirement. Finally, the long windows voting scheme is used to make the final VAD decisions.

2.2 Speech Enhancement algorithms

SE is concerned with improving some perceptual aspect of speech that has been degraded by additive noise. SE techniques have a broad range of applications, from hearing aids to mobile communication, voice-controlled systems, multi-party teleconferencing, and automatic speech recognition (ASR) systems. In most applications, the aim of speech enhancement is to improve the quality and intelligibility of degraded speech.

The algorithms can be summarized into four classes: spectral subtractive, subspace, statistical model based and Wiener type algorithms.

2.2.1 Spectral-Subtraction

The spectral subtractive algorithm is historically one of the first algorithms proposed for noise reduction [4]. It is based on a simple principle. Assuming additive noise, one can obtain an estimate of the clean signal spectrum by subtracting an estimate of the noise spectrum from the noisy speech spectrum. The noise spectrum can be estimated, and updated, during periods when the signal is absent. The assumption made is that noise is stationary or a slowly varying process, and that the noise spectrum does not change significantly between the updating periods. The enhanced signal is obtained by computing the inverse discrete Fourier transform of the estimated signal spectrum using the phase of the noisy signal. The algorithm is computationally simple as it only involves a forward and an inverse Fourier transform.

The subtraction process needs to be done carefully to avoid any speech distortion. If too much is subtracted, then some speech information might be

removed, whereas if too little is subtracted, then much of the interfering noise remains. Many methods have been proposed to alleviate or eliminate most of the speech distortion introduced by the spectral subtraction process. Please refer [20] for the details of the basic spectral subtraction algorithm.

2.2.2 Subspace method: Karhunen-Loeve transform (KLT)

The idea behind subspace methods is to project the noisy signal into two subspaces: the signal subspace and the noise subspace. The clean signal can be estimated by removing the components of the signal in the noise subspace because it contains signals from the noise process only. There are two methods to decompose the space into two subspaces: singular value decomposition (SVD) [21] or the eigenvalue decomposition (EVD) [22, 23].

The KLT based subspace approach proposed by Ephraim and Van Trees [23] is undoubtedly one of the most important work in speech enhancement area. By assuming that the additive noise is wide-band and that the speech signal only occupies a portion of the vector space, they sought for an optimal estimator that would minimize the speech distortion subject to the constraint that the residual noise fell below a preset threshold. Using the EVD of the covariance matrix, Ephraim and Van Trees showed that the decomposition of the vector space of the noisy signal into a signal and noise subspace can be obtained by applying the theorem Karhunen-Loeve transform (KLT) to the noisy signal.

A generalization of the Ephraim and Van Trees approach for white noise was derived in [9]. It is a generalized subspace approach with built-in prewhitening for enhancing speech corrupted with colored noise. A nonunitary transform based on the simultaneous diagonalization of the clean speech and noise covariance matrices is used in this approach to project the noisy signal into a signal subspace and a noise subspace. Two estimators were derived based on this nonunitary transform, one based on time-domain constraints and one based on spectral domain constraints. This KLT algorithm was proved in [2] and [10] by subjective tests to perform well in terms of preserving speech intelligibility for normal hearing listeners, and improving speech intelligibility significantly

for cochlear implant users in regards to recognition of sentences corrupted by stationary noises, respectively.

2.2.3 Statistical-model based methods

Statistical-model-based methods focus on nonlinear estimators of the magnitude using various statistical models and optimization criteria. These nonlinear estimators take the probability density function (PDF) of the noise and the speech discrete Fourier transform (DFT) coefficients into account and are often combined with soft-decision gain modifications that take the probability of speech presence into account. Several statistical-model-based methods including maximum-likelihood estimator, Minimum Mean Square Error (MMSE) magnitude estimator and log-MMSE estimator for optimal spectral magnitude estimation were discussed in [4].

2.2.4 Wiener-type

The optimal filter that minimizes the estimation error is caller the Wiener filter, named after the mathematician Norbert Wiener [24], who first formulated and solved this filtering problem in the continuous domain. The Wiener filtering approach derives the enhanced signal by optimizing a mathematically tractable error criterion: the mean-square error. More specifically, the Wiener filters were derived by minimizing the speech distortion subject to the noise distortion falling below a given threshold level (e.g., masking threshold).

According to [4], the Wiener filters are considered to be linear estimators of the clean signal spectrum, and they are optimal in the mean-square sense. These filters are constrained to be linear which means that the enhanced time domain signal is obtained by convolving the noisy speech signal with a linear (Wiener) filter. Equivalently, in the frequency domain the enhanced spectrum is obtained by multiplying the input noisy spectrum by the Wiener filter. The linear estimators, however, are not necessarily the best estimators of the clean signal spectrum. Nonlinear estimators of the clean signal spectrum could po-

tentially yield better performance.

The Wiener filter can also be expressed as a function of the ratio of the clean signal power spectrum to the noise power spectrum, i.e., the *a priori* SNR. Several algorithms have attempted to estimate the *a priori* SNR, rather than the clean signal power spectrum [25, 26, 27].

2.3 Speech and Noise Database

2.3.1 VAD: TIMIT corpus and NOISEX-92 database

In this subsection, the test speech corpus and noise database used in our research are introduced. They are the DAPRA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) [28] and the NOISEX 92 noise database [29].

2.3.1.1 TIMIT corpus

The TIMIT corpus is designed for the development and evaluation of automatic speech recognition systems. The sampling frequency of the corpus is 16 kHz. TIMIT corpus contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. It includes the training corpus and test corpus. In the training corpus, there are 462 speakers with 4620 sentences while there are 168 speakers with 1680 sentences in the test corpus. The components of the TIMIT corpus are described by the Table 2.2.

2.3.1.2 NOISEX 92 database

The noises used in this research are from the NOISEX 92 database. In total there are 11 types of noise which include the followings:

- Stationary noise: White and Pink.
- Non-stationary noise: Tank, Military vehicle, Jet cockpit, HFchannel, F16, Factory, Car Interior, Machinegun, Babble.

Table 2.2: The components in each file of TRAIN and TEST

File name	TRAIN Number	TEST Number
DR1	38	11
DR2	76	26
DR3	76	26
DR4	68	32
DR5	70	28
DR6	35	11
DR7	77	23
DR8	22	11

2.3.2 SE: NOIZEUS corpus and AURORA database

2.3.2.1 NOIZEUS corpus

NOIZEUS [30] was s a noisy speech corpus developed by [4] to facilitate comparison of SE algorithms among different research groups [6]. The noisy database contains thirty IEEE sentences [31] which were recorded in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The sentences were produced by three male and three female speakers (5 sentences/speaker). The IEEE database was used as it contains phonetically-balanced sentences with relatively low word-context predictability. The thirty sentences were selected from the IEEE database so as to include all phonemes in the American English language. The sentences were originally sampled at 25 kHz and downsampled to 8 kHz. The details about the thirty sentences were given in Table 2.3.

2.3.2.2 AURORA database

The AURORA database [32] include the following recordings from eight different places: Train, Babble (crowd of people), Car, Exhibition Hall, Restaurant, Street, Airport and Train Station.

2.4 Conclusions

In this chapter, for the comparison with our proposed VAD scheme, a brief review of the three standard VAD schemes, namely, G.729B, AMR1 and AMR2 were provided. Also, the newly proposed LTSV based VAD is presented in this chapter. Four kinds of SE algorithms including spectral-subtractive, subspace, statistical-model based and Wiener-type methods were also introduced briefly. Finally, the test speech corpus and noise database for VAD and SE algorithms are also described. Although significant progress has been made in the above mentioned methods, each of them has its own drawbacks and limitations. In the next two chapters, we will derive our own VAD and SE algorithms.

Table 2.3: List of sentences used in NOIZEUS.

Filename	Speaker	Gender	Sentence text
sp01	CH	M	The birch canoe slid on the smooth planks.
sp02	CH	M	He knew the skill of the great young actress.
sp03	CH	M	Her purse was full of useless trash.
sp04	CH	M	Read verse out loud for pleasure.
sp05	CH	M	Wipe the grease off his dirty face.
sp06	DE	M	Men strive but seldom get rich.
sp07	DE	M	We find joy in the simplest things.
sp08	DE	M	Hedge apples may stain your hands green.
sp09	DE	M	Hurdle the pit with the aid of a long pole.
sp10	DE	M	The sky that morning was clear and bright blue.
sp11	JE	F	He wrote down a long list of items.
sp12	JE	F	The drip of the rain made a pleasant sound.
sp13	JE	F	Smoke poured out of every crack.
sp14	JE	F	Hats are worn to tea and not to dinner.
sp15	JE	F	The clothes dried on a thin wooden rack.
sp16	KI	F	The stray cat gave birth to kittens.
sp17	KI	F	The lazy cow lay in the cool grass.
sp18	KI	F	The friendly gang left the drug store.
sp19	KI	F	We talked of the sideshow in the circus.
sp20	KI	F	The set of china hit the floor with a crash.
sp21	SI	M	Clams are small, round, soft and tasty.
sp22	SI	M	The line where the edges join was clean.
sp23	SI	M	Stop whistling and watch the boys march.
sp24	SI	M	A cruise in warm waters in a sleek yacht is fun.
sp25	SI	M	A good book informs of what we ought to know.
sp26	TI	F	She has a smart way of wearing clothes.
sp27	TI	F	Bring your best compass to the third class .
sp28	TI	F	The club rented the rink for the fifth night.
sp29	TI	F	The flint sputtered and lit a pine torch.
sp30	TI	F	Let's all join as we sing the last chorus.

All the sentences were used in the evaluation.

Chapter 3

Efficient Voice Activity Detection Algorithm using Long-term Spectral Flatness

Contents

3.1 Introduction

Voice activity detection (VAD) is a method to discriminate speech segments from input noisy speech. It is an integral part to many speech and audio processing applications and is widely used within the field of speech communication for achieving high coding efficiency and low bit rate transmission. Examples include noise reduction for digital hearing aid devices [33], mobile communication services [34], voice recognition systems [35], compression [36], and speech coding [37].

A typical VAD system consists of two core parts: feature extraction and speech/non-speech decision mechanism. Researchers have proposed a variety of features exploiting different properties of speech and noise to achieve better VAD performance. In early VAD algorithms, short-term energy [38] and zero-crossing rate [39] were widely used features because of their simplicity. However, the performance degrades easily when faced with low signal-to-noise ratio (SNR) or non-stationary background noise. To solve this problem, robust acoustic features such as spectrum [40], autocorrelation [41], power in the band-limited region [42], and higher-order statistics [43] have been proposed for VAD. Most of those methods assume the background noise to be stationary during a certain period; thus, they are sensitive to changes in SNR of the observed signal. Some works [44, 45] proposed noise estimation and adaptation for improving VAD robustness, but those methods are computationally expensive. Most of those features mentioned work sufficiently well in stationary noise and higher than 10-dB SNR cases. When facing with lower SNR cases or when the background noise contains complex audible events appearing occasionally, such as babble noise in a cafeteria and machinery noise in a factory, there will be cases when most of the speech spectrum is corrupted, which destroys the overall statistical as well as structural properties of the speech signal [46]. In general, VAD algorithms based on a particular feature or specific set of features are still far from efficient especially when they are operating in adverse acoustic conditions. Therefore, the VAD algorithm in low SNRs and some specific noises

such as speech babble noise and machine gun noise still remains challenging and requires the design of further robust features and algorithms.

All VAD features mentioned are extracted from the short-term analysis frames (usually 20 to 40 ms), and decisions are made at each frame. In contrast with the use of frame level features, Ramirez et al. [44] proposed the use of a long-term spectral divergence (LTSD) feature to discriminate speech from noise. The speech/non-speech decision rule was formulated by comparing the long-term spectral envelope to the average noise spectrum. A high discriminating decision rule was achieved and the average number of decision errors were minimized. This LTSD-based algorithm requires average noise spectrum magnitude information which is not accurately available in practice. Moreover, Ghosh et al. [19] proposed a long-term signal variability (LTSV)-based VAD which uses a very long window to estimate the averaged spectrogram as well as for computing long-term entropies of each frequency band. This LTSV-based VAD yields a great improvement for SNRs smaller than 5 dB but becomes saturated when SNRs are higher than 5 dB.

Spectral flatness is a measure of the width, uniformity, and noisiness of the power spectrum. A high spectral flatness indicates that the spectrum has a similar amount of power in all spectral bands, and the graph of the spectrum would appear relatively flat and smooth; this would sound similar to white noise. A low spectral flatness indicates that the spectral power is less uniform in frequency structure, and this would typically sound like speech. Therefore, the analysis over a long window for exploiting the spectral flatness of the signal will be beneficial for distinguishing speech from noise. In this chapter, we propose a novel VAD algorithm based on long-term spectral flatness measure (LSFM). The discriminative power of the proposed LSFM feature will be verified by researching the distribution of LSFM measure for speech and non-speech in terms of their misclassification rate for various noises. We have experimentally evaluated its performance under a variety of noise types and SNR conditions.

The structure of the rest of this chapter is arranged as follows. Section 3.2 discusses the LSFM feature and its discriminative ability. Section 3.3 presents

the proposed LSFM-based VAD algorithm including the choice for proper parameters and the design of an adaptive threshold. Section 3.4 contains the speech and noise database and metrics used in the evaluation. Section 3.5 provides the experimental results. Finally, a conclusion of this work and the discussion are given in Section 3.6.

3.2 Long-term Spectral Flatness Measure and Its Discriminative Power

Speech is a highly non-stationary signal, while background noise can be considered to be stationary over relatively long periods. The rationale behind the LSFM feature is that the observed signal spectrum evinces more structure when the signal of interests is present compared to when it is absent. This increase in the structure of the signal may be characterized by a reduction in the flatness of the magnitude spectrum of the short-time Fourier representation of the signal [47].

3.2.1 Long-term Spectral Flatness Measure

The LSFM feature is computed using the spectra of the last R frames of the input signal $x(n)$. The LSFM feature, $L_x(m)$, at the mth frame and across all the chosen frequency is then calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum. To expand the dynamic range, it is measured on a logarithmic scale, ranging from zero to minus infinity as:

$$L_x(m) = \sum_k \log_{10} \frac{\mathrm{GM}(m, \omega_k)}{\mathrm{AM}(m, \omega_k)}, \tag{3.1}$$

where the geometric mean $\mathrm{GM}(m,\omega_k)$ and arithmetic mean $\mathrm{AM}(m,\omega_k)$ of the power spectrum is calculated as:

$$\mathrm{GM}(m,\omega_k) = \sqrt[R]{\prod_{n=m-R+1}^{m} S(n,\omega_k)}, \tag{3.2}$$

$$\mathrm{AM}(m,\omega_k) = \frac{1}{R}\sum_{n=m-R+1}^{m} S(n,\omega_k). \tag{3.3}$$

The short-time spectrum $S(n,\omega_k)$ used in this research is estimated using the Welch-Bartlett method which averages the spectral estimates of M consecutive frames. The expressions are

$$S(n,\omega_k) = \frac{1}{M}\sum_{p=n-M+1}^{n} |X(p,\omega_k)|^2, \tag{3.4}$$

$$X(p,\omega_k) = \sum_{l=(p-1)N_{\mathrm{sh}}+1}^{N_w+(p-1)N_{\mathrm{sh}}} w(l-(p-1)N_{\mathrm{sh}}-1)x(l)e^{-j\omega_k l}, \tag{3.5}$$

where $X(p,\omega_k)$ is the short-time Fourier transform coefficient at frequency ω_k of the pth frame. $w(i)$ is the short-time Hann window, where $i \in [0, N_w)$. N_w is the frame length and N_{sh} is the frame shift duration in terms of samples.

According to AM-GM inequality, the geometric mean, $\mathrm{GM}(m,\omega_k)$, is smaller than or equal to the arithmetic mean, $\mathrm{AM}(m,\omega_k)$, with equality being achieved if and only if all $S(n,\omega_k)$ are the same. Therefore, from Eq. (3.1) we can conclude that the LSFM feature, $L_x(m)$, is in the range $(-\infty, 0]$ with the maximum value acquired when the geometric mean is equal to the arithmetic mean.

3.2.2 The LSFM Feature Distributions of Speech and Non-Speech

In this subsection, the distributions of the LSFM feature are investigated in order to clarify the motivation for utilizing the proposed LSFM feature as a VAD algorithm and demonstrate the discriminative power of this feature.

The test set consisting of 16 individual speakers (8 male, 8 female), each speaking 10 phonetically balanced English sentences, is randomly chosen from the TIMIT training corpus [28]. The LSFM feature values were computed at every frame from noisy speech. The LSFM measure, $L_x(m)$, is considered to be $L_{S+N}(m)$ if there are speech samples between $(m-R+1)$th and mth frame. Otherwise, it is decided to be $L_N(m)$ which contains only noise information. The overlap area of the two distributions (L_{S+N} and L_N) is considered to be the error caused by misclassification. The lower the misclassification rate is, the better the separation. The sampling frequency of the test signal is 16 kHz, and the Hann window has a length of 20 ms and 10-ms shift. M is fixed to be 10, and $\{\omega_k\}$ is uniformly distributed between the frequency range 500 Hz to 4 kHz. The total misclassification error among these realizations of L_{S+N} and L_N was computed by comparing with the phonetic level transcription [28] of the TIMIT training corpus.

First, the distributions of the LSFM feature as a function of the long-term window length (R=2, 5, 10, 20, 30, and 40) for white noise at five SNR levels (-10, -5, 0, 5, and 10 dB) were studied. The results are shown in Figures 3.1, 3.2, 3.3, 3.4, and 3.5. The total misclassification error (Error), accuracy rate (Correct), speech detection error (SE), and non-speech detection error (NE) are displayed on the upper or lower right of each subfigure. The total misclassification error was reduced by 61.4% (-10 dB), 74.5% (-5 dB), 79.0% (0 dB), 84.3% (5 dB), and 82.9% (10 dB) when the window length R was increased from 2 to 30 frames. The percentage is the ratio between the reduced misclassification error (when R was changed from 2 to 30 frames) and the misclassification error when R was 2.

The distributions of the LSFM feature for all 12 kinds of noises at 0-dB SNR were investigated. M is fixed to be 10, and R is chosen to be 30. The discriminative power of this LSFM feature can be measured by the separateness of its distribution for speech and non-speech. As shown in Figures 3.6, 3.7, and 3.8, there is overlap between the histograms of $\log_{10}(L_{S+N})$ and $\log_{10}(L_N)$. We calculated the total misclassification error which is the sum of the speech detection

error and non-speech detection error. From the figures, we can conclude that for most noises considered (9 out of 12 kinds of noises), the proposed LSFM feature resulted in a misclassification error smaller than 10%: white (7.86%), pink (7.75%), tank (7.47%), military vehicle (7.75%), jet cockpit (9.32%), HF channel (8.30%), F-16 cockpit (8.89%), car interior (8.14%), and speech shaped (7.84%). For factory floor (25.86%), machine gun (45.42%), and speech babble (24.08%), the misclassification errors were comparatively high. The factory floor is that of cutting noise, that of the machine gun is impulsive in nature, and that of speech babble is speechlike. One possible reason for the poor performance is the mismatch of M and R.

3.3 The Proposed LSFM-based VAD Algorithm

The proposed VAD algorithm assumes that the signal spectrum is more organized during speech segments than during noise segments [48]. It adopts the average spectrum over a long-term window instead of instantaneous values of the spectrum. Typically, a periodogram is commonly employed for spectrum estimation, but it is well known that the periodogram is an inconsistent spectral estimator. According to [40], the Welch-Bartlett method [49] was found to give a good trade-off between variance reduction and spectral resolution reduction. Therefore, in our proposed algorithm, the signal spectrum is estimated using the Welch-Bartlett method.

A block diagram of the proposed LSFM-based VAD algorithm is shown in Figure 3.9. The algorithm can be described as follows. The input speech signal is decomposed into frames of 20 ms in length with an overlap of 10 ms by the Hann window. The spectrum of the segmented signal is estimated using the Welch-Bartlett method. At the mth frame, the LSFM feature $L_x(m)$ is computed using the previous R frames. The initial decision about whether there contains speech in the last R frames is made through the comparison with an adaptive threshold. The initial decision is denoted by V_{INL}. If there is a speech frame existing over the previous R frames ending at the mth frame,

$V_{\mathrm{INL}}(m) = 1$; otherwise, $V_{\mathrm{INL}}(m) = 0$ and there are only non-speech frames over the previous R frames. We adopt the voting scheme proposed by Ghosh et al. [19] to make the final VAD decisions on a 10-ms interval. First, the initial decisions, $V_{\mathrm{INL}}(m)$, $V_{\mathrm{INL}}(m+1)$, ..., $V_{\mathrm{INL}}(m+R-1)$, are collected for those long windows which overlap with the target 10-ms interval. Then, the target 10-ms interval is marked to be speech if there is 80% or above of those initial decisions that contain speech; otherwise, it is marked as non-speech. The 80% was gotten empirically, which provided the maximum VAD accuracy for most noises tested over five SNR levels.

In general, speech is a low-pass signal, and the frequency range of 500 Hz to 4 kHz is crucial for speech intelligibility [50]. Hence, for a better discrimination, the start bin, k_s, and the end frequency bin, k_e, are calculated by:

$$k_s = N_{\mathrm{DFT}}(\frac{500}{f_s}), \tag{3.6}$$

$$k_e = N_{\mathrm{DFT}}(\frac{4,000}{f_s}), \tag{3.7}$$

in which f_s is the sampling frequency and N_{DFT} is the order of Discrete Fourier Transform (DFT) which is used to calculate the spectral estimate of the observed signal. In our experiment, $f_s = 16$ kHz and $N_{\mathrm{DFT}} = 512$. The frequencies, ω_k , are uniformly distributed between 500 Hz and 4 kHz.

An illustrative example of the VAD output is shown in Figure 3.10. A high spectral flatness indicates that the spectrum has a similar amount of power in all spectral bands, which would sound similar to white noise, and the graph of the spectrum would appear relatively flat and smooth. A low spectral flatness indicates that the spectral power is concentrated in a relatively small number of bands; this means that the spectrum is more organized, and the graph of the spectrum would appear 'spiky'. Hence, the spectral flatness measure is a good feature for VAD.

3.3.1 Selection of M and R

M and R are parameters used for computing the LSFM feature L_x. We want to choose the appropriate M and R so that the separateness of the distribution for noise and speech is maximized since the more it is separated, the better the final VAD decision. The total misclassification errors (sum of speech detection error and non-speech detection error) for all combinations of M (1, 5, 10, 20, 30, and 40) and R (5, 10, 20, 30, and 40) are computed over 12 types of noise for five SNR levels (-10, -5, 0, 5, and 10 dB). The test speech set is the same with the one we used for the demonstration of the discriminative power of the proposed LSFM feature in Section 2.2.

The total misclassification error as a function of different combinations of M and R is shown in Figures 3.11, 3.12, 3.13, and 3.14. The best combination of M and R for each noise at each SNR level is written on the upper or lower right of each subfigure. After the summed up frequency of each M and R that appeared in the subfigures, we conclude that (10, 30) is the optimal combination which appeared most frequently ($M = 10$ appeared 25 times out of the 60 subfigures in total; $R = 30$ also appeared 25 times out of 60 subfigures in total). This fixed combination (10, 30) is then adopted for all the following tests.

Furthermore, from Figures 3.11, 3.12, 3.13, and 3.14, we observe that for the same R value, if M is increased from 1 to 10, the total misclassification error is decreased for most cases tested. However, when M is larger than 10, even if M is increased, the total misclassification error stops decreasing any further. This observation verified the choice of 10 to be the optimal value of M.

It is also worth mentioning that for those noise types and SNR levels whose optimal M and R combination is not (10, 30), the fixed combination (10, 30) still works well. Table 3.1 shows the points that the total misclassification error of adopting the fixed combination (10, 30) is worse (higher error value) than utilizing the best combination of M and R for each noise type and SNR level shown in the subfigures. Except for cutting factory floor noise and impulsive machine gun noise, the differences are all less than five points.

Table 3.1: The total misclassification error difference between adopting the fixed combination (10, 30) and utilizing the best (M, R) combination

Noise type	−10 dB	−5 dB	0 dB	5 dB	10 dB
White	2.85	0.69	0	0	0.87
Pink	4.49	0.64	0	0	1.19
Tank	2.32	0	1.59	2.20	2.17
Military vehicle	2.36	2.59	2.62	2.82	3.31
Jet cockpit	0	0	0	0	1.36
HF channel	1.15	0	0.57	1.02	2.27
F-16 cockpit	0	0	0.10	1.94	2.30
Factory floor	6.04	7.35	7.59	7.04	6.25
Car interior	2.33	2.82	3.16	2.33	2.21
Machine gun	12.45	11.33	9.83	8.74	7.56
Speech babble	2.11	0	2.068	0	0
Speech shaped	3.80	0	0	0	1.84

The numbers are the points that worse (higher error value) than utilizing the best (M, R) combination for each noise type and SNR level shown in the subfigures.

For machine gun noise, the optimal choice of R is 5 for all SNR levels. Machine gun noise is an impulsive noise which consists of two types of sounds, namely gunshot and silence between gunshots [19]. When R is 30, the long analysis window would include both types of sounds. Therefore, the spectral power over these 30 frames will be less uniform; the LSFM feature value will then be small, and there will be more classification errors compared to the case when R is 5. Similarly, for factory floor noise, the optimal choice of M is 1 for all SNR levels. Factory floor noise [29] was recorded near plate-cutting and electrical welding equipment which shows a repetitive pattern. According to [49], the variance of estimated power spectrum will not be obviously reduced if the overlapped frames are highly correlated with each other. Therefore, averaging over M overlapped frames will cause more misclassification errors compared to the case when M is 1.

3.3.2 Adaptive threshold

Unlike M and R, a fixed threshold would lose its efficiency when facing varying acoustic environments. Therefore, it is more suitable to design an adaptive threshold [51]. From Equations 3.2 to 3.4, we can conclude that $(R + M - 1)$ frame (0.39 s for fixed $R = 30$ and $M = 10$) information is needed to acquire the first LSFM feature value. In our implementations, the initial 1.39 s of the test signal $x(n)$ is always assumed to be non-speech. From this 1.39 s of $x(n)$, 100 realizations of L_N can be collected and saved to ψ_{INL}.

The threshold is initialized to be

$$\text{THR}_{\text{INL}} = \min(\psi_{\text{INL}}). \tag{3.8}$$

To update the threshold at the mth frame, we used two buffers ψ_{S+N} and ψ_N. ψ_{S+N} stores the LSFM measures of the last 100 long window ending at the mth frame which was decided as containing speech; similarly, ψ_N stores the LSFM measures of the last 100 long window ending at the mth frame which was decided as including non-speech information only. The adaptive threshold for the mth frame is then updated as:

$$\text{THR}(m) = \lambda \times \min(\psi_N) + (1 - \lambda) \times \max(\psi_{S+N}), \tag{3.9}$$

where λ is the parameter of the convex combination. We experimentally found that $\lambda = 0.55$ results in the maximum accuracy rate in VAD decisions over the TIMIT training set.

3.4 Evaluation setup

The proposed VAD algorithm was trained and tested using a speech database that is phonetically balanced. The system was evaluated using the error rate and accuracy rate metrics.

3.4.1 Database description

For the evaluation of VAD algorithms, TIMIT corpus is preferred since it provides manual transcription down to word and phoneme levels. The reference labels are computed using the start and end times of the utterance obtained from the TIMIT transcription (.phn files). Some experiments are carried out on the core TIMIT test set consisting of 24 individual speakers (16 males, 8 females) of eight different dialects, each speaking 10 phonetically balanced English sentences. The utterances of TIMIT corpus are short (about 3.5 s), and around 90% of which are speech; this may introduce a bias when comparing the distributions of speech and non-speech. To reduce this effect and make it closer to real-world scenarios, 2-s silence was added before and after each utterance to simulate a typical telephone conversation [52, 40, 19] in which the ratio of speech to non-speech is almost 40% to 60%.

The noise of 11 categories taken from the NOISEX-92 database [29] and speech-shaped noise are added at five different SNR levels (−10, −5, 0, 5, and 10 dB) to the signal concatenated by all 240 sentences. The noise samples from the NOISEX-92 database are resampled to 16 kHz according to the experiment requirement. Among the 12 kinds of noises, white noise and pink noise are stationary noises while others are all non-stationary noises, namely tank, military vehicle, jet cockpit, HF channel, F-16 cockpit, factory floor, car interior, machine gun, speech babble, and speech-shaped noises. The test set for each noise and SNR thus consisted of 28.10 min of noisy speech of which 62.51% was only noise.

3.4.2 Performance evaluation

Performance of a VAD algorithm can be evaluated both subjectively and objectively. In general, subjective evaluation is done through a listening test, and VAD decision errors are detected based on human perception [53]. On the other hand, objective evaluation relies on a mathematical criterion for judging. However, subjective listening tests like ABC [53] fail to consider the effect

of the false alarm which is inappropriate for a thorough evaluation of a VAD algorithm [40]. Therefore, the objective evaluation scheme proposed by Freeman et al. [34] was adopted to evaluate the performance of the proposed VAD algorithm.

3.4.2.1 Error rate

The four traditional parameters that describe the error rate are as follows:

- *Front-end clipping (FEC).* Clipping introduced in passing from noise to speech activity.
- *Mid-speech clipping (MSC).* Clipping due to speech misclassified as noise in an utterance.
- *Noise detected as speech (NDS).* Noise detected as speech within a silence period.
- *Carry over (OVER).* Noise interpreted as speech due to the VAD flag remaining active in passing from speech activity to noise.

These four parameters are illustrated in Figure 3.15. Among them, FEC and MSC are indicators of true rejection, while NDS and OVER are indicators of false acceptance. Thus, in order to obtain the best overall system performance, all four parameters should be minimized.

3.4.2.2 Accuracy rate

Although the method described above provides useful objective information concerning the performance of a VAD algorithm, it only gives the error rate of the system. Parameters which describe the accuracy rate are needed for a thorough analysis of the detection results. Three parameters concerning the accuracy rate are described as follows:

- *CORRECT*. They are correct decisions made by VAD algorithm.

- *Speech hit rate (HR1).* Speech frames that are correctly detected among all speech frames.
- *Non-speech hit rate (HR0).* Non-speech frames that are correctly detected among all non-speech frames.

Among the three parameters, HR1 and HR0 define the fraction of all actual speech frames or non-speech frames that are correctly detected as speech frames or non-speech frames, respectively [44]. The speech hit rate and non-speech hit rate are calculated as follows:

$$\text{HR1} = \frac{N_{1,1}}{N_1^{\text{ref}}} \qquad \text{HR0} = \frac{N_{0,0}}{N_0^{\text{ref}}}, \tag{3.10}$$

where N_1^{ref} and N_0^{ref} are the numbers of real speech and non-speech frames in the whole database, respectively, while $N_{1,1}$ and $N_{0,0}$ are the numbers of speech and non-speech frames correctly classified. The overall accuracy rate (CORRECT) is then defined as:

$$\text{CORRECT} = \frac{N_{1,1} + N_{0,0}}{N_1^{\text{ref}} + N_0^{\text{ref}}}. \tag{3.11}$$

All three parameters should be maximized to get the best performance.

3.5 Simulation results

In order to gain a comparative analysis of the proposed LSFM-based VAD performance, three modern standardized VAD schemes a statistical model-based VAD and one recent long-term algorithm, namely ETSI adaptive multi-rate (AMR) VAD options 1 and 2 (AMR1 and AMR2) [16], the G.729B VAD [13], Sohn [18] and LTSV [19], were also evaluated. The implementations of these three schemes were taken from the authors' C implementations [54, 55], respectively.

One important aspect of the comparison is the different frame lengths used. The proposed schemes, the G.729B VAD, Sohn and LTSV-based VAD, produce a decision every 10 ms, while the AMR VADs need 20 ms. In order to be

comparable, the frame-wise VAD decisions produced by the AMR VADs were compared to a set of reference labels generated every 20 ms from the TIMIT phonetic level transcription. Meanwhile, the proposed schemes, the G.729B VAD, Sohn and LTSV-based VAD, were compared to a set of reference labels generated every 10 ms from the TIMIT phonetic level transcription. The TIMIT utterances were down-sampled to 8 kHz for the software implementations of the G.729B VAD and AMR VADs. The final VAD decisions were made, and the accuracy rate and error rate were computed for 12 noises and five SNRs.

3.5.1 Performance average over all twelve kinds of noises

In Figure 3.16, the proposed LSFM-based VAD is compared with three standards, Sohn VAD and LTSV-based VAD in terms of accuracy rate and error rate for SNR levels ranging from −10 to 10 dB. Note that the results in Figure 3.16 are averaged values for all 12 noises. The first row of the figure shows the accuracy rates which include CORRECT, HR1, and HR0. The behavior of the different VADs is analyzed. G.729B suffers poor CORRECT(62.74% at −10 dB) and HR1 (33.62% at −10 dB) with the increasing noise level, while it keeps a steady and relatively high HR0 for the whole range of SNRs (80.33% on average). The Sohn algorithm outperformed the G.729B VAD in terms of both CORRECT and HR1. AMR1 performs much better than Sohn algorithm for both CORRECT and HR1 while suffering degradation of HR0 when the SNR level is increased. AMR2 improves considerably over AMR1 in CORRECT mainly because of the high HR0 over all SNRs (88.96% on average) while yielding similar HR1 with AMR1. LTSV performs very well under low SNR conditions (80.73% CORRECT at −10 dB) but becomes saturated (around 91% since 5 dB) at higher SNRs. Our proposed LSFM-based VAD yields the best CORRECT for all SNRs and shows a steady improvement with the increased SNR.

Similarly, the second row of Figure 3.16 shows the error rates which include FEC, MSC, OVER, and NDS. The Sohn algorithm performs the worst on average in terms of true rejection rate (FEC and MSC). However, it also achieved

the lowest OVER among the six VADs tested. The NDS of G.729B is the highest of all. AMR1 and AMR2 yield similar true rejection rates for all tested SNRs, while AMR2 gives smaller false alarm rate (NDS and OVER) especially for NDS (around four points less than AMR1 for all SNRs). LTSV leads to the lowest true rejection rate, while LSFM achieved the best performance in terms of NDS. The proposed LSFM-based VAD acquires a comparatively higher FEC in low SNRs (smaller than -5 dB) because of the averaging property of this algorithm shown in Equations 3.2, 3.3, and 3.4.

Table 3.2 summarizes the results provided by LSFM-based VAD over the other five different VAD methods being evaluated by comparing them in terms of the average accuracy rate and error rate for all 12 noises over five SNR levels ranging from -10 to 10 dB. LSFM achieves the best CORRECT (88.95%) and HR0 (91.00%), while LTSV yields the best HR1 (88.04%).

Table 3.2: Average performance comparison for all 12 noises over five SNR levels ranging from -10 to 10 dB

VAD	AMR1	AMR2	G.729B	Sohn	LTSV	LSFM
CORRECT	81.00	86.07	70.87	75.10	88.08	88.95
HR1	78.96	81.25	55.16	54.96	88.04	85.53
HR0	82.22	88.96	80.33	87.18	88.10	91.00
FEC	1.46	1.09	2.28	3.11	0.41	0.49
MSC	6.42	5.94	14.56	13.77	4.07	4.93
OVER	2.53	2.15	1.06	0.40	2.46	1.39
NDS	8.59	4.75	11.24	7.62	4.98	4.24

3.5.2 Performance average over five SNRs

Figure 3.17 shows the three different accuracy rate evaluation metrics averaged over five SNRs for 12 kinds of noises computed for AMR1, AMR2, G.729B, Sohn, LTSV, and LSFM-based VAD algorithms. From Figure 3.17, it is clear that in terms of CORRECT, LTSV is the best among all five reference VAD algorithms considered here. Hence, the proposed LSFM-based VAD is com-

pared with the LTSV-based VAD. We observe that on average, the LSFM-based VAD is better than the LTSV-based VAD in terms of CORRECT for tank (0.52%), military vehicle (1.40%), F-16 cockpit (0.34%), car interior (2.12%), machine gun (1.88%), and speech babble (7.63%) noises, and it is worse for white (1.10%), pink (0.79%), jet cockpit (0.64%), HF channel (0.15%), factory floor (0.15%), and speech-shaped (0.58%) noises. The number in the bracket indicates the absolute CORRECT by which the proposed LSFM-based VAD is better or worse than the LTSV-based VAD. The mean CORRECT over all 12 noise types of our proposed LSFM-based VAD is 0.87% higher than that of the LTSV-based VAD. Furthermore, the proposed LSFM-based VAD outperforms LTSV-based VAD in terms of HR0 over most noises (11 out of 12) that were considered.

Figure 3.18 shows the four different error rate evaluation metrics (FEC, MSC, OVER, and NDS), averaged over five SNRs for 12 kinds of noises, computed for AMR1, AMR2, G.729B, Sohn, LTSV, and LSFM algorithms. From Figure 3.18, it is clear that the performance of LSFM-based VAD outperforms the LTSV-based VAD in terms of OVER (all 12 noises) and NDS (9 out of 12) which means that our proposed algorithm performs better in terms of false alarm rate. For example, The proposed LSFM-based VAD has a smaller NDS score for tank (0.26%), military vehicle (0.10%), jet cockpit (0.35%), HF channel(0.44%), F16 cockpit (0.80%), factory floor (0.42%), car interior (0.31%), machine gun (2.15%), and babble (5.83%) noises. The number in the bracket indicates the absolute NDS by which the proposed LSFM-based VAD is smaller than the LTSV-based VAD. Moreover, values of standard deviation of our proposed LSFM-based VAD in terms of MSC, OVER, and NDS are all smaller than that of the LTSV-based VAD.

Thus, in consideration of both accuracy rate and error rate, the proposed VAD algorithm achieved the best compromise when compared with the four representative VADs analyzed.

3.5.3 Statistical Significance test of the six VAD algorithms

The accuracy and error rate evaluation metrics obtained from six VAD algorithms were subjected to statistical analysis in order to assess their significance differences. Table 3.3 shows the ANOVA (analysis of variance) results by SPSS for all twelve noise types. A highly significant effect (p <0.005) was found for most noises tested except the FEC case.

Table 3.3: ANOVA test results for all six VAD algorithms

Noises	CORRECT	HR1	HR0	FEC	MSC	NDS	OVER
White	.004	.002	.000	.392	.001	.009	.000
Pink	.015	.006	.000	.245	.004	.034	.000
Tank	.000	.001	.001	.182	.000	.000	.000
Military	.000	.000	.000	.057	.000	.000	.000
Jet cockpit	.007	.001	.000	.149	.003	.013	.001
HF channel	.000	.000	.000	.215	.000	.000	.000
F16	.022	.006	.000	.327	.002	.004	.000
Factory	.048	.135	.003	.960	.078	.001	.000
Car Interior	.000	.000	.000	.007	.000	.000	.000
Machinegun	.000	.000	.000	.004	.000	.000	.000
Babble	.000	.000	.000	.046	.000	.000	.000
SSN	.059	.005	.001	.359	.003	.009	.000

The value is marked in green if it is smaller than 0.005.

Following the ANOVA, multiple comparison statistical tests according to Tukey's HSD test were also done to assess significance between algorithms. The difference was deemed significant if the p value was smaller than 0.05. Tables 3.4, 3.5, 3.6, 3.7, and 3.8 show the comparisons between our proposed LSFM based VAD and the other five VADs. If our proposed LSFM based VAD gives better results, the mean difference (MD) and p value are marked in green, blanks stand for no significant difference between two algorithms. From these tables we can conclude that our proposed method performed better for most noises tested when comparing with G729B and Sohn methods. As

for AMR1, LSFM performed better for machinegun and babble noise in terms of both accuracy rate and false alarm rate (NDS+OVER). When compares with AMR2 and LTSV, the LSFM algorithm achieved similar results for most conditions tested.

Table 3.4: Statistical significance test results for accuracy and rate between LSFM and AMR1 VADs

LSFM-AMR1	**CORRECT**		**HR1**		**HR0**	
Noise	MD	p	MD	p	MD	p
White						
Pink						
Tank						
Military						
Jet cockpit						
HFchannel	20.77	.002			36.28	.000
F16						
Factory						
Car Interior						
Machinegun	5.97	.002	-20.62	.000	21.91	.000
Babble	24.24	.000			51.94	.000
SSN					-7.83	.008

LSFM-AMR1	**FEC**		**MSC**		**NDS**		**OVER**	
Noise	MD	p	MD	p	MD	p	MD	p
White								
Pink								
Tank								
Military								
Jet cockpit								
HFchannel					-17.55	.000	-5.13	.003
F16								
Factory								
Car Interior								
Machinegun			7.51	.000	-12.31	.000		
Babble					-26.08	.000	-6.39	.000
SSN								

Table 3.5: Statistical significance test results for accuracy and rate between LSFM and AMR2 VADs

LSFM-AMR2	**CORRECT**		**HR1**		**HR0**	
Noise	MD	p	MD	p	MD	p
White						
Pink						
Tank						
Military					3.07	.017
Jet cockpit						
HFchannel						
F16						
Factory						
Car Interior						
Machinegun			-21.68	.000	17.17	.000
Babble					10.58	.030
SSN					-6.62	.033

LSFM-AMR2	**FEC**		**MSC**		**NDS**		**OVER**	
Noise	MD	p	MD	p	MD	p	MD	p
White								
Pink								
Tank							-1.13	.047
Military								
Jet cockpit								
HFchannel								
F16								
Factory								
Car Interior								
Machinegun			7.91	.000	-11.07	.000	-2.38	.021
Babble								
SSN								

Table 3.6: Statistical significance test results for accuracy and rate between LSFM and G729B VADs

LSFM-G729B	**CORRECT**		**HR1**		**HR0**	
Noise	MD	p	MD	p	MD	p
White			47.37	.036	-2.13	.008
Pink					-2.54	.031
Tank	15.70	.003	38.88	.011		
Military	43.93	.000	17.29	.001	59.90	.000
Jet cockpit					-3.63	.038
HFchannel	16.31	.020	49.90	.007		
F16						
Factory						
Car Interior	30.55	.000	14.88	.000	39.81	.000
Machinegun	8.09	.000	-16.73	.000	22.86	.000
Babble	27.45	.000	29.16	.035	26.14	.000
SSN					-8.22	.005

LSFM-G729B	**FEC**		**MSC**		**NDS**		**OVER**	
Noise	MD	p	MD	p	MD	p	MD	p
White			-13.54	.018				
Pink								
Tank			-13.66	.007	-1.99	.000		
Military			-6.37	.000	-33.54	.000	-3.90	.000
Jet cockpit			-12.69	.042				
HFchannel			-15.16	.005				
F16			-12.52	.049				
Factory							1.34	.012
Car Interior			-5.40	.000	-26.96	.000	1.94	.000
Machinegun			6.23	.000	-13.98	.000		
Babble			-10.93	.011	-16.43	.000		
SSN					4.11	.039	1.03	.028

Table 3.7: Statistical significance test results for accuracy and rate between LSFM and Sohn VADs

LSFM-Sohn	**CORRECT**		**HR1**		**HR0**	
Noise	MD	p	MD	p	MD	p
White	18.06	.029	51.70	.019	-2.11	.009
Pink					-2.71	.019
Tank	12.73	.022	34.04	.032		
Military	5.17	.002	19.23	.000	-3.27	.010
Jet cockpit			52.83	.011	-3.75	.030
HFchannel			44.96	.017		
F16					-3.13	.013
Factory					15.28	.007
Car Interior			10.61	.000		
Machinegun	5.20	.010	-17.82	.000	19.00	.000
Babble					38.78	.000
SSN			59.53	.012	-8.38	.004

LSFM-Sohn	**FEC**		**MSC**		**NDS**		**OVER**	
Noise	MD	p	MD	p	MD	p	MD	p
White			-14.33	.011				
Pink			-13.60	.024				
Tank			-12.07	.022				
Military			-7.01	.000				
Jet cockpit			-12.86	.038				
HFchannel			-14.75	.007				
F16			-14.23	.019				
Factory					-10.60	.001		
Car Interior			-3.90	.000			2.02	.000
Machinegun			6.45	.000	-14.40	.000	2.52	.001
Babble					-24.49	.000		
SSN			-14.73	.010	4.18	.034	1.06	.022

Table 3.8: Statistical significance test results for accuracy and rate between LSFM and LTSV VADs

LSFM-LTSV	**CORRECT**		**HR1**		**HR0**	
Noise	MD	p	MD	p	MD	p
White						
Pink						
Tank					2.29	.034
Military						
Jet cockpit						
HFchannel						
F16					2.86	.028
Factory						
Car Interior						
Machinegun			-5.76	.001	6.45	.049
Babble					11.39	.016
SSN						

LSFM-LTSV	**FEC**		**MSC**		**NDS**		**OVER**	
Noise	MD	p	MD	p	MD	p	MD	p
White								
Pink								
Tank							-1.17	.039
Military								
Jet cockpit								
HFchannel								
F16								
Factory								
Car Interior							-1.75	.001
Machinegun			2.30	.000			-1.88	.019
Babble								
SSN								

3.6 Conclusions

The main contribution of this chapter was the introduction of an efficient long-term spectral flatness measure-based VAD algorithm. The motivation of exploring flatness measure along time frames using a long window was clarified by the LSFM feature distributions as a function of the long-term window length R. The discriminative power of the LSFM feature was verified in terms of the separateness of its distribution for noisy speech and non-speech signals. The decision threshold was adapted according to the previous 100 LSFM measures of speech and non-speech. Experiments were done on core TIMIT test set for 12 kinds of noises (11 from NOISEX-92 database and speech-shaped noise) across five different SNRs ranging from -10 to 10 dB. No *a priori* knowledge of noise characteristics was needed for training purposes. The performance of our proposed method was compared with the three standards (namely G.729B, AMR1, and AMR2) Sohn algorithm, and an emerging LTSV-based VAD algorithm. The results were analyzed by accuracy rate and error rate. Statistical tests were done on those results. Through extensive experiments, we showed that our proposed LSFM-based VAD achieved better results for most noises tested when comparing with G729B and Sohn methods. Furthermore, our proposed LSFM-based VAD outperformed AMR1 for non-stationary impulsive machine gun noise and speechlike babble noise in terms of both accuracy rate and false alarm rate (NDS+OVER). However, when compares with AMR2 and LTSV algorithms, the LSFM algorithm achieved similar results for most conditions tested.

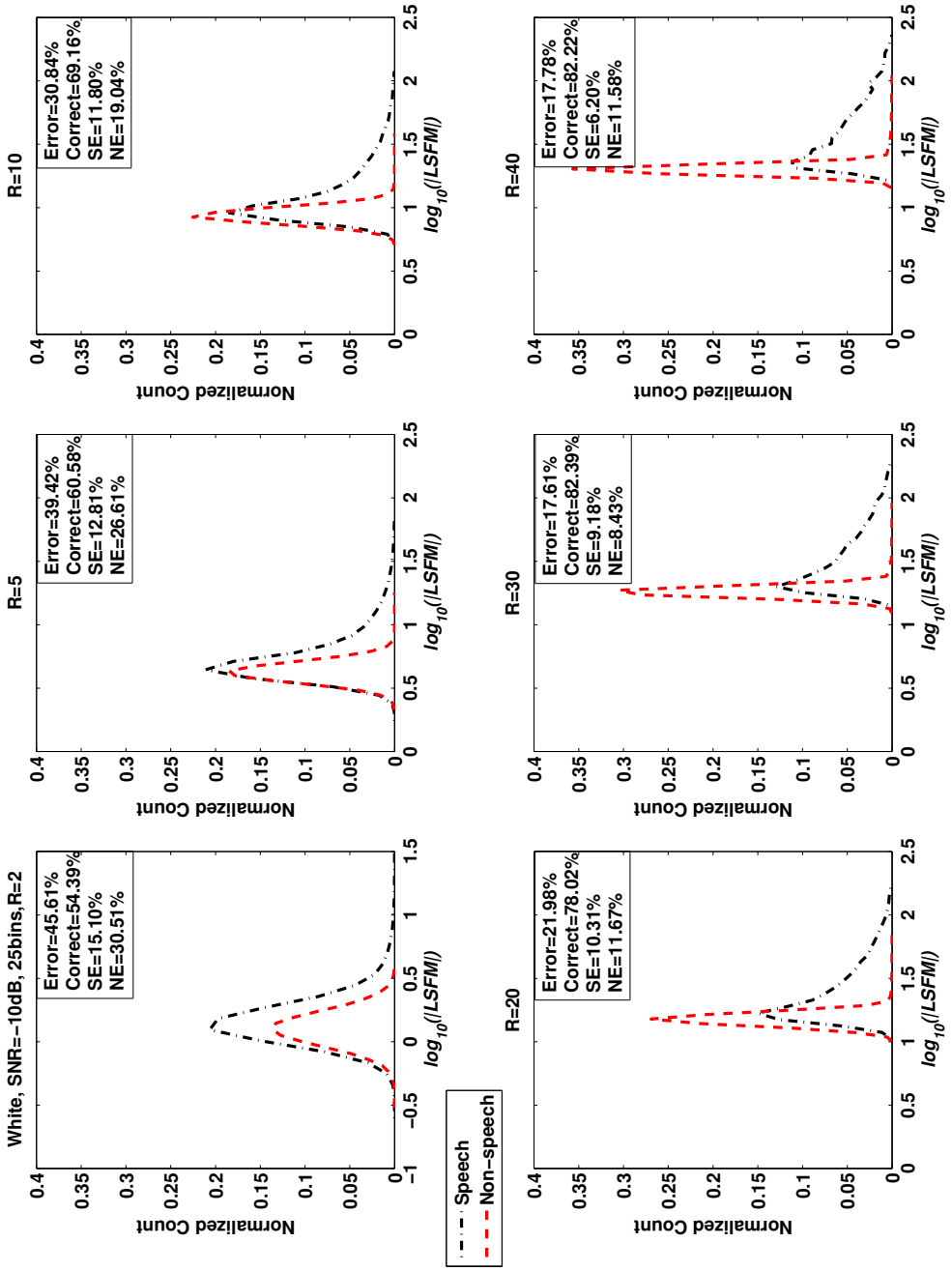


Figure 3.1: LSFM measure as a function of long-term window length (R) in additive white noise (SNR = -10 dB).

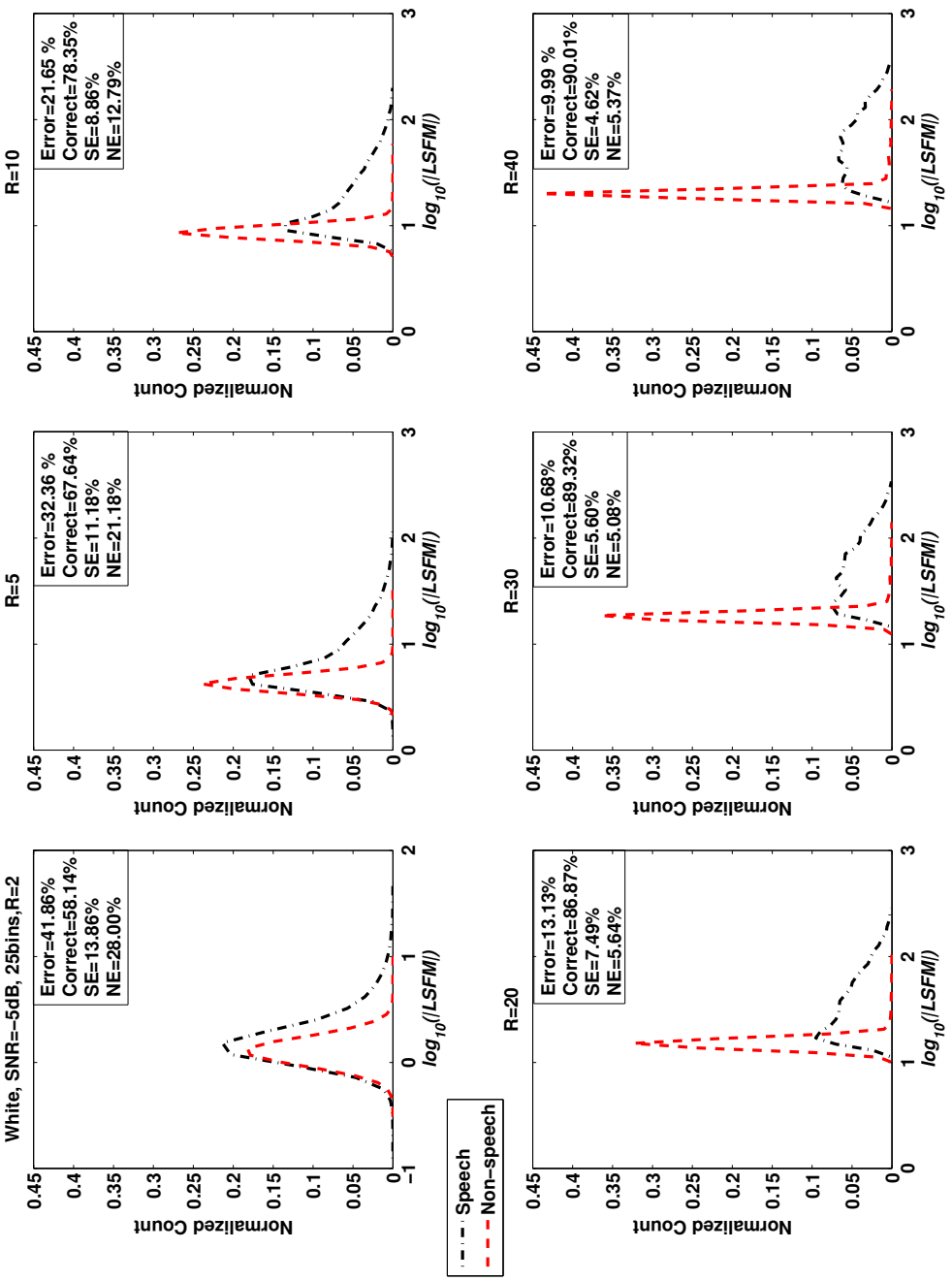


Figure 3.2: LSFM measure as a function of long-term window length (R) in additive white noise (SNR = -5 dB).

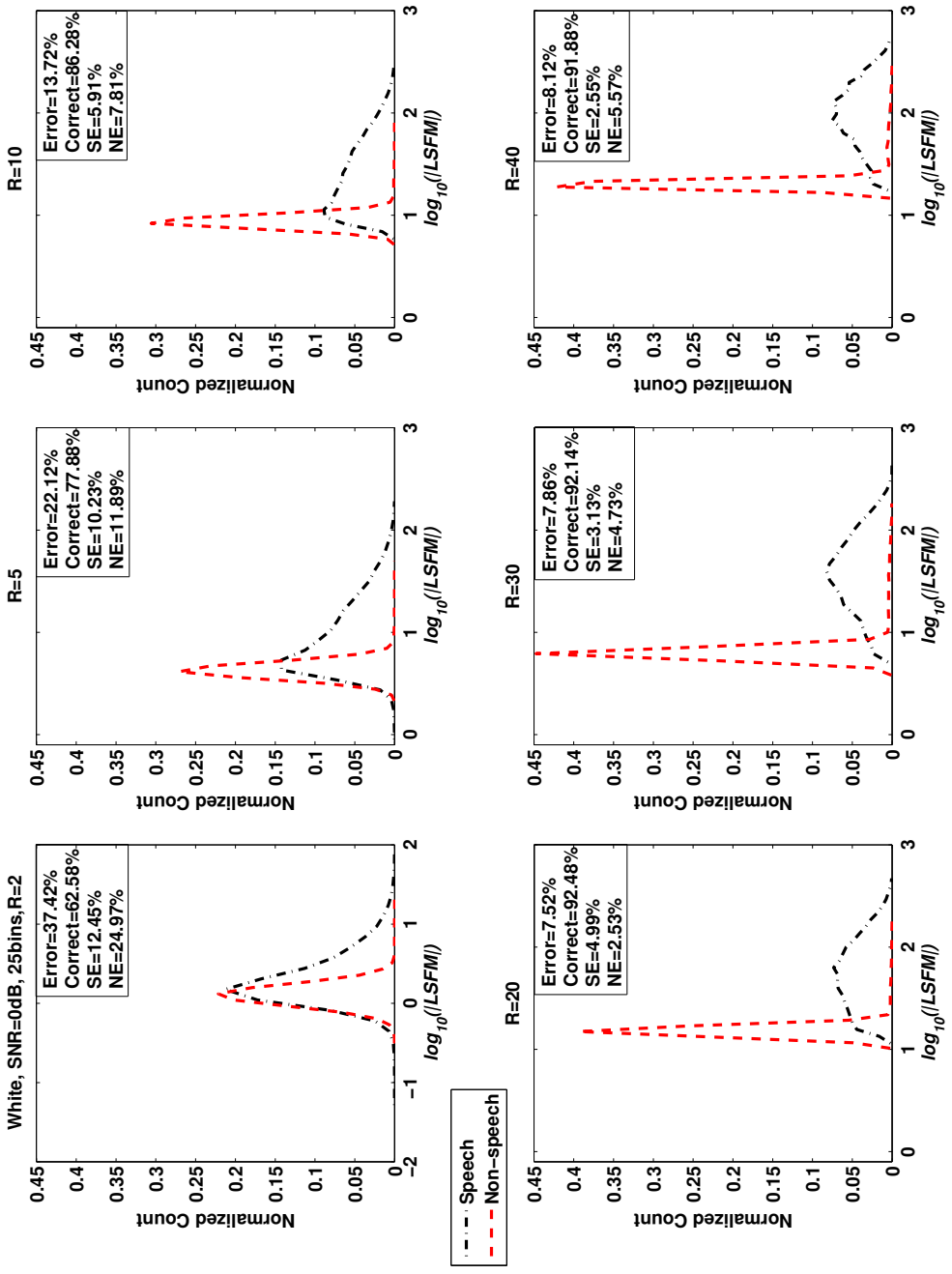


Figure 3.3: LSFM measure as a function of long-term window length (R) in additive white noise (SNR = 0 dB).

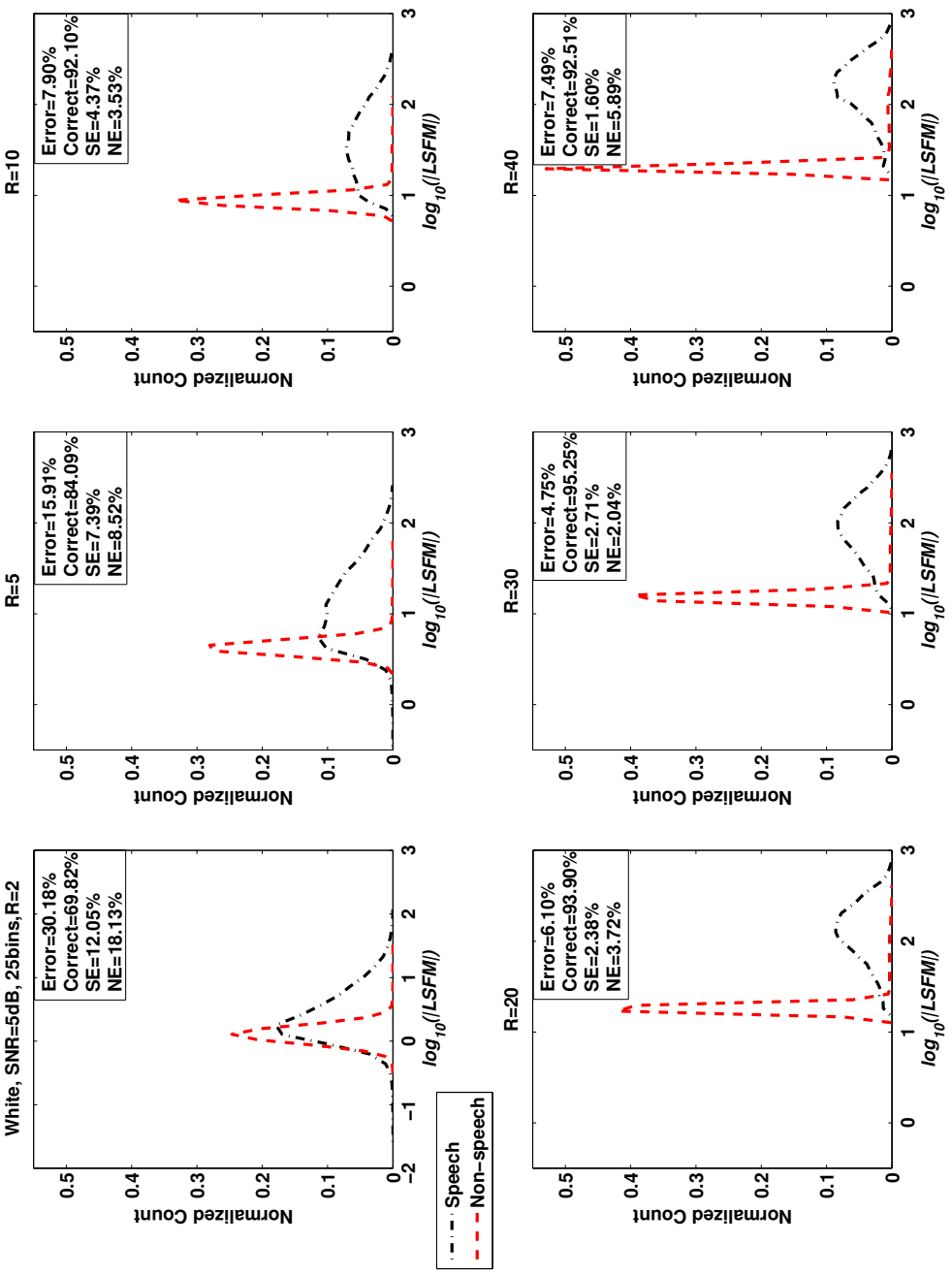


Figure 3.4: LSFM measure as a function of long-term window length (R) in additive white noise (SNR = 5 dB).

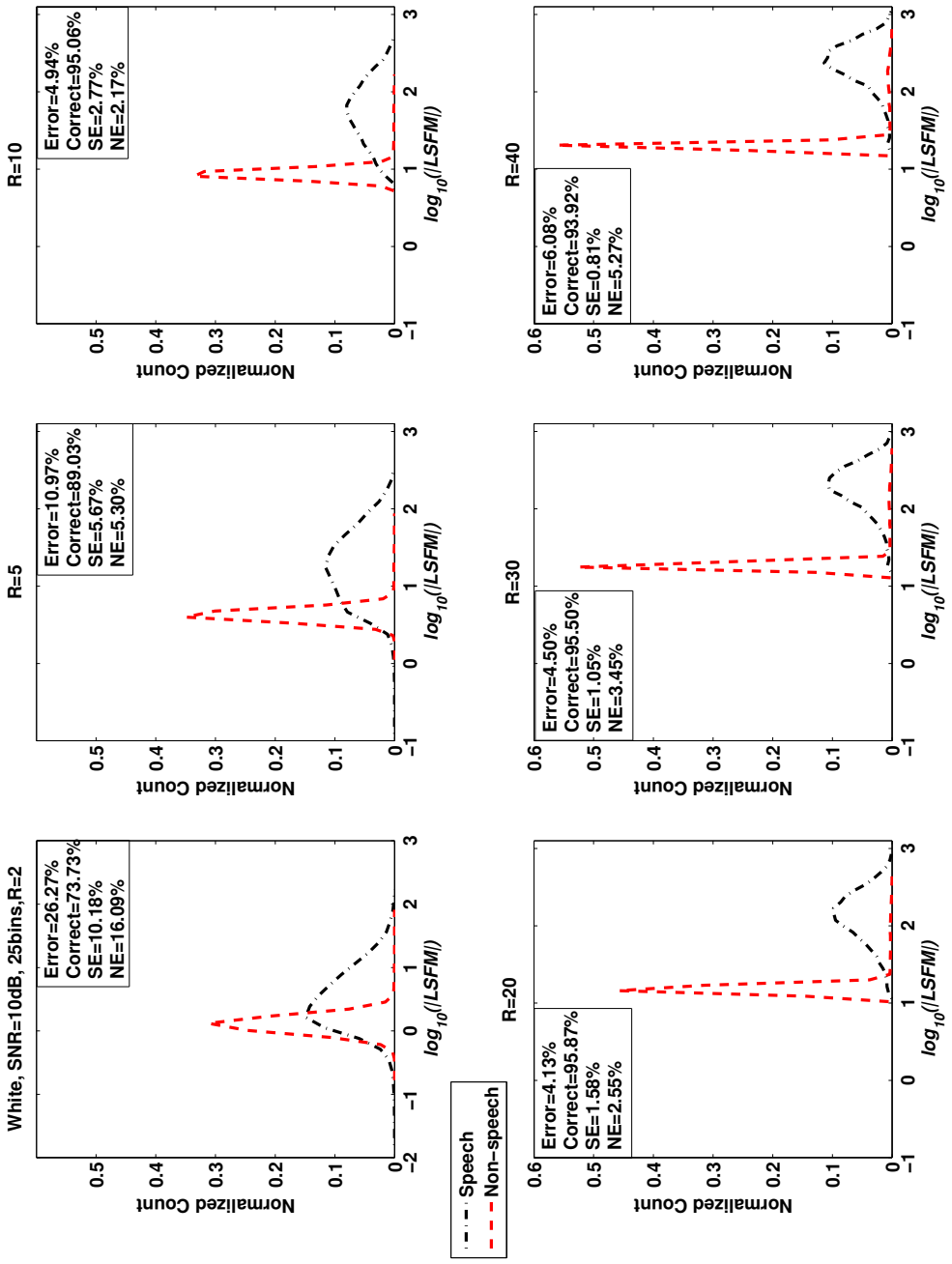


Figure 3.5: LSFM measure as a function of long-term window length (R) in additive white noise (SNR = 10 dB).

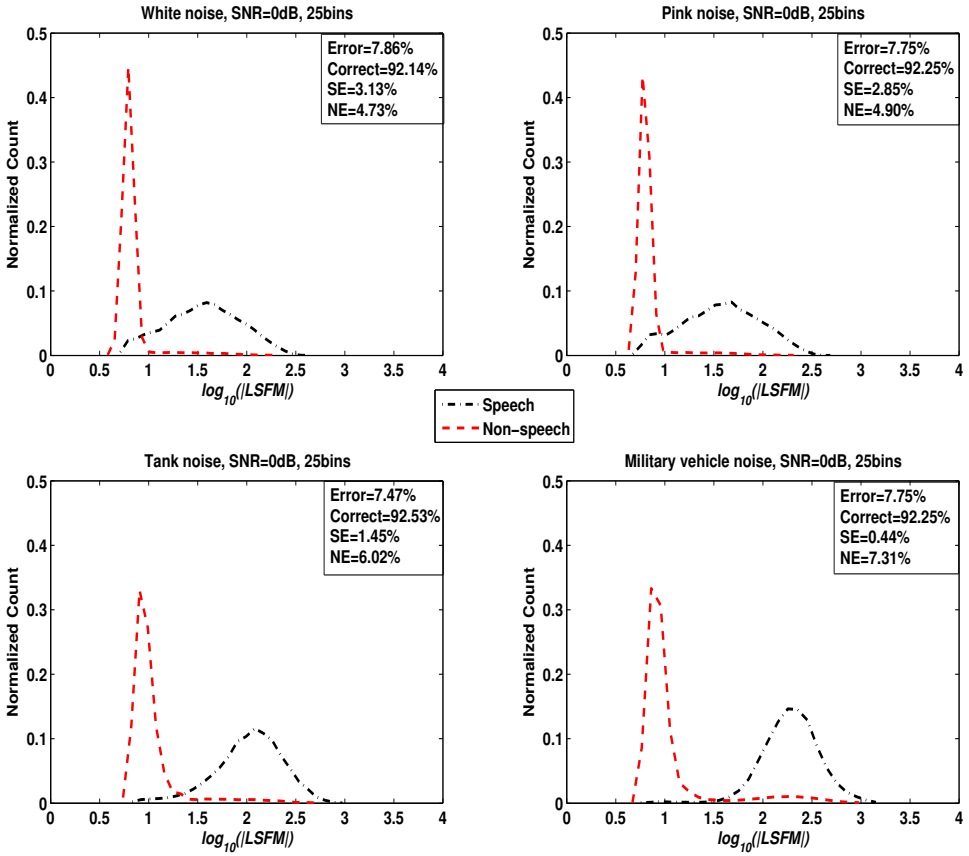


Figure 3.6: Histogram of the logarithmic LSFM measure for white, pink, tank and military vehicle noises (SNR = 0 dB). Upper left: white noise, upper right: pink noise, lower left: tank noise, and lower right: military vehicle noise.

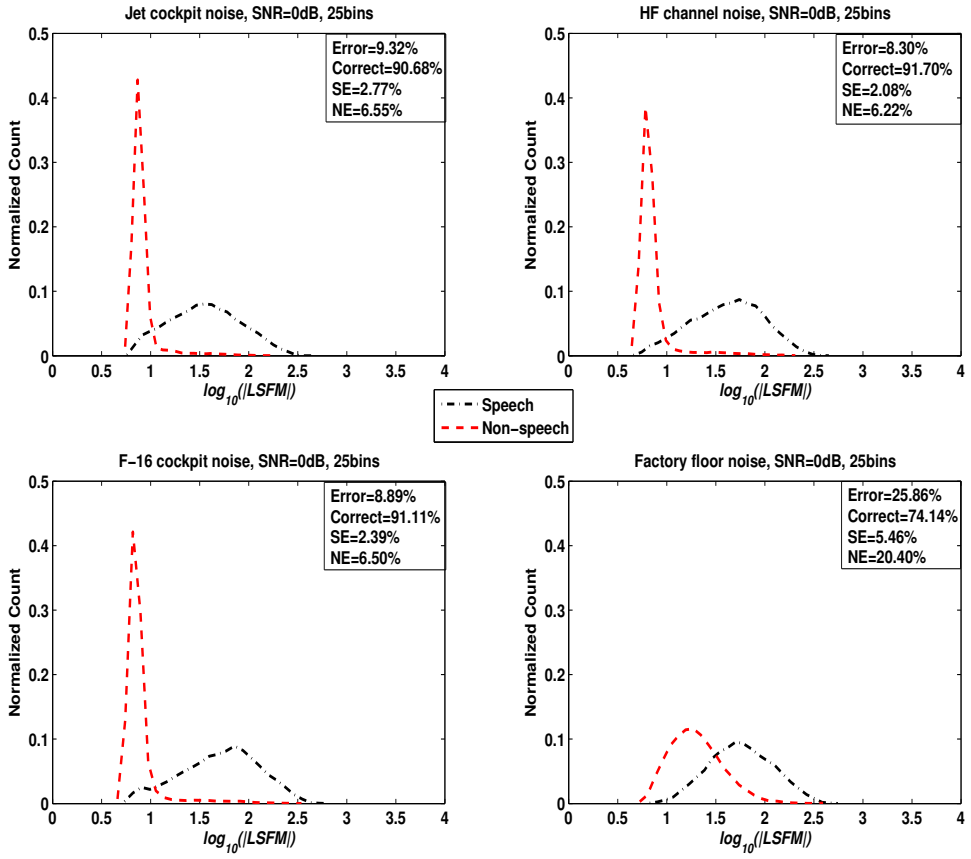


Figure 3.7: Histogram of the logarithmic LSFM measure for jet cockpit, HF channel, F-16 cockpit and factory floor noises (SNR = 0 dB). Upper left: jet cockpit noise, upper right: HF channel noise, lower left: F-16 cockpit noise, and lower right: factory floor noise.

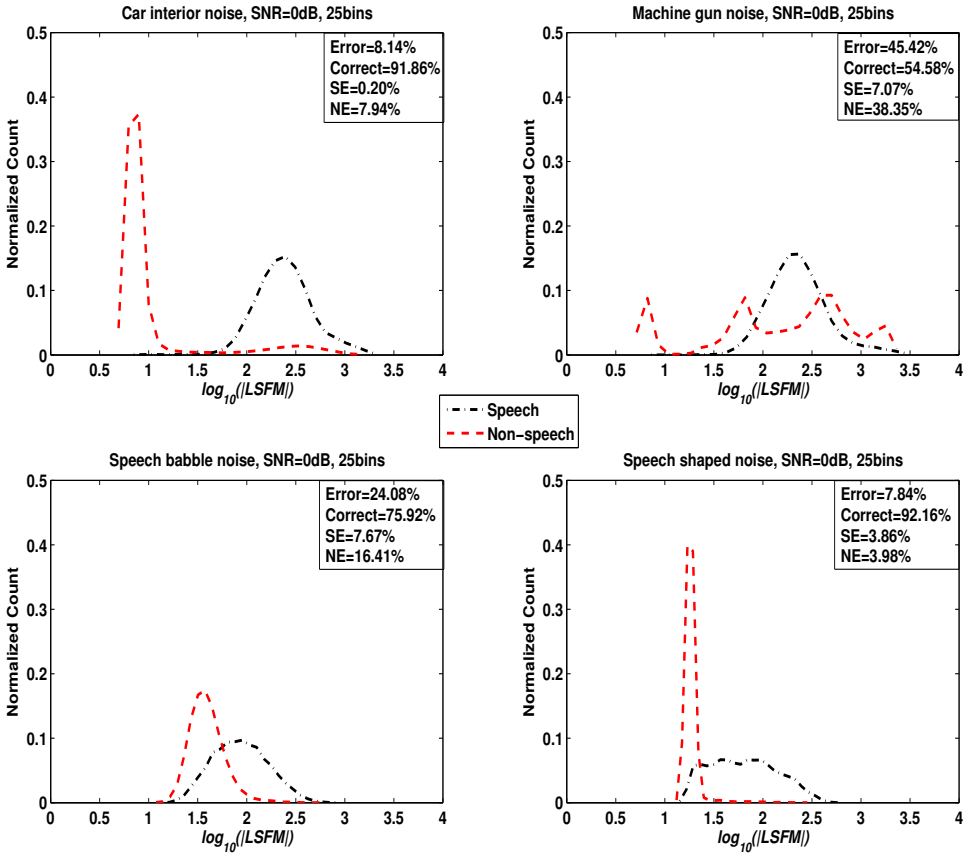


Figure 3.8: Histogram of the logarithmic LSFM measure for car interior, machine gun, speech babble and speech-shaped noises (SNR = 0 dB). Upper left: car interior noise, upper right: machine gun noise, lower left: speech babble noise, and lower right: speech-shaped noise.

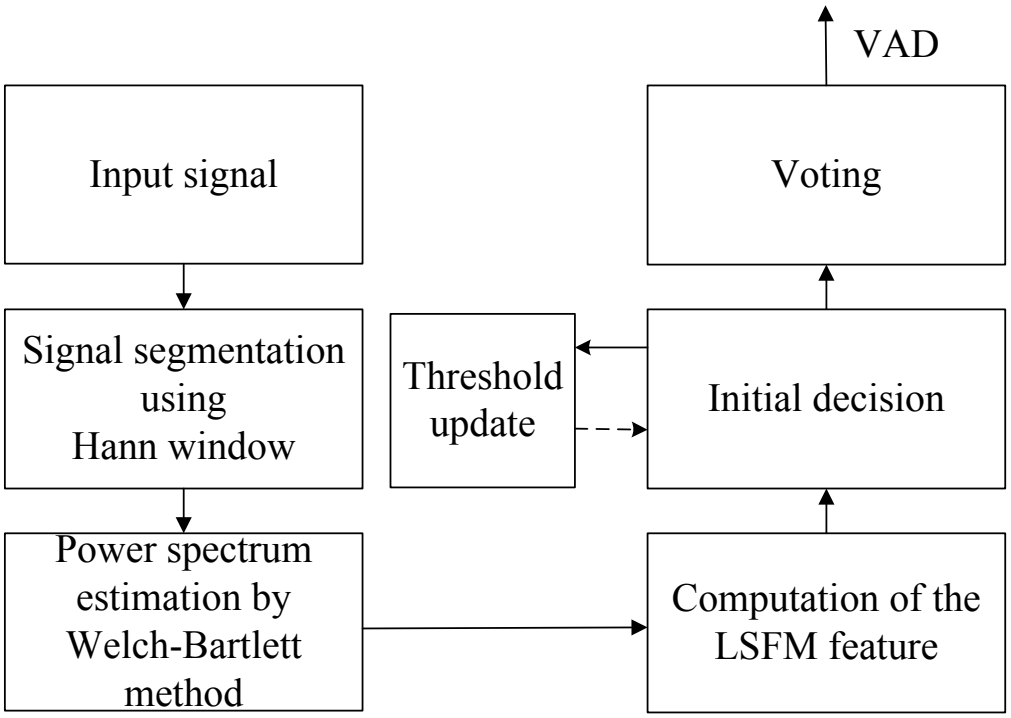


Figure 3.9: Block diagram of the proposed LSFM-based VAD algorithm.

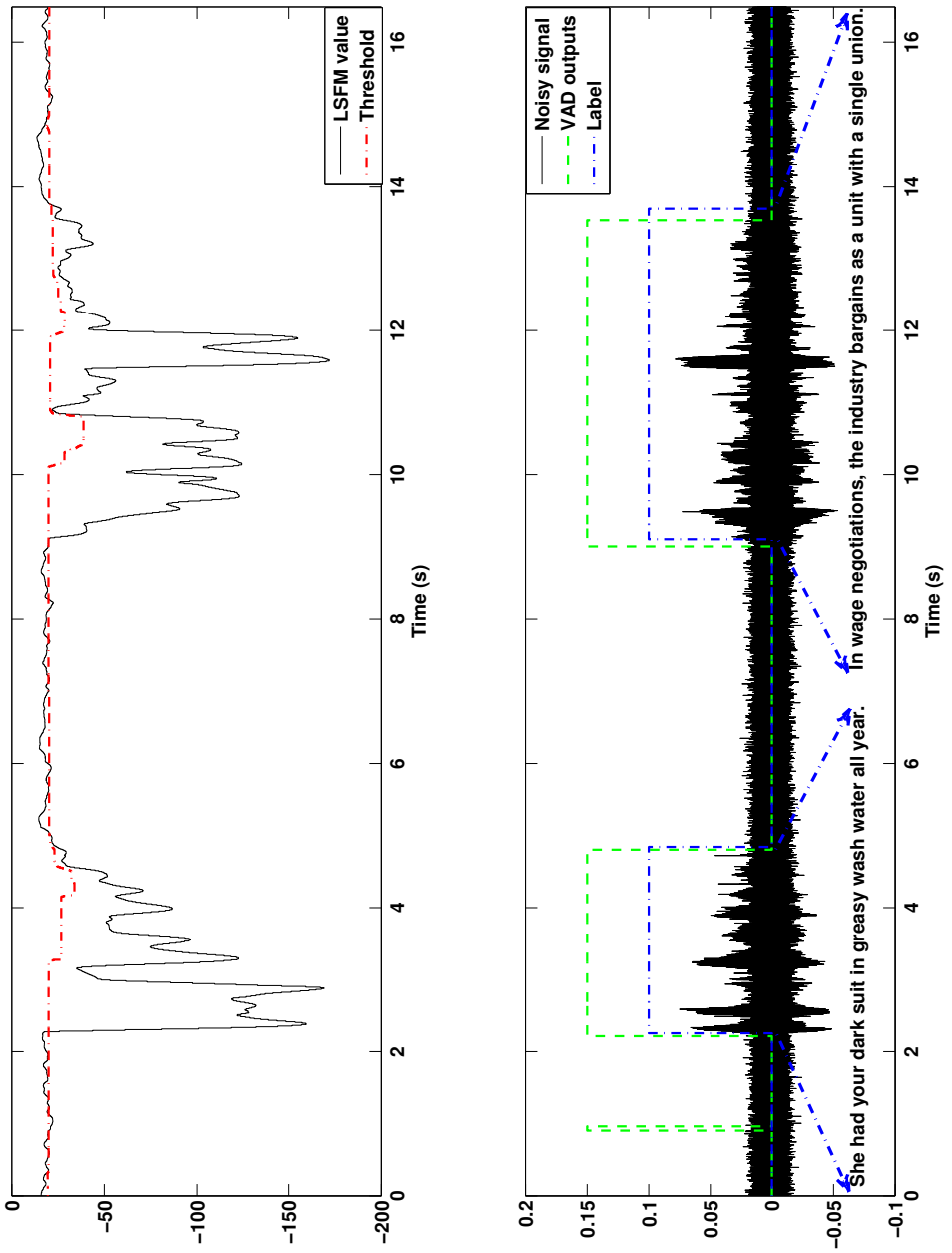


Figure 3.10: Illustrative example of the adaptive threshold and the VAD output, white noise, SNR = 0 dB. The upper figure shows the LSFM value and the corresponding adaptive threshold for each frame. The lower figure shows the VAD output decisions and the ground truth, namely 'Label'. The two sentences are as follows: (1) She had your dark suit in greasy wash water all year; (2) in wage negotiations, the industry bargains as a unit with a single union.

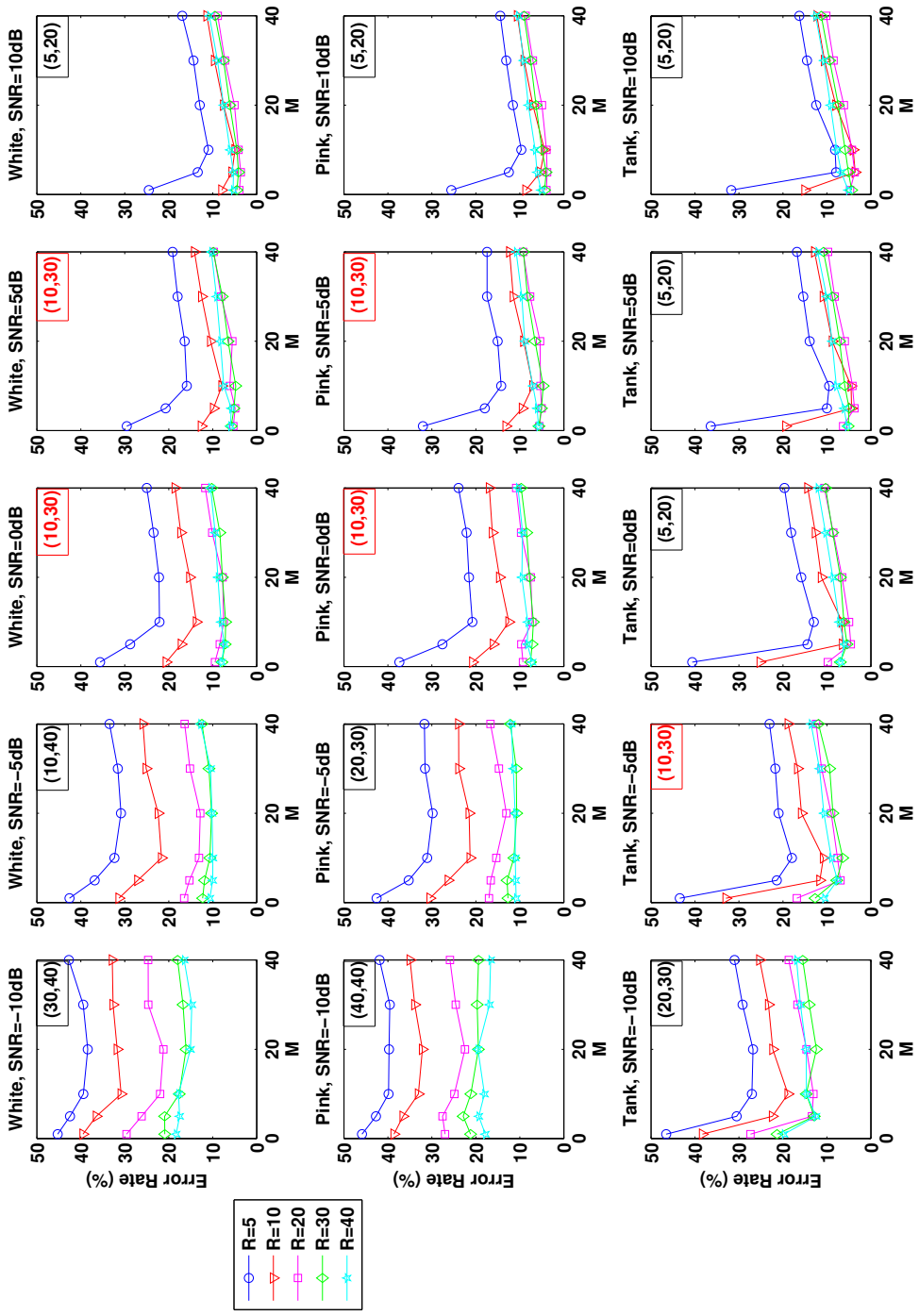


Figure 3.11: Total misclassification error as a function of M and R combination for white, pink and tank noises. Upper row: white noise, middle row: pink noise, and lower row: tank noise. SNR= -10, -5, 0, 5, and 10 dB. The best combination of M and R for each noise at each SNR level is written on the upper or lower right of each subfigure.

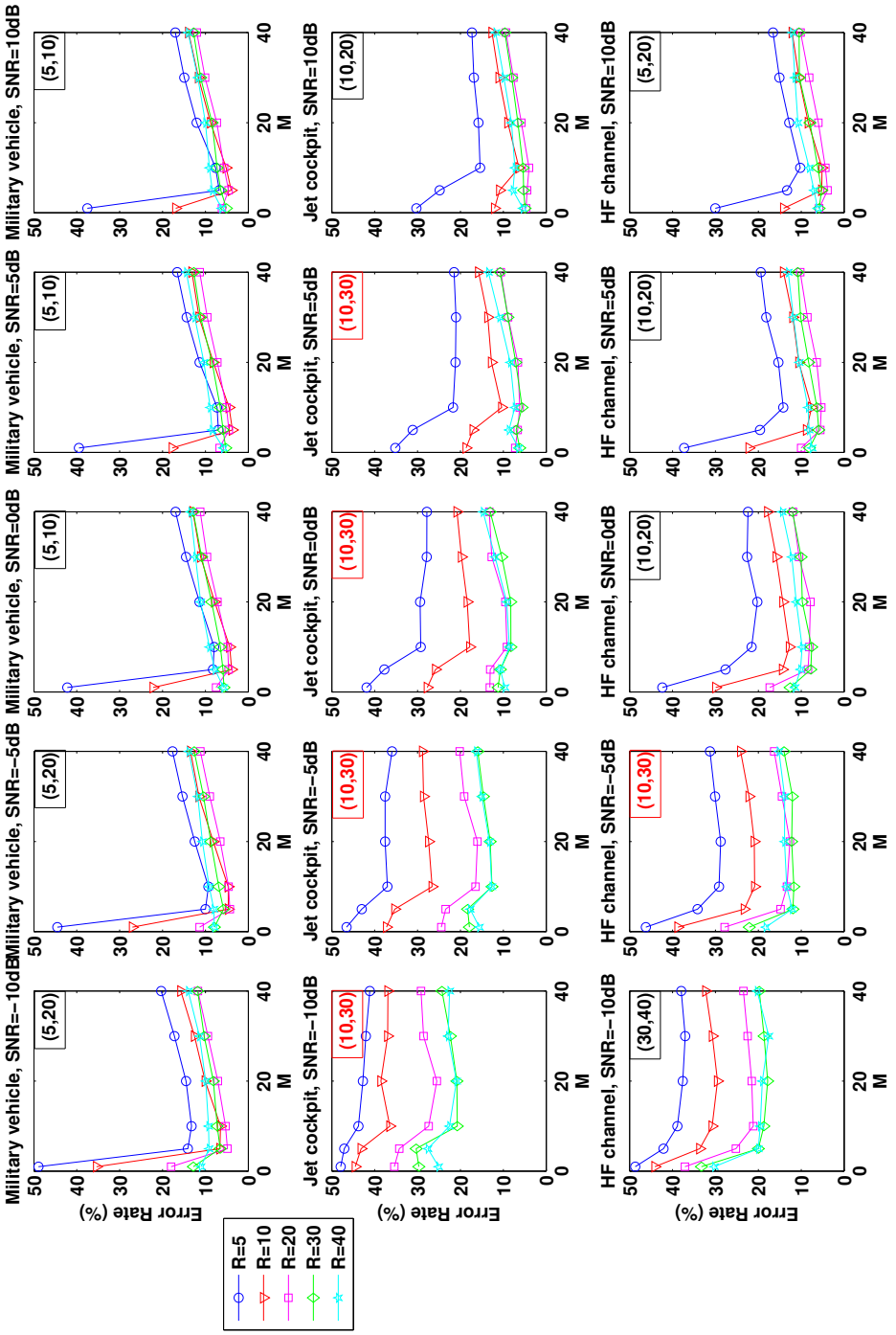


Figure 3.12: Total misclassification error as a function of M and R combination for military, jet cockpit and HF channel noises. Upper row: military noise, middle row: jet cockpit noise, and lower row: HF channel noise. SNR= -10, -5, 0, 5, and 10 dB.

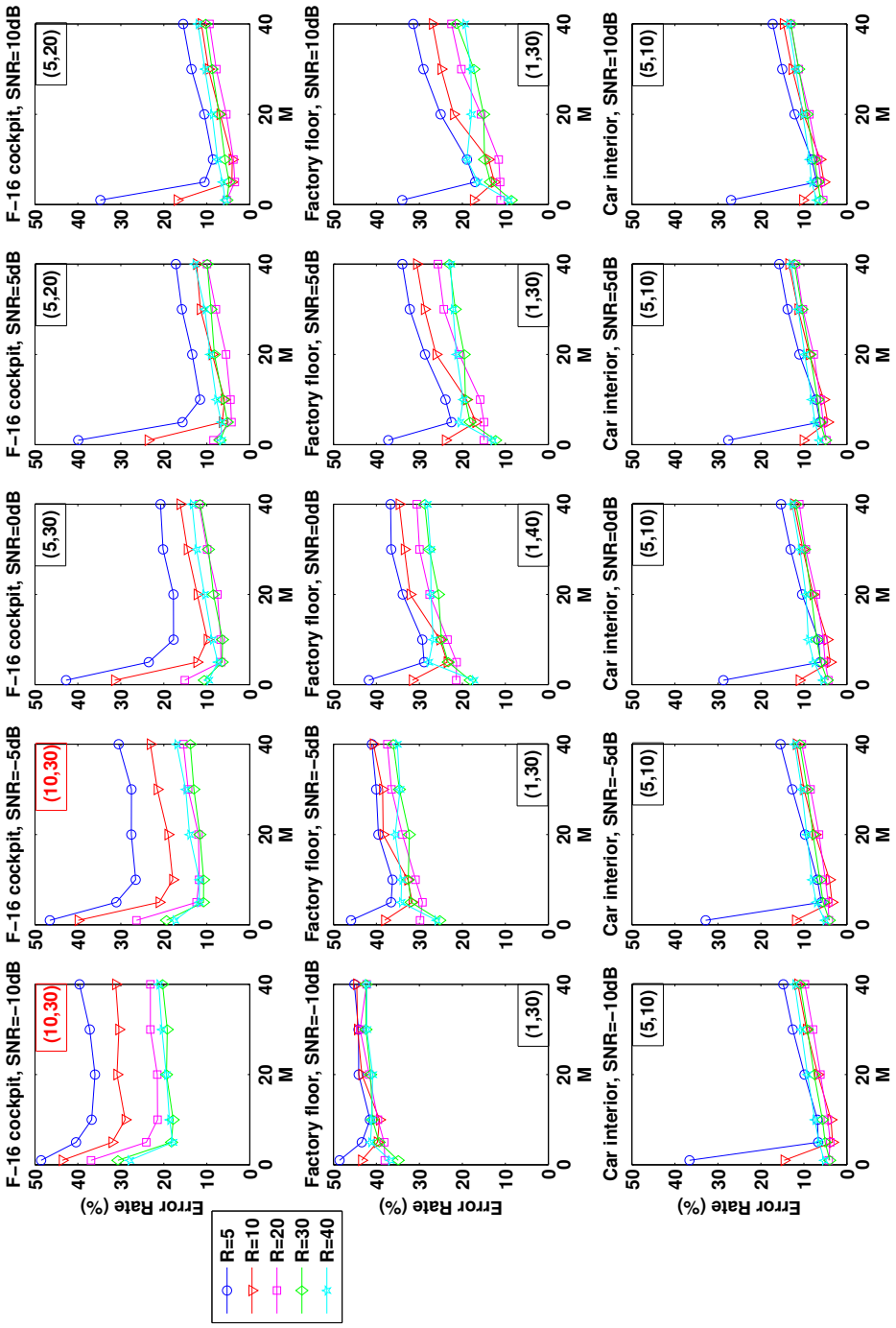


Figure 3.13: Total misclassification error as a function of M and R combination for F-16 cockpit, factory floor and car interior noises. Upper row: F-16 cockpit noise, middle row: factory floor noise, and lower row: car interior noise. SNR= -10, -5, 0, 5, and 10 dB.

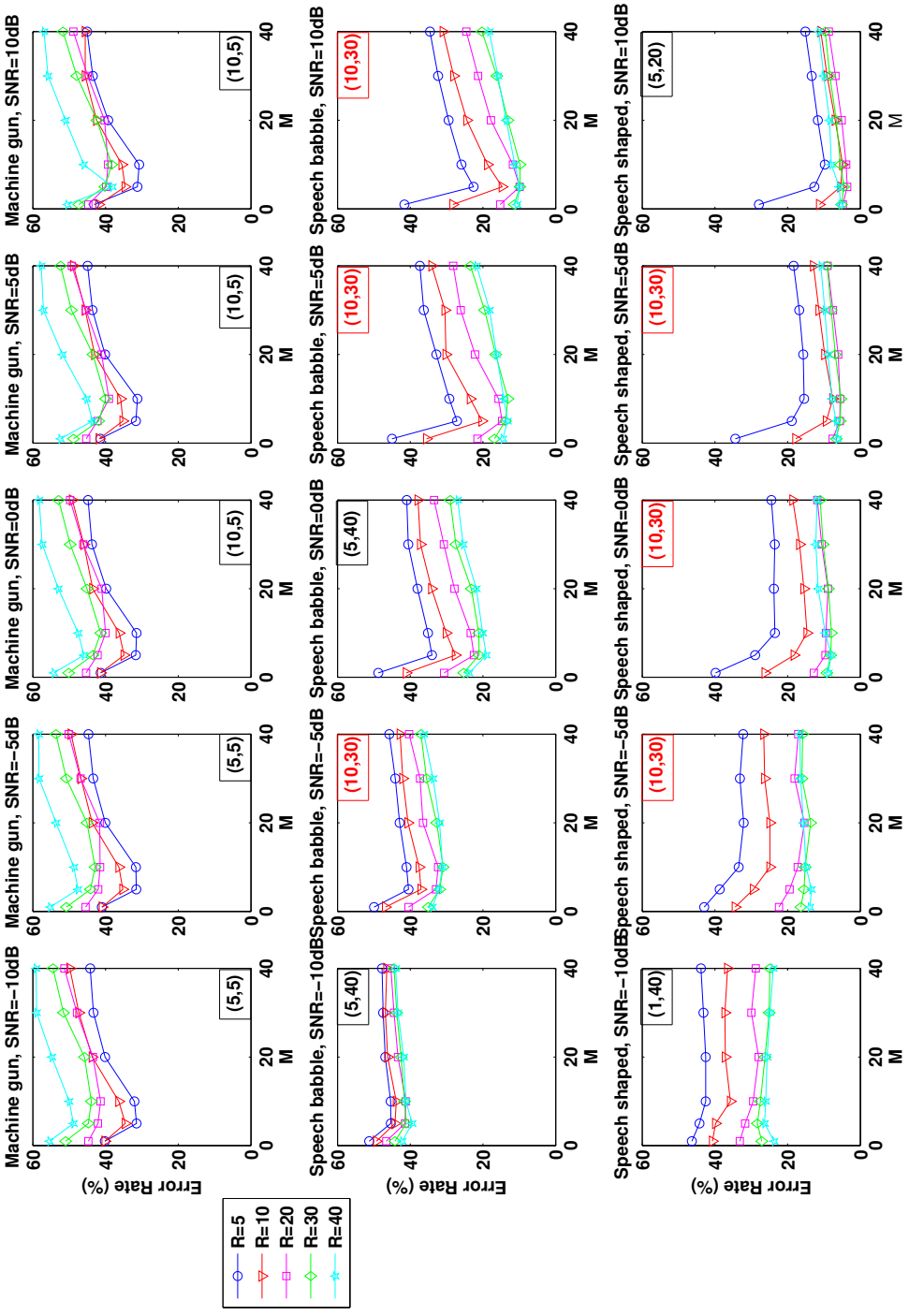


Figure 3.14: Total misclassification error as a function of M and R combination for machine gun, speech babble and speech-shaped noises. Upper row: machine gun noise, middle row: speech babble noise, and lower row: speech-shaped noise. SNR= -10, -5, 0, 5, and 10 dB.

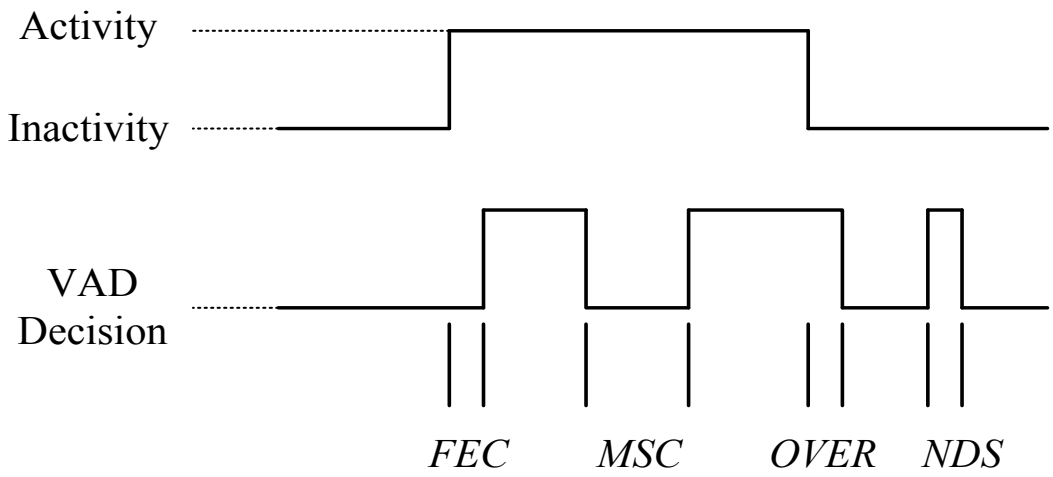


Figure 3.15: Objective parameters for performance evaluation.

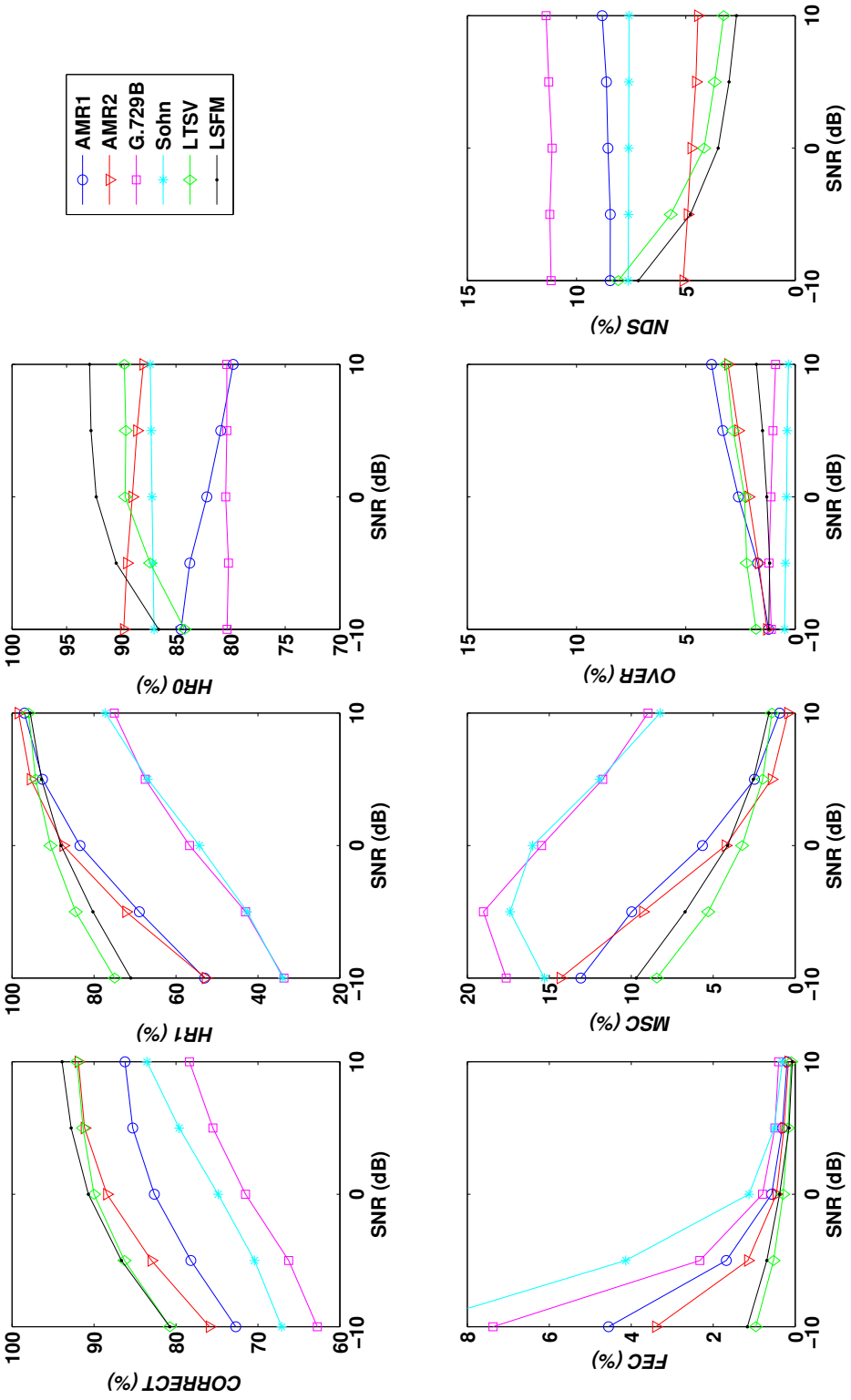


Figure 3.16: Accuracy and error rate comparisons for six VAD schemes averaged over 12 noises for five SNRs. Accuracy rate: CORRECT, HR1, and HR0; error rate: FEC, MSC, OVER, and NDS. Six VAD schemes: AMR1, AMR2, G.729B, Sohn, LTSV, and LSFM. Five SNRs (-10, -5, 0, 5, and 10 dB).

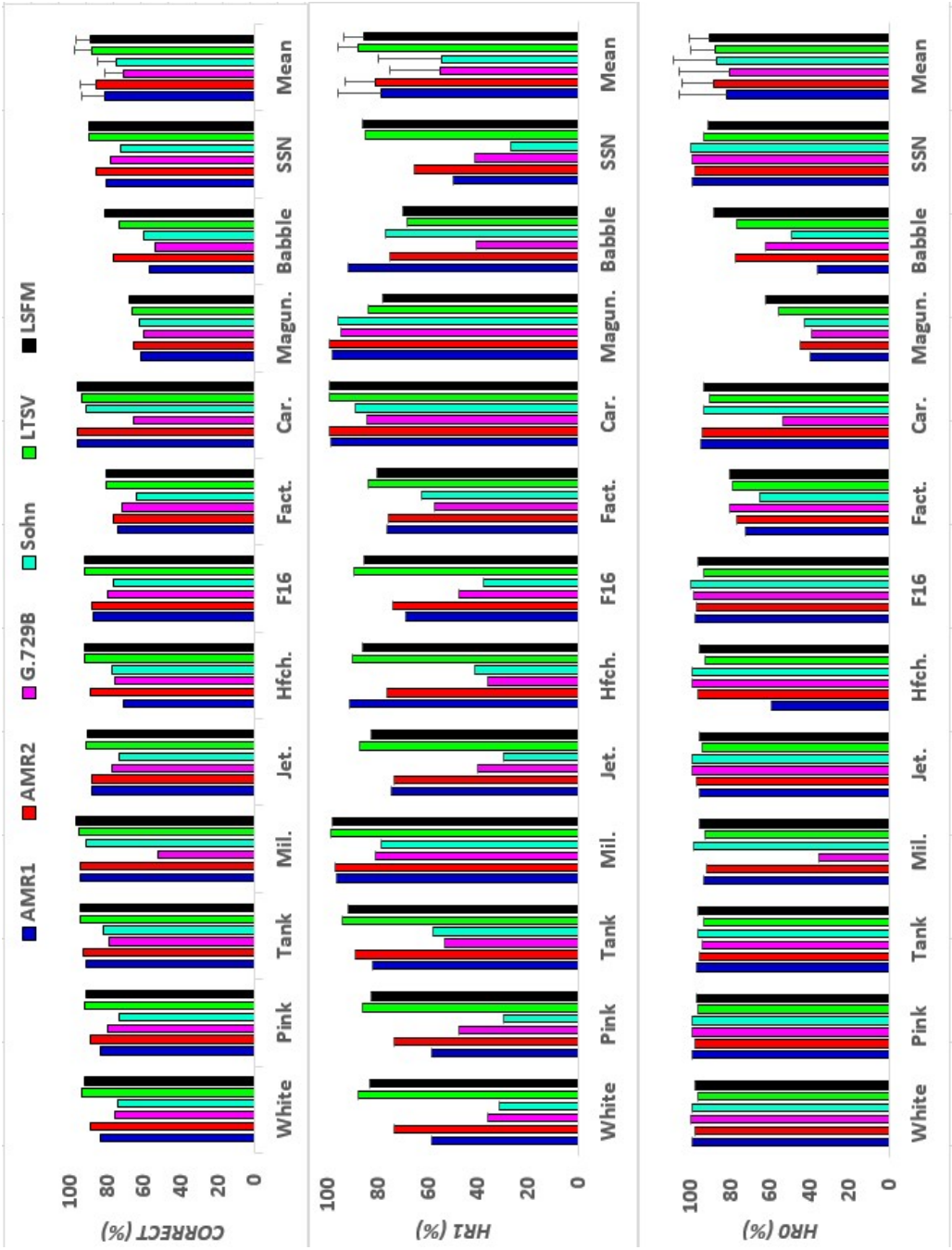


Figure 3.17: Accuracy rate comparisons for six VAD schemes averaged over five SNRs for 12 kinds of noises. Accuracy rate: CORRECT, HR1, and HR0. Five VAD schemes: AMR1, AMR2, G.729B, Sohn, LTSV, and LSFM.

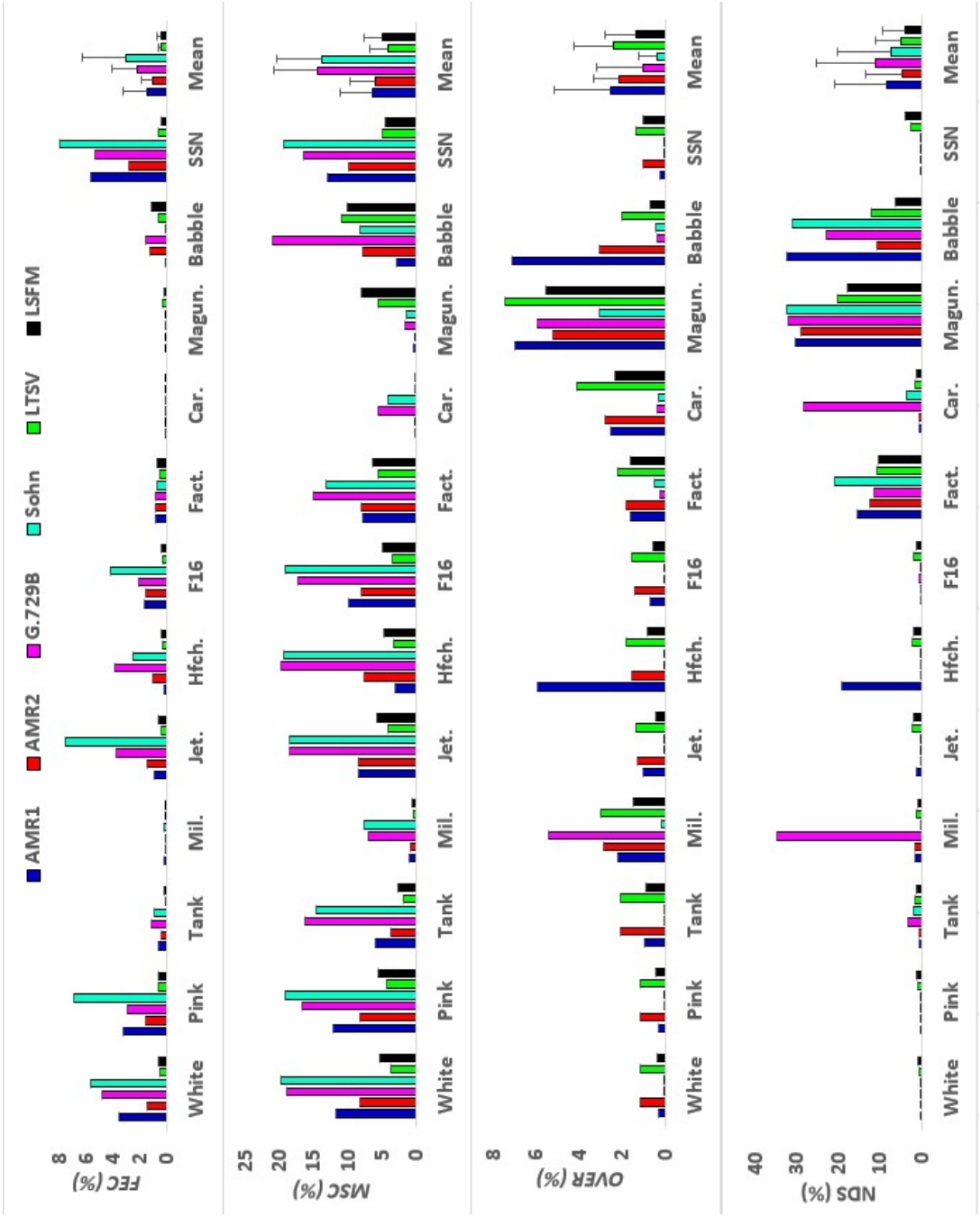


Figure 3.18: Error rate comparison of six VAD schemes averaged over five SNRs for 12 kinds of noises. Error rate: FEC, MSC, OVER, and NDS. Six VAD schemes: AMR1, AMR2, G.729B, Sohn, LTSV, and LSFM.

Chapter 4

A Modified Wiener Filtering Method Combined with Wavelet Thresholding Multitaper Spectrum for Speech Enhancement

Contents

4.1 Introduction

The objective of speech enhancement (SE, also called noise reduction) algorithms is to improve one or more perceptual aspects of the noisy speech by decreasing the background noise without affecting the intelligibility of the speech [2]. Research on SE can be traced back to 40 years ago with 2 patents by Schroeder [56], where an analog implementation of the spectral magnitude subtraction method was described. Since then the problem of enhancing speech degraded by uncorrelated additive noise, when only the noisy speech is available, has become an area of active research [27]. Researchers and engineers have approached this challenging problem by exploiting different properties of speech and noise signals to achieve better performance [57].

SE techniques have a broad range of applications, from hearing aids to mobile communication, voice-controlled systems, multiparty teleconferencing, and automatic speech recognition (ASR) systems [57]. The algorithms can be summarized into four classes: spectral subtractive [20, 58, 59, 60], sub-space [9, 61], statistical model based [62, 8, 63] and Wiener-type [64, 65, 27, 25] algorithms.

Much progress has been made in the development of SE algorithms capable of improving speech quality [3, 4] which was evaluated mainly by the objective performance criteria such as signal-to-noise ratio (SNR) [5]. However, SE algorithm that improves speech quality may not perform well in real-world listening situations where background noise level and characteristics are constantly changing [6]. The first intelligibility study done by Lim [7] in the late 1970s found no intelligibility improvement with the spectral subtraction algorithm for speech corrupted in white noise at -5 to 5 dB SNR. Thirty years later, study conducted by Hu and Loizou [2] found that none of the examined eight different algorithms improved speech intelligibility relative to unprocessed (corrupted) speech. Moreover, according to [2], the algorithms with the highest overall speech quality may not perform the best in terms of speech intelligibility (e.g.,logMMSE [8]). And the algorithm which performs the worst in terms

of overall quality may perform well in terms of preserving speech intelligibility (e.g.,KLT [9]). To our knowledge, very few speech enhancement algorithms [10, 11, 12] claimed to improve speech intelligibility by subjective tests for either normal-hearing listeners or hearing-impaired listeners. Hence, we focused in this chapter on improving performance on speech intelligibility of SE algorithm.

From [5] we know that the perceptual effects of attenuation and amplification distortion on speech intelligibility are not equal. Amplification distortion in excess of 6.02 dB (Region III) bear the most detrimental effect on speech intelligibility while the attenuation distortion (Region I) was found to yield the least effect on intelligibility. Region I+II constraints are the most robust in terms of yielding consistently large benefits in intelligibility independent of the SE algorithm used. However, in order to divide those three Regions [5], the estimated magnitude spectrum needs to be compared with the clean spectrum which we usually don't have in real circumstances.

In this chapter, we explored the multitaper spectrum which was shown in [66] to have good bias and variance properties. The spectral estimate was further refined by wavelet thresholding the log multitaper spectrum in [25]. The refined spectrum was proposed in this chapter to be used as an alternative of the clean spectrum. Then the Region I+II constraints were imposed and incorporated in the derivation of the gain function of the Wiener algorithm based on *a priori* SNR [27]. We have experimentally evaluated its performance under a variety of noise types and SNR conditions.

The structure of the rest of this chapter is organized as follows. Section 4.2 provides background information on wavelet thresholding the multitaper spectrum, and Section 4.3 introduces the amplification and attenuation distortion in details. Section 4.4 presents the proposed approach which imposes constraints on the Wiener filtering gain function. Section 4.5 contains the speech and noise database and metrics used in the evaluation. The simulation results are given in Section 4.6. Finally, a conclusion of this work and the discussion are given in Section 4.7.

4.2 Wavelet Thresholding the Multitaper Spectrum

In real-world scenarios, the background noise level and characteristics are constantly changing [6]. Better estimation of the spectrum is required to alleviate the distortion casued by SE algorithms. For speech enhancement, the most frequently used power spectrum estimator is direct spectrum estimation based on Hann windowing. However, windowing reduces only the bias not the variance of the spectral estimate [67]. The multitaper spectrum estimator [66], on the other hand, can reduce this variance by computing a small number (L) of direct spectrum estimators (eigenspectra) each with a different taper (window), and then averaging the L spectral estimates. The underlying philosophy is similar to Welch's method of modified periodogram [67].

The multitaper spectrum estimator is given by

$$\widehat{S}^{mt}(\omega) = \frac{1}{L}\sum_{k=0}^{L-1}\widehat{S}_k^{mt}(\omega) \tag{4.1}$$

with

$$\widehat{S}_k^{mt}(\omega) = \left|\sum_{m=0}^{N-1} a_k(m)x(m)e^{-j\omega m}\right|^2, \tag{4.2}$$

where N is the data length and a_k is the kth sine taper used for the spectral estimate $\widehat{S}_k^{mt}(\cdot)$, which is proposed by Riedel and Sidorenko [68] and defined by:

$$a_k(m) = \sqrt{\frac{2}{N+1}}sin\frac{\pi k(m+1)}{N+1}, m = 0, \cdots, N-1. \tag{4.3}$$

The sine tapers were proved in [68] to produce smaller local bias than the Slepian tapers, with roughly the same spectral concentration.

The multitaper estimated spectrum can be further refined by wavelet thresholding techniques [69, 70, 71]. Improved periodogram estimates were proposed in [69] and improved multitaper spectrum estimates were proposed in [70, 71]. The underlying idea behind those techniques is to represent the log periodogram as "signal" plus the "noise", where the signal is the true spectrum and "noise" is the estimation error [72]. It was shown in [73] that if the eigenspectra defined

in Eq(4.2) are assumed to be uncorrelated, the ratio of the estimated multitaper spectrum $\widehat{S}^{mt}(\omega)$ and the true power spectrum $S(\omega)$ conforms to a chi-square distribution with $2L$ degrees of freedom, i.e.,

$$\upsilon(\omega) = \frac{\hat{S}^{mt}(\omega)}{S(\omega)} \sim \frac{\chi^2_{2L}}{2L}, \quad 0 < \omega < \pi. \tag{4.4}$$

Taking the log of both sides, we get

$$\log\hat{S}^{mt}(\omega) = \log S(\omega) + \log\upsilon(\omega). \tag{4.5}$$

From Eq(4.5) we know that the log of the multitaper spectrum can be represented as the sum of the true log spectrum plus a $\log\chi^2$ distributed noise term. It follows from Bartlett and Kendall [74] that the distribution of $\log\upsilon(\omega)$ is with mean $\phi(L) - \log(L)$ and variance $\phi'(L)$, where $\phi(\cdot)$ and $\phi'(\cdot)$ denote, respectively, the digamma and trigamma functions. For $L \geq 5$, the distribution of $\log\upsilon(\omega)$ will be close to a normal distribution [75]. Hence, provided L is at least 5, the random variable $\eta(\omega)$

$$\eta(\omega) = \log\upsilon(\omega) - \phi(L) + \log(L) \tag{4.6}$$

will be approximately Gaussian with zero mean and variance $\sigma^2_\eta = \phi'(L)$. If $Z(\omega)$ is defined as

$$Z(\omega) = \log\hat{S}^{mt}(\omega) - \phi(L) + \log(L), \tag{4.7}$$

then we have

$$Z(\omega) = \log S(\omega) + \eta(\omega), \tag{4.8}$$

i.e., the log multitaper power spectrum plus a known constant $(\log(L) - \phi(L))$ can be written as the true log power spectrum plus approximately Gaussian noise $\eta(\omega)$ with zero mean and known variance σ^2_η [70].

The model in Eq(4.8) is well suited for wavelet denoising techniques [76, 77, 78, 79] for eliminating the noise $\eta(\omega)$ and obtaining a better estimate of the log spectrum. The idea behind refining the multitaper spectrum by wavelet thresholding can be summarized into four steps [25].

- Obtain the multitaper spectrum using Eq(4.1) to Eq(4.3), and calculate $Z(w)$ using Eq(4.7).

- Apply a standard periodic Discrete Wavelet Transform (DWT) out to level q_0 to $Z(w)$ to get the empirical DWT coefficients $z_{j,k}$ at each level j, where q_0 is specified in advance [80].

- Apply a thresholding procedure to $z_{j,k}$.

- The inverse DWT is applied to the thresholded wavelet coefficients to obtain the refined log spectrum.

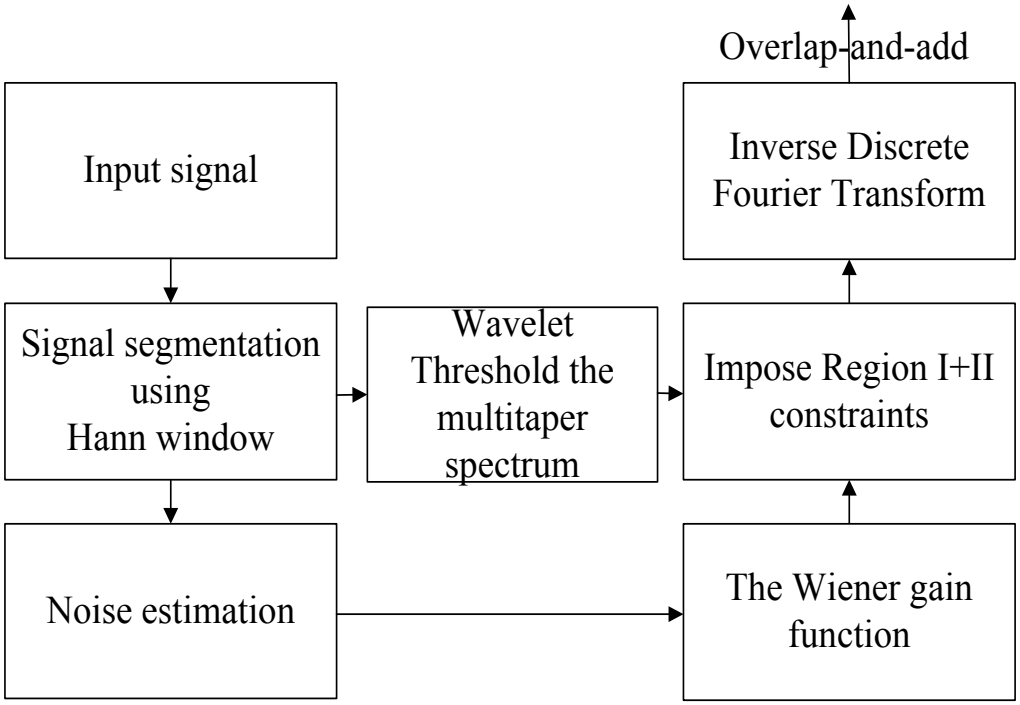


Figure 4.1: Block diagram of the proposed speech enhancement algorithm.

4.3 Amplification distortion and Attenuation distortion

In [5], they mentioned that a positive difference between the clean and estimated spectra would signify attenuation distortion, while a negative spectral difference would signify amplification distortion. The perceptual effect of these two distortions on speech intelligibility cannot be assumed to be equivalent.

Let $\mathrm{SNR}_{\mathrm{ESI}}$ denote the signal-to-residual spectrum ratio at frequency bin k

$$\mathrm{SNR}_{\mathrm{ESI}}(k) = \frac{X^2(k)}{(X(k) - \hat{X}(k))^2} \tag{4.9}$$

where $X(k)$ denotes the clean magnitude spectrum and $\hat{X}(k)$ denotes the magnitude spectrum estimated by a speech-enhancement algorithm. Dividing both numerator and denominator by $D^2(k)$, where $D(k)$ denotes the noise magnitude spectrum, we get

$$\mathrm{SNR}_{\mathrm{ESI}}(k) = \frac{\mathrm{SNR}(k)}{(\sqrt{\mathrm{SNR}(k)} - \sqrt{\mathrm{SNR}_{\mathrm{ENH}}(k)})^2} \tag{4.10}$$

where $\mathrm{SNR}(k) = \frac{X^2(k)}{D^2(k)}$ is the true instantaneous SNR at bin k, and $\mathrm{SNR}_{\mathrm{ENH}}(k) = \frac{\hat{X}^2(k)}{D^2(k)}$ is the enhanced SNR. The correlation of the $\mathrm{SNR}_{\mathrm{ESI}}$ measure with speech intelligibility was found to be 0.81 [81] and the correlation with speech quality was found to be 0.85 [82].

Three regions were divided according to the distortions introduced.

- Region I: $\hat{X}(k) \leq X(k)$, suggesting only attenuation distortion.
- Region II: $X(k) < \hat{X}(k) \leq 2 \cdot X(k)$, suggesting amplification distortion up to 6.02 dB.
- Region III: $\hat{X}(k) > 2 \cdot X(k)$, suggesting amplification distortion of 6.02 dB or greater.

Intelligibility listening tests verified the hypothesis that the estimated magnitude spectra need to be contained in regions I and II in order to maximize the speech intelligibility.

In summary, the SE algorithms need to treat the two types of distortions differently in order to improve speech intelligibility. More specifically, SE algorithms need to be designed so as to minimize the amplification distortions (in excess of 6.02 dB) which were found to bear the most detrimental effects on speech intelligibility.

4.4 Speech enhancement based on constrained Wiener filtering algorithm

Among the numerous techniques that were developed, the Wiener filter can be considered as one of the most fundamental SE approaches, which has been delineated in different forms and adopted in various applications [57]. The Wiener gain function is the least aggressive, in terms of suppression, providing small attenuation even at extremely low SNR levels.

A block diagram of the proposed SE algorithm is shown in Figure 4.1. The initial four frames are assumed to be noise only. The algorithm can be described as follows. The input noisy speech signal is decomposed into frames of 20ms length with an overlap of 10ms by the Hann window. Each segment was transformed using a 160-point discrete Fourier transform (DFT). The spectrum of the segmented noisy and noise signal are estimated by the multitaper method and then further refined by wavelet thresholding technique. The estimated "clean" spectrum was gotten from the refined multitaper estimated noisy and noise spectrum. On the other hand, the noise-corrupted sentences were enhanced by the Wiener algorithm based on *a priori* SNR estimation [27]. The Region I+II constraints were then imposed on the enhanced spectrum. Finally, the inverse FFT was applied to obtain the enhanced speech signal.

The implementation details of the proposed method can be described in the following four steps. For each speech frame:

- Compute the multitaper power spectrum $\widehat{S}_{\mathbf{y}}^{mt}$ of the noisy speech $\mathbf{y}$ using Eq(4.1), and estimate the multitaper power spectrum $\widehat{S}_{\mathbf{x}}^{mt}$ of the clean speech signal by: $\widehat{S}_{\mathbf{x}}^{mt} = \widehat{S}_{\mathbf{y}}^{mt} - \widehat{S}_{\mathbf{n}}^{mt}$, where $\widehat{S}_{\mathbf{n}}^{mt}$ is the multitaper power

spectrum of the noise. $\widehat{S}_{\mathbf{n}}^{mt}$ can be obtained using noise samples collected during speech absent frames. Here L is set to 16. Any negative elements of $\widehat{S}_{\mathbf{x}}^{mt}$ are floored as follows:

$$\widehat{S}_{\mathbf{x}}^{mt} = \begin{cases} \widehat{S}_{\mathbf{y}}^{mt} - \widehat{S}_{\mathbf{n}}^{mt}, & \text{if } \widehat{S}_{\mathbf{y}}^{mt} > \widehat{S}_{\mathbf{n}}^{mt} \\ \beta \widehat{S}_{\mathbf{n}}^{mt}, & \text{if } \widehat{S}_{\mathbf{y}}^{mt} \leq \widehat{S}_{\mathbf{n}}^{mt} \end{cases} \tag{4.11}$$

where β is the spectral floor set to $\beta = 0.002$.

- Compute $Z(\omega) = log\hat{S}_{\mathbf{y}}^{mt}(\omega) - \phi(L) + log(L)$ and then apply the Discrete Wavelet Transform (DWT) (Daubechies 4) of $Z(\omega)$ out to level q_0 to obtain the empirical DWT coefficients $z_{j,k}$ for each level j, where q_0 is specified to be 5 [80]. Threshold the wavelet coefficients $z_{j,k}$ and apply the inverse DWT to the thresholded wavelet coefficients to obtain the refined log spectrum, $log\widehat{S}_{\mathbf{y}}^{\omega mt}(\omega)$, of the noisy singal. Repeat the above procedure to obtain the refined log spectrum, $log\widehat{S}_{\mathbf{n}}^{\omega mt}(\omega)$, of the noise signal. The estimated power spectrum $\widehat{S}_{\mathbf{x}}^{\omega mt}(\omega)$ of the clean speech signal can be estimated using

$$\widehat{S}_{\mathbf{x}}^{\omega mt}(\omega) = \widehat{S}_{\mathbf{y}}^{\omega mt}(\omega) - \widehat{S}_{\mathbf{n}}^{\omega mt}(\omega) \tag{4.12}$$

- Let $Y(\omega, t)$ denote the magnitude of the noisy spectrum at time frame t and frequency bin ω estimated by the method in [83]. Then, the estimate of the signal spectrum magnitude is obtained by multiplying $Y(\omega, t)$ with a gain function $G(\omega, t)$ as: $\hat{X}(\omega, t) = G(\omega, t) \cdot Y(\omega, t)$. The Wiener gain function is based on the *a priori* SNR and is given by

$$G(\omega, t) = \sqrt{\frac{\text{SNR}_{\text{prio}}(\omega, t)}{1 + \text{SNR}_{\text{prio}}(\omega, t)}} \tag{4.13}$$

where SNR_{prio} is the *a priori* SNR estimated using the decision-directed approach [27, 5] as follows:

$$\text{SNR}_{\text{prio}} = \alpha \cdot \frac{X_M^2(\omega, t-1)}{\hat{P}_D^2(\omega, t-1)} + (1-\alpha) \cdot \max\left[\frac{Y^2(\omega, t)}{\hat{P}_D^2(\omega, t)} - 1,\ 0\right] \tag{4.14}$$

where $\hat{P}_D^2(\omega,t)$ is the estimate of the power spectral density of background noise and α is a smoothing constant (typically set to α=0.98).

- To maximize speech intelligibility, the final enhanced spectrum, $X_M(\omega,t)$, can be obtained by utilizing the Region I+II constraints to the enhanced spectrum $\hat{X}(\omega,t)$ as follows:

$$X_M(\omega,t) = \begin{cases} \hat{X}(\omega,t), & \text{if } \hat{X}(\omega,t) < 2\widehat{S}_{\mathbf{x}}^{\omega mt}(\omega) \\ 0 & \text{else} \end{cases} \tag{4.15}$$

  Finally, the enhanced speech signal can be obtained by apply the inverse FFT of $X_M(\omega,t)$.

The above estimator was applied to 20ms duration frames of the noisy signal with 50% overlap between frames. The enhanced speech signal was combined using the overlap and add method.

4.5 Evaluation Setup

The proposed SE algorithm was tested using a speech database that was corrupted by eight different real-world noises at different SNRs. The system was evaluated using both the composite evaluation measures proposed in [84] and the SNR_{LOSS} measure proposed in [85].

4.5.1 Database Description

For the evaluation of SE algorithms, NOIZEUS [30] is preferred since it is a noisy speech corpus recorded by [4] to facilitate comparison of SE algorithms among different research groups [6]. The noisy database contains thirty IEEE sentences [31] which were recorded in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The sentences were produced by three male and three female speakers (5 sentences/speaker). The IEEE database was used as it contains phonetically-balanced sentences with relatively low word-context predictability. The thirty sentences were selected from the IEEE database so

as to include all phonemes in the American English language. The sentences were originally sampled at 25 kHz and downsampled to 8 kHz.

To simulate the receiving frequency characteristics of the telephone handsets, the intermediate Reference System (IRS) filter used in ITU-T P.862 [86] for evaluation of the PESQ measures was independently applied to the clean and noise signals [3]. Then noise segment of the same length as the speech signal was randomly cut out of the noise recordings, appropriately scaled to reach the desired SNR levels (-8 dB, -5 dB, -2 dB, 0 dB, 5 dB, 10 dB, 15 dB) and finally added to the filtered clean speech signal. Noise signals were taken from the AURORA database [32] and included the following recordings from different places: Train, Babble (crowd of people), Car, Exhibition Hall, Restaurant, Street, Airport and Train Station. Therefore, in total there are 1680 (30 sentences$\times$ 8 noises$\times$ 7 SNRs) noisy speech segments in the test set.

4.5.2 Performance Evaluation

Performance of an SE algorithm can be evaluated both subjectively and objectively. In general, subjective listening test is the most accurate and preferable method for evaluating speech quality and intelligibility. However, it is time consuming and cost expensive. Recently, many researchers have placed much effort on developing objective measures that would predict subjective quality and intelligibility with high correlation [84, 82, 85, 81] with subjective listening test. Among them, the composite objective measures [84] were proved to have high correlation with subjective ratings, and at the same time, capture different characteristics of the distortions present in the enhanced signals [75] while the SNR_{LOSS} measure [85] was found appropriate in predicting speech intelligibility in fluctuating noisy conditions by yielding a high correlation for predicting sentence recognition.

Therefore, the composite objective measures and the SNR_{LOSS} measure were adopted to predict the performance of the proposed SE algorithm on subjective quality and speech intelligibility, respectively.

4.5.2.1 The composite measures to predict subjective speech quality

The composite objective measures are obtained by linearly combining existing objective measures that highly correlate with subjective ratings. The objective measures include: segmental SNR ($segSNR$) [4], weighted-slope spectral (WSS) [87], perceptual evaluation of speech quality ($PESQ$) [88] and log likelihood ration (LLR) [4].

Table 4.1: Scale of signal distortion C_{sig}, background intrusiveness C_{bak} and overall quality C_{ovl}

Scale of signal distortion	Scale of background intrusiveness
5- Very natural, no degradation	5- Not noticeable
4- Fairly natural, little degradation	4- Somewhat noticeable
3- Somewhat natural, somewhat degraded	3- Noticeable but not intrusive
2- Fairly unnatural, fairly degraded	2- Fairly conspicuous, somewhat intrusive
1- Very unnatural, very degraded	1- Very conspicuous, very intrusive

Scale of overall quality C_{ovl}
5- Excellent
4- Good
3- Fair
2- Poor
1-Bad

The three new composite measures obtained from multiple linear regression analysis are given below:

- C_{sig}: A five-point scale of signal distortion (SIG) formed by linearly combining the LLR, $PESQ$, and WSS measures (Table 4.1).

- C_{bak}: A five-point scale of noise intrusiveness (BAK) formed by linearly combining the $segSNR$, $PESQ$, and WSS measures (Table 4.1).

- C_{ovl}: The mean opinion score of overall quality (OVRL) formed by linearly combining the $PESQ$, LLR, and WSS measures.

The three new composite measures obtained from multiple linear regression analysis are given below:

$$C_{sig} = 3.093 - 1.029 \cdot LLR + 0.603 \cdot PESQ - 0.009 \cdot WSS \tag{4.16}$$

$$C_{bak} = 1.634 + 0.478 \cdot PESQ - 0.007 \cdot WSS + 0.063 \cdot segSNR \tag{4.17}$$

$$C_{ovl} = 1.594 + 0.805 \cdot PESQ - 0.512 \cdot LLR - 0.007 \cdot WSS \tag{4.18}$$

The correlation coefficients between the three composite measures and real subjective measures are given in Table 4.2 [84]. All three parameters should be maximized in order to get the best performance.

Table 4.2: Correlation coefficients between the composite measures and subjective measure

	C_{sig}	C_{bak}	C_{ovl}
SIG	0.7		
BAK		0.58	
OVRL			0.73

4.5.2.2 The SNR_{LOSS} measure to predict speech intelligibility

The SNR loss in band j and frame m is defined as follows [85]:

$$SL(j,m) = \text{SNR}_X(j,m) - \text{SNR}_{\hat{X}}(j,m) \tag{4.19}$$

where $\text{SNR}_X(j,m)$ is the input SNR in band j, $\text{SNR}_{\hat{X}}(j,m)$ is the SNR of the enhanced signal in the j_{th} frequency band at the m_{th} frame.

Assuming the SNR range is restricted to $[-\text{SNR}_{\text{Lim}}, \text{SNR}_{\text{Lim}}]$ dB (SNR_{Lim}=3 in this chapter), the $SL(j,m)$ term is then limited as follows:

$$\hat{SL}(j,m) = \min(\max(SL(j,m), -\text{SNR}_{\text{Lim}}), \text{SNR}_{\text{Lim}}) \tag{4.20}$$

and subsequently mapped to the range of [0, 1] using the following equation:

$$\mathrm{SNR}_{\mathrm{LOSS}}(j,m) = \begin{cases} -\frac{C_-}{\mathrm{SNR}_{\mathrm{Lim}}}\hat{SL}(j,m), & \text{if } \hat{SL}(j,m) < 0 \\ \frac{C_+}{\mathrm{SNR}_{\mathrm{Lim}}}\hat{SL}(j,m), & \text{if } \hat{SL}(j,m) \geq 0 \end{cases} \tag{4.21}$$

where C_- and C_+ are parameters (fixed to be 1 in this chapter) controlling the slopes of the mapping function which was defined in the range of [0, 1], therefore, the frame $\mathrm{SNR}_{\mathrm{LOSS}}$ is normalized to the range of $0 \leq \mathrm{SNR}_{\mathrm{LOSS}}(j,m) \leq 1$. The average $\mathrm{SNR}_{\mathrm{LOSS}}$ is finally computed by averaging $\mathrm{SNR}_{\mathrm{LOSS}}(j,m)$ over all frames in the signal as follows:

$$\overline{\mathrm{SNR}}_{\mathrm{LOSS}} = \frac{1}{M}\sum_{m=0}^{M-1} f\mathrm{SNR}_{\mathrm{LOSS}}(m) \tag{4.22}$$

where M is the total number of data segments in the signal and $f\mathrm{SNR}_{\mathrm{LOSS}}(m)$ is the average (across bands) SNR loss computed as follows:

$$f\mathrm{SNR}_{\mathrm{LOSS}}(m) = \frac{\sum_{j=1}^{K} W(j)\cdot \mathrm{SNR}_{\mathrm{LOSS}}(j,m)}{\sum_{j=1}^{K} W(j)} \tag{4.23}$$

where $W(j)$ is the weight (i.e., band importance function [89]) placed on the j_{th} frequency band and was taken from Table B.1 in the ANSI standard [89].

The implementation of the $\mathrm{SNR}_{\mathrm{LOSS}}$ measure was supplied in the website[1] of the authors in [85]. The smaller the value of the $\mathrm{SNR}_{\mathrm{LOSS}}$ measure is, the better performance of the SE algorithm is achieved. The correlation with real subjective test on speech recognition was 0.82 [85].

4.6 Simulation Results

The evaluation of the subjective quality and intelligibility of the speech enhanced by our proposed SE algorithm are reported in this section. Three other SE schemes, namely wavelet thresholding (WT) [25], KLT [9] and Wiener algorithm with clean signal present (Wiener_Clean) [5], were also evaluated in

[1] http://ecs.utdallas.edu/loizou/cimplants/cpubs.htm

order to gain a comparative analysis of the proposed SE algorithm. The KLT algorithm was proved in [2] and [10] by subjective tests to perform well in terms of preserving speech intelligibility for normal hearing listeners, and improving speech intelligibility significantly for cochlear implant users in regards to recognition of sentences corrupted by stationary noises, respectively. The Wiener_Clean algorithm was taken as the ground truth in this chapter because there is clean signal used in the algorithm. The unprocessed noisy signal (UP) were also evaluated by the SNR_{LOSS} measure for comparison purpose. The implementations of these three schemes were taken from the implementations in [4].

4.6.1 Performance of predicting subjective quality

4.6.1.1 Performance average over all eight kinds of noise

In Figure 4.2, the proposed algorithm is compared with WT and KLT algorithms in terms of the composite measures averaging over all eight noises for seven SNRs. The four objective measures (*LLR*, *segSNR*, *WSS* and *PESQ*) that composed of the composite measures were also given in the first row for reference. The Wiener_Clean algorithm, as the ground truth, performed the best for all four objective evaluation measures. According to [84], the *LLR* measure performed the best in terms of predicting signal distortion while the *PESQ* measure gave the best prediction for both noise intrusiveness and overall speech quality. From the first row of Figure 4.2 we can notice that our proposed algorithm gives better performance than both WT and KLT in terms of the *LLR* measure for all seven SNRs tested. Moreover, when SNR is smaller than 5dB, our proposed algorithm also performed better than both WT and KLT for the *PESQ* measure.

The second row of Figure 4.2 shows the composite measures, which include C_{sig}, C_{bak} and C_{ovl}, estimated by the combination of all those four objective measures expressed in the first row. In terms of both signal distortion C_{sig} and overall quality C_{ovl}, our proposed method performs the best when SNR is less

than 10dB. Specifically speaking, for overall quality measure C_{ovl}, the proposed algorithm improved 10.94%, 18.94%, 21.63%, 23.66%, and 6.67% for -8dB, -5dB, -2dB, 0dB and 5dB, respectively when compared with the KLT method. In general, the proposed algorithm achieved 13.88% and 6.40% improvement for C_{sig} and C_{ovl} when average over all seven tested SNR levels. However, for C_{bak}, the WT and KLT algorithms give similar and better results than our proposed one when SNR is no smaller than 0dB. The improvement was 0.98%, 6.98%, 11.11% and 16.55% for 0dB, 5dB, 10dB and 15dB respectively. In average, the WT and KLT methods were 5.14% better than our proposed algorithm in terms of background intrusiveness C_{bak}.

The composite measure comparisons for four SE schemes (WT, KLT, Proposed and Wiener_Clean) averaged over seven SNRs (-8dB, -5dB, -2dB, 0dB, 5dB, 10dB, 15dB) for eight kinds of noise.

4.6.1.2 Performance average over seven SNRs

Figure 4.3 shows the three different composite measures averaged over seven SNRs for eight kinds of noises computed for WT, KLT, Wiener_Clean and Proposed SE algorithms. The Wiener_Clean algorithm still works as the ground truth here. From Figure 4.3, it is clear that in terms of C_{sig}, the KLT works much better than that of WT. Hence, the proposed algorithm is compared with only the KLT method here. We observe that on average, the proposed algorithm is better than the KLT method in terms of C_{sig} for Train (9.19%), Babble (15.93%), Car (14.51%), Exhibition Hall (7.74%), Restaurant (16.74%), Street (13.42%), Airport (16.64%) and Train Station (16.23%) noises. The number in the bracket indicates the C_{sig} by which our proposed algorithm is better than the KLT method. The mean C_{sig} over all eight noise types of our proposed SE algorithm is 13.88% better than that of the KLT method. Furthermore, the proposed SE algorithms outperforms the KLT in terms of C_{ovl} by an average of 6.40% over all eight kinds of noise that were considered. However, in terms of background intrusiveness C_{bak}, the KLT algorithm gives an average of 5.14% better results than our proposed algorithm.

Thus, in conclusion, the proposed SE algorithm was predicted to be able to achieve the best overall subjective quality for most SNRs and all noise types considered when comparing with WT and KLT algorithms.

4.6.2 Performance of predicting speech intelligibility

The $\mathrm{SNR}_{\mathrm{LOSS}}$ measure values obtained from each algorithm (include UP) were subjected to statistical analysis in order to assess their significance differences. A highly significant effect (p <0.005) was found in all SNR levels and all types of noise by analysis of variance (ANOVA). Following the ANOVA, multiple comparison statistical tests according to Tukey's HSD test were done to assess significance between algorithms. The difference was deemed significant if the p value was smaller than 0.05.

Figure 4.4 gives the $\mathrm{SNR}_{\mathrm{LOSS}}$ measure value under seven SNRs for eight kinds of noises computed for UP, WT, KLT, Wiener_Clean and Proposed SE algorithms. It is easy to conclude that our proposed SE algorithm gave small $\mathrm{SNR}_{\mathrm{LOSS}}$ measure value and better performance than UP, WT and KLT for all eight noises when SNR smaller than 5dB.

Table 4.3 and Table 4.4 gives the statistical comparisons of the $\mathrm{SNR}_{\mathrm{LOSS}}$ measure between unprocessed noisy sentences (UP) and enhanced sentences by four SE algorithms (WT, KLT, Wiener_Clean and Proposed). At the same time, the comparisons between our proposed SE algorithm and the other three algorithms were also given in Table 4.5 and Table 4.6. From Table 4.3 and Table 4.4 we know that when compared with the UP, our proposed algorithm was predicted by the $\mathrm{SNR}_{\mathrm{LOSS}}$ measure to be able to improve the intelligibility in low SNRs for most noises tested (marked in green color). The R in the table gives the percentage by which our algorithm is better than others, the value is minus because better performance gave smaller $\mathrm{SNR}_{\mathrm{LOSS}}$ measure. Furthermore, our proposed SE algorithm was also compared with the WT and KLT algorithms in Table 4.5 to Table 4.6 and was proved to supply better performance for most conditions tested.

Table 4.3: Statistical comparisons of the SNR_{LOSS} measure between unprocessed noisy sentences and enhanced sentences by four SE algorithms (WT, KLT, Wiener_Clean (Clean) and Proposed) for Train, Babble, Car and Exhibition Hall noises.

Comparisons		WT-UP		KLT-UP		Clean-UP		Proposed-UP	
Noise	SNR(dB)	R(%)	p-value	R(%)	p-value	R(%)	p-value	R(%)	p-value
Train	-8	1.30	.000	0.98	.011	-7.05	.000	-0.87	.031
	-5	1.61	.000	1.26	.004	-6.87	.000	-0.91	.076
	-2	1.93	.000	1.43	.005	-6.74	.000	-0.96	.127
	0	3.41	.000	2.33	.002	-9.83	.000	-1.67	.054
	5	3.44	.001	2.12	.084	-8.95	.000	0.45	.983
	10	2.96	.024	2.47	.091	-7.06	.000	3.74	.002
	15	3.54	.034	4.96	.001	-1.85	.553	10.76	.000
Babble	-8	1.61	.000	0.73	.128	-7.69	.000	-1.16	.002
	-5	2.25	.000	1.21	.010	-7.23	.000	-1.11	.022
	-2	2.42	.000	1.22	.047	-6.90	.000	-1.06	.118
	0	3.25	.000	1.55	.046	-9.57	.000	-1.85	.009
	5	4.88	.000	2.76	.002	-8.06	.000	0.10	1.000
	10	6.11	.000	4.74	.000	-4.76	.000	4.72	.000
	15	6.84	.000	7.72	.000	1.91	.687	12.76	.000
Car	-8	1.18	.000	0.12	.984	-8.38	.000	-1.89	.000
	-5	1.90	.000	0.46	.424	-8.20	.000	-2.11	.000
	-2	2.25	.000	0.56	.321	-8.20	.000	-2.48	.000
	0	3.38	.000	0.78	.373	-11.23	.000	-3.45	.000
	5	3.46	.000	-0.21	.997	-10.70	.000	-2.33	.001
	10	5.32	.000	1.90	.345	-7.52	.000	1.55	.554
	15	5.10	.000	2.78	.075	-2.68	.093	8.25	.000
Exh.Hall	-8	0.58	.026	-0.42	.201	-7.79	.000	-1.44	.000
	-5	1.01	.000	-0.17	.947	-7.70	.000	-1.61	.000
	-2	1.41	.000	0.03	1.000	-7.48	.000	-1.83	.000
	0	4.00	.000	0.77	.738	-10.52	.000	-1.95	.020
	5	3.68	.000	-0.62	.889	-9.69	.000	0.15	.999
	10	4.86	.001	2.14	.393	-6.57	.000	4.34	.004
	15	6.17	.000	4.88	.003	-1.24	.880	11.13	.000

Table 4.4: Statistical comparisons of the SNR_{LOSS} measure between unprocessed noisy sentences and enhanced sentences by four SE algorithms (WT, KLT, Wiener_Clean (Clean) and Proposed) for Restaurant, Street, Airport and Train Station noises.

Comparisons		WT-UP		KLT-UP		Clean-UP		Proposed-UP	
Noise	SNR(dB)	$R(\%)$	p-value	$R(\%)$	p-value	$R(\%)$	p-value	$R(\%)$	p-value
Restaurant	-8	1.81	.000	0.95	.033	-7.42	.000	-1.14	.005
	-5	2.15	.000	1.10	.044	-7.09	.000	-1.07	.053
	-2	2.18	.000	1.12	.113	-6.67	.000	-1.05	.159
	0	4.76	.000	3.21	.000	-8.82	.000	-0.88	.746
	5	4.06	.002	2.53	.142	-7.90	.000	0.83	.939
	10	7.40	.000	6.27	.000	-3.21	.053	6.53	.000
	15	8.27	.000	9.01	.000	-3.77	.101	15.18	.000
Street	-8	1.11	.045	0.72	.363	-7.26	.000	-1.27	.015
	-5	1.48	.017	0.89	.329	-7.24	.000	-1.32	.044
	-2	1.58	.030	0.86	.502	-7.12	.000	-1.36	.088
	0	4.76	.000	2.90	.004	-9.14	.000	-1.13	.622
	5	4.64	.000	2.85	.036	-8.59	.000	0.74	.944
	10	6.28	.001	5.29	.007	-5.61	.003	5.36	.006
	15	7.55	.000	7.49	.000	0.47	.998	12.22	.000
Airport	-8	1.80	.000	0.96	.147	-7.89	.000	-1.51	.004
	-5	2.31	.000	1.37	.048	-7.42	.000	-1.37	.048
	-2	2.65	.000	1.60	.035	-6.78	.000	-1.13	.249
	0	4.53	.000	1.72	.225	-9.43	.000	-1.45	.392
	5	5.26	.000	3.36	.007	-7.69	.000	0.60	.972
	10	7.72	.000	5.88	.000	-2.90	.113	6.58	.000
	15	7.76	.000	7.43	.000	2.61	.270	13.81	.000
Train Station	-8	1.67	.000	0.89	.034	-8.37	.000	-1.91	.000
	-5	2.30	.000	1.22	.010	-7.89	.000	-1.90	.000
	-2	2.64	.000	1.47	.006	-7.53	.000	-1.88	.000
	0	3.19	.000	0.58	.857	-10.72	.000	-2.97	.000
	5	4.25	.003	1.24	.812	-9.47	.000	-1.63	.608
	10	5.43	.000	2.38	.192	-5.94	.000	2.79	.082
	15	7.60	.000	6.30	.000	0.38	.999	11.19	.000

Table 4.5: Statistical comparisons of the SNR_{LOSS} measure between sentences enhanced by our proposed algorithm and unprocessed noise sentences (UP) and enhanced sentences by three other SE algorithms (WT, KLT, and Wiener_Clean (Clean)) for Train, Babble, Car and Exhibition Hall noises.

Comparisons		Proposed-UP		Proposed-WT		Proposed-KLT		Proposed-Clean	
Noise	SNR(dB)	R(%)	p-value	R(%)	p-value	R(%)	p-value	R(%)	p-value
Train	-8	-0.87	.031	-2.15	.000	-1.83	.000	6.64	.000
	-5	-0.91	.076	-2.47	.000	-2.13	.000	6.41	.000
	-2	-0.96	.127	-2.83	.000	-2.35	.000	6.20	.000
	0	-1.67	.054	-4.92	.000	-3.92	.000	9.04	.000
	5	0.45	.983	-2.89	.004	-1.63	.264	10.32	.000
	10	3.74	.002	0.76	.930	1.25	.687	11.63	.000
	15	10.76	.000	6.97	.000	5.52	.000	12.85	.000
Babble	-8	-1.16	.002	-2.73	.000	-1.88	.000	7.07	.000
	-5	-1.11	.022	-3.28	.000	-2.30	.000	6.60	.000
	-2	-1.06	.118	-3.39	.000	-2.25	.000	6.28	.000
	0	-1.85	.009	-4.93	.000	-3.34	.000	8.54	.000
	5	0.10	1.000	-4.56	.000	-2.59	.004	8.87	.000
	10	4.72	.000	-1.31	.684	-0.02	1.000	9.95	.000
	15	12.76	.000	5.54	.001	4.67	.006	10.65	.000
Car	-8	-1.89	.000	-3.04	.000	-2.01	.000	7.09	.000
	-5	-2.11	.000	-3.94	.000	-2.56	.000	6.64	.000
	-2	-2.48	.000	-4.62	.000	-3.03	.000	6.23	.000
	0	-3.45	.000	-6.61	.000	-4.19	.000	8.77	.000
	5	-2.33	.001	-5.59	.000	-2.12	.005	9.38	.000
	10	1.55	.554	-3.58	.003	-0.34	.997	9.81	.000
	15	8.25	.000	2.99	.030	5.32	.000	11.22	.000
Exh.Hall	-8	-1.44	.000	-2.01	.000	-1.03	.000	6.88	.000
	-5	-1.61	.000	-2.60	.000	-1.45	.000	6.60	.000
	-2	-1.83	.000	-3.20	.000	-1.86	.000	6.10	.000
	0	-1.95	.020	-5.72	.000	-2.70	.000	9.58	.000
	5	0.15	.999	-3.40	.000	0.77	.779	10.90	000
	10	4.34	.004	-0.49	.993	2.15	.365	11.67	.000
	15	11.13	.000	4.67	.002	5.96	.000	12.53	.000

Table 4.6: Statistical comparisons of the SNR_{LOSS} measure between sentences enhanced by our proposed algorithm and unprocessed noise sentences (UP) and enhanced sentences by three other SE algorithms (WT, KLT, and Wiener_Clean (Clean)) for Restaurant, Street, Airport and Train Station noises.

Comparisons		Proposed-UP		Proposed-WT		Proposed-KLT		Proposed-Clean	
Noise	SNR(dB)	R(%)	p-value	R(%)	p-value	R(%)	p-value	R(%)	p-value
Restaurant	-8	-1.14	.005	-2.90	.000	-2.07	.000	6.78	.000
	-5	-1.07	.053	-3.16	.000	-2.15	.000	6.47	.000
	-2	-1.05	.159	-3.16	.000	-2.14	.000	6.03	.000
	0	-0.88	.746	-5.38	.000	-3.96	.000	8.70	.000
	5	0.83	.939	-3.10	.028	-1.65	.525	9.49	.000
	10	6.53	.000	-0.81	.945	0.24	1.000	10.06	.000
	15	15.18	.000	6.38	.000	5.67	.001	11.00	.000
Street	-8	-1.27	.015	-2.35	.000	-1.98	.000	6.46	.000
	-5	-1.32	.044	-2.76	.000	-2.19	.000	6.39	.000
	-2	-1.36	.088	-2.90	.000	-2.20	.001	6.21	.000
	0	-1.13	.622	-5.62	.000	-3.91	.000	8.82	.000
	5	0.74	.944	-3.73	.001	-2.05	.210	10.20	.000
	10	5.36	.006	-0.86	.976	0.07	1.000	11.62	.000
	15	12.22	.000	4.34	.014	4.41	.012	11.70	.000
Airport	-8	-1.51	.004	-3.25	.000	-2.44	.000	6.93	.000
	-5	-1.37	.048	-3.60	.000	-2.70	.000	6.53	.000
	-2	-1.13	.249	-3.68	.000	-2.68	.000	6.06	.000
	0	-1.45	.392	-5.72	.000	-3.12	.001	8.81	.000
	5	0.60	.972	-4.42	.000	-2.67	.042	8.99	.000
	10	6.58	.000	-1.05	.876	0.66	.977	9.76	.000
	15	13.81	.000	5.62	.000	5.95	.000	10.92	.000
Train Station	-8	-1.91	.000	-3.52	.000	-2.78	.000	7.04	.000
	-5	-1.90	.000	-4.11	.000	-3.09	.000	6.50	.000
	-2	-1.88	.000	-4.41	.000	-3.30	.000	6.11	.000
	0	-2.97	.000	-5.98	.000	-3.54	.000	8.67	.000
	5	-1.63	.608	-5.64	.000	-2.84	.092	8.65	.000
	10	2.79	.082	-2.50	.115	0.41	.995	9.29	.082
	15	11.19	.000	3.33	.065	4.60	.004	10.76	.000

4.7 Conclusions and Discussions

The main contribution of this chapter was the introduction of a new SE algorithm based on imposing constraint on Wiener gain function. Experiments were done on NOIZEUS database for eight kinds of noise (AURORA database) across seven different SNRs ranging from -8dB to 15dB. The Wiener_Clean algorithm was taken as the ground truth. The performance of our proposed algorithm was compared with WT and KLT methods. The results were analyzed mainly by three composite measures and the SNR_{LOSS} measure to predict the performance on subjective quality and speech intelligibility, respectively. Through extensive experiments, we showed that when averaged over all eight kinds of noises, our proposed SE algorithm achieved the best results in terms of predicting signal distortion C_{sig} and overall quality C_{ovl} when SNR is no more than 10dB. Furthermore, we investigated the individual performance on each noise type. Our proposed SE algorithm outperformed the KLT algorithm for all noise types tested in terms of both C_{sig} and C_{ovl}. On the other hand, the SNR_{LOSS} measure comparisons with both the UP and other SE algorithms predicted that our proposed algorithm was able to improve speech intelligibility for low SNR levels and outperform WT and KLT algorithms for most conditions examined.

It is important to point out that the three composite measures and the SNR_{LOSS} measure used in this chapter are adopted for predicting the subjective quality and intelligibility of noisy speech enhanced by noise suppression algorithms because of their high correlation with real subjective tests [84, 85]. Further subjective tests on both normal-hearing listeners and hearing impaired listeners are needed to verify the effectiveness of the proposed algorithm on improving both subjective quality and speech intelligibility. It is also worth mentioning that depending on the nature of the application, some practical SE systems may require very high quality speech, but can tolerate a certain amount of noise while other systems may want speech as clean as possible even with some degree of speech distortion. Therefore, it should be noted that according

to different applications, different SE algorithms should be chosen to meet the variant requirement.

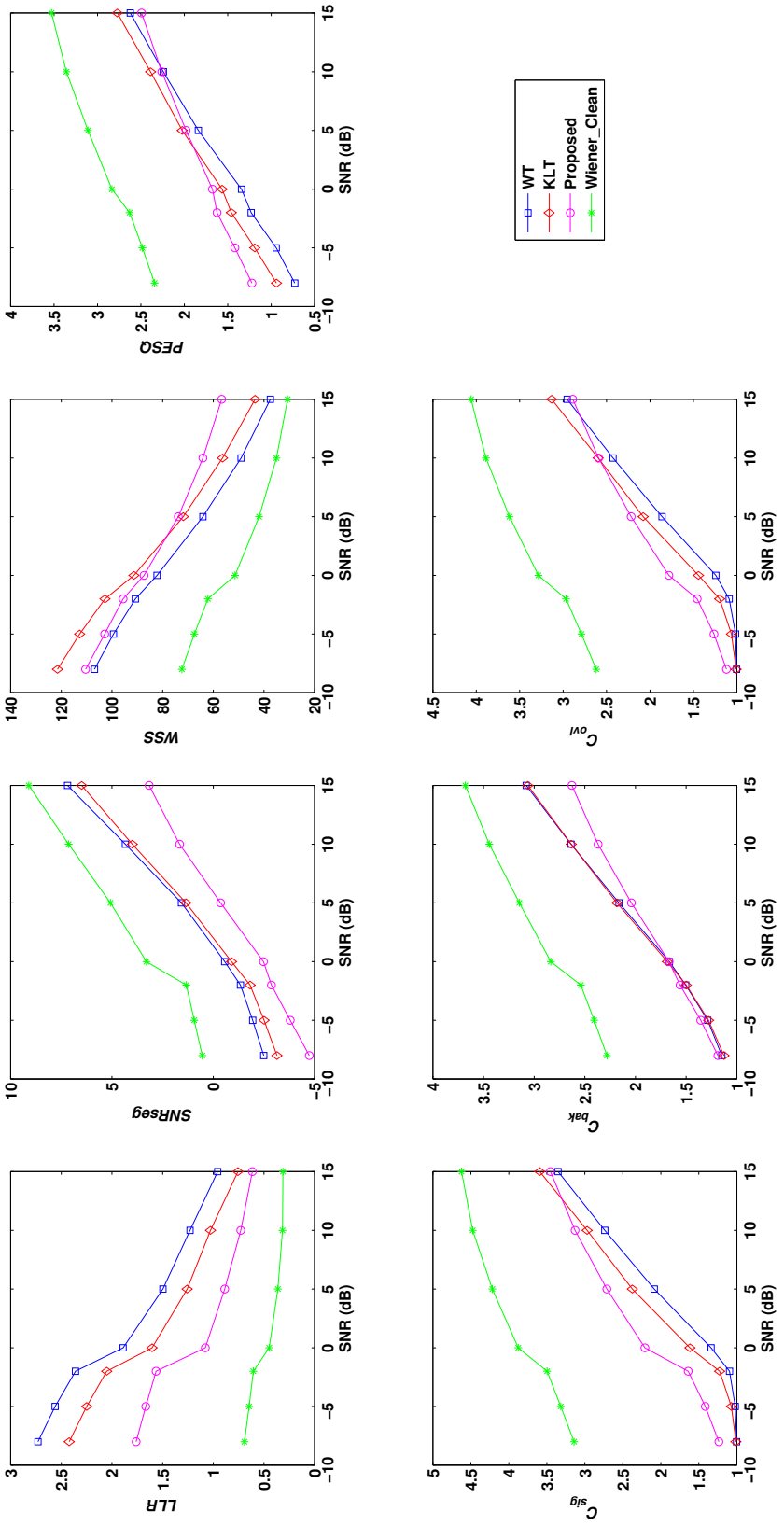


Figure 4.2: The composite measure comparisons for four SE schemes (WT, KLT, Proposed and Wiener_Clean) averaged over seven SNRs (-8dB, -5dB, -2dB, 0dB, 5dB, 10dB, 15dB) for eight kinds of noise.

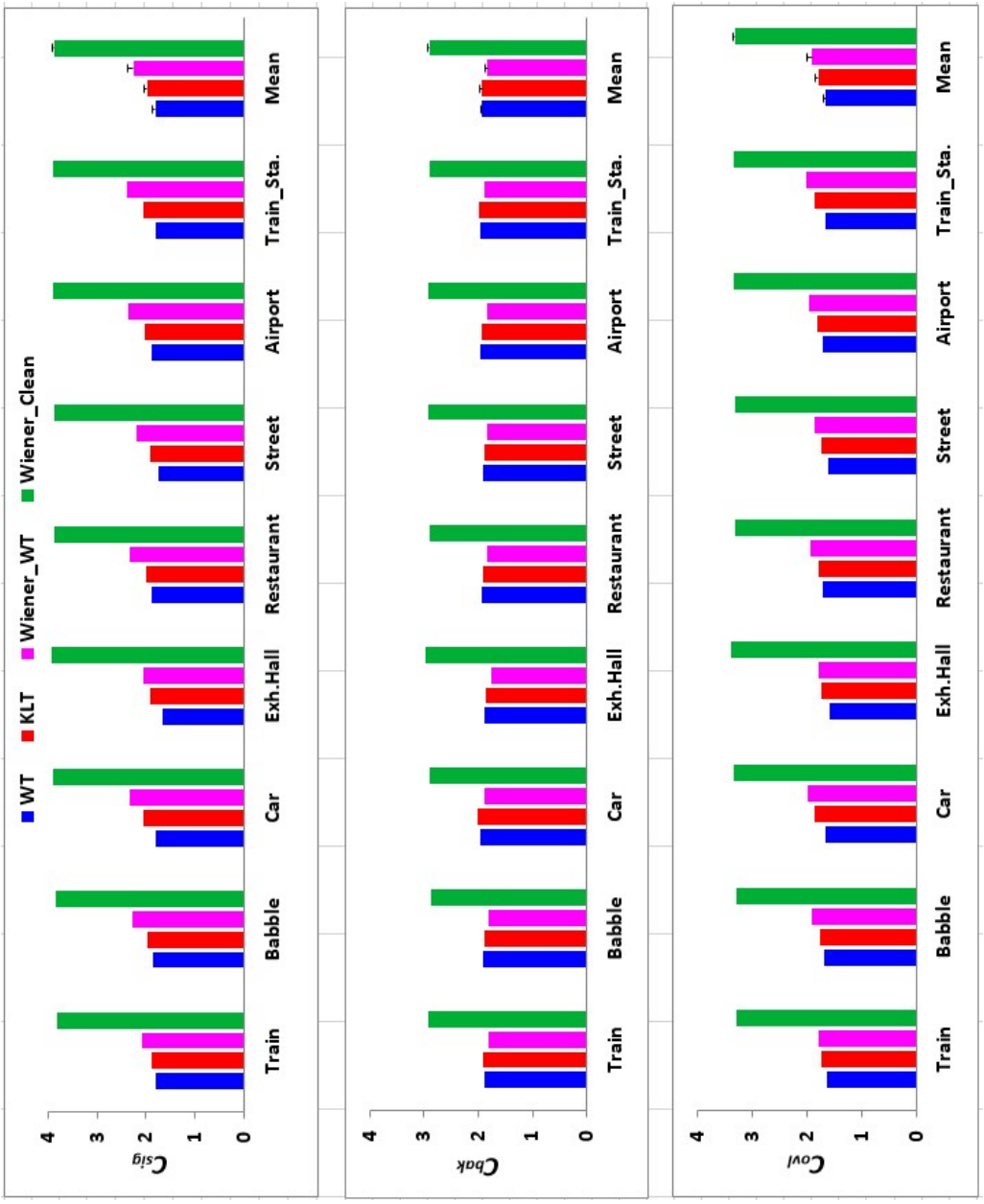


Figure 4.3: The composite measure comparisons for four SE schemes (WT, KLT, Proposed and Wiener_Clean) averaged over eight kinds of noise for seven SNRs (-8dB, -5dB, -2dB, 0dB, 5dB, 10dB, 15dB).

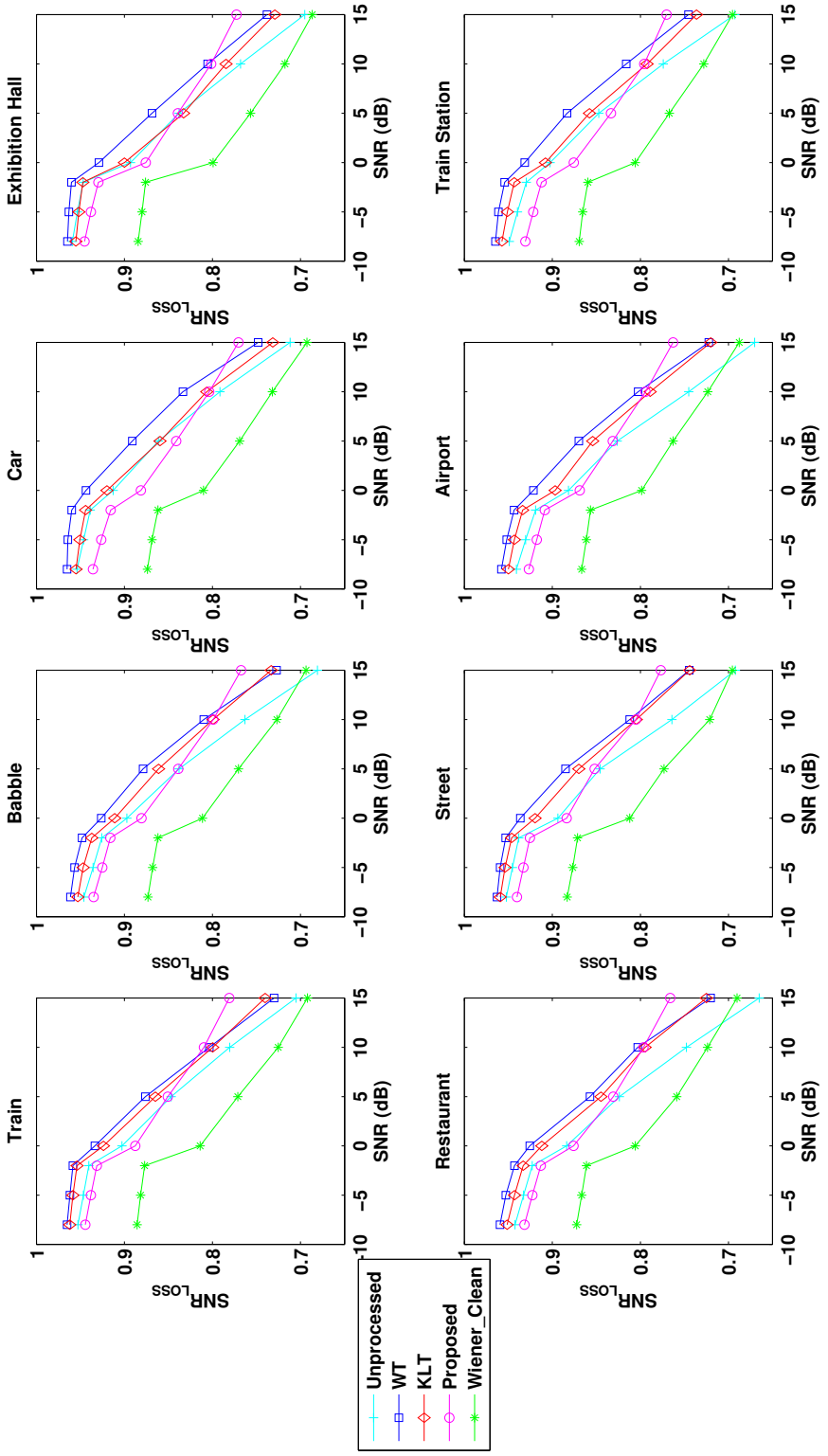


Figure 4.4: The SNR_{LOSS} measure comparisons for the unprocessed noisy (UP) sentences and the enhanced sentences by four SE schemes (WT, KLT, Proposed and Wiener_Clean) under seven SNRs (-8dB, -5dB, -2dB, 0dB, 5dB, 10dB, 15dB) for eight kinds of noise.

Chapter 5

Conclusion and Discussions

Contents

The main contribution of this dissertation was the introduction of a novel long-term spectral flatness measure-based VAD algorithm and a new and efficient SE algorithm based on imposing constraint on Wiener gain function.

5.1 VAD

Our main contributions for VAD algorithm are as follows:

- The effectiveness of the flatness measure along time frames by using a long window was clarified by the LSFM feature distributions as a function of the long-term window length R.
- The proposed LSFM-based VAD algorithm was explained in details and the effect of the M and R combination on the misclassification error was also presented.
- The simulation results of our proposed method proved its robustness by comparing with the three standards and an emerging LTSV-based VAD algorithm in terms of both accuracy rate and error rate.

The proposed VAD algorithm proved its discriminating ability. However, there are several things to be noticed. Firstly, the test database used in the

implementations was created to simulate typical conversational speech by inserting 2-s silence before and after each utterance from core TIMIT test corpus so that the ratio of speech to non-speech was almost 40% to 60%. While this simulates a conversational speech statistically, this is not very realistic in terms of randomness of pauses, hesitations, etc.

Furthermore, depending on the choice of the long-term window length (R and M combination), the LSFM-based VAD application is expected to suffer a delay equal to the duration of the window ($R + M - 1$ frames). Therefore, a trade-off between the delay and robustness of VAD should be carefully considered before utilizing the proposed LSFM-based VAD algorithm.

Moreover, it is worth mentioning that there is a trade-off between HR1 and HR0. The increase of one may lead to a decrease of the other. Therefore, it should be noted that according to different applications, different (R, M) combinations and thresholds for voting scheme can be chosen to meet the variant requirement for HR1 and HR0. For example, HR1 is a crucial factor for speech coding, while high HR0 rate is necessary for most speech recognition-oriented systems.

5.2 SE

As for SE algorithm, our main contributions are as follows:

- The different perceptual effects of attenuation and amplification distortion on speech intelligibility and wavelet thresholded multitaper spectrum were introduced in details.

- The proposed SE algorithm based on constrained Wiener filtering algorithm was presented and the performance of this algorithm on improving subjective quality and speech intelligibility were predicted by the composite measures and the $\mathrm{SNR}_{\mathrm{LOSS}}$ measure, respectively.

- The performance of our proposed algorithm was compared with unprocessed noisy speech (UP), WT and KLT methods on NOIZEUS database

for eight kinds of noise (AURORA database) across seven different SNRs ranging from -8dB to 15dB. The results predicted that our proposed algorithm was bale to improve speech intelligibility for low SNR levels and outperform WT and KLT algorithms for most conditions examined.

It should be noticed that the three composite measures and the $\mathrm{SNR}_{\mathrm{LOSS}}$ measure used in this thesis are adopted for predicting the subjective quality and intelligibility of noisy speech enhanced by noise suppression algorithms because of their high correlation with real subjective tests [84, 85]. Further subjective tests on both normal-hearing listeners and hearing impaired listeners are needed to verify the effectiveness of the proposed algorithm on improving both subjective quality and speech intelligibility.

Bibliography

[1] Yu, R.: A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference On, pp. 4421–4424 (2009). doi:10.1109/ICASSP.2009.4960610 2

[2] Hu, Y., Loizou, P.C.: A comparative intelligibility study of single-microphone noise reduction algorithms. The Journal of the Acoustical Society of America **122**(3), 1777–1786 (2007) 3, 14, 66, 79

[3] Hu, Y., Loizou, P.C.: Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. **49(7)**, 588–601 (July, 2007) 3, 66, 75

[4] Loizou, P.C.: Speech Enhancement: Theory and Practice. FL: CRC, Boca Raton (2007) 3, 13, 15, 17, 66, 74, 76, 79

[5] Loizou, P.C., Kim, G.: Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. Audio, Speech, and Language Processing, IEEE Transactions on **19**(1), 47–56 (2011) 3, 4, 66, 67, 71, 73, 78

[6] Hu, Y., Loizou, P.C.: Subjective comparison of speech enhancement algorithms. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference On, vol. 1, pp. 153–156 (2006) 3, 17, 66, 68, 74

[7] Lim, J.S.: Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise. Acoustics, Speech and Signal Processing, IEEE Transactions on **26**(5), 471–472 (1978) 3, 66

[8] Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. Acoustics, Speech and Signal Processing, IEEE Transactions on **33**(2), 443–445 (1985) 3, 66

[9] Hu, Y., Loizou, P.C.: A generalized subspace approach for enhancing speech corrupted by colored noise. Speech and Audio Processing, IEEE Transactions on **11**(4), 334–341 (2003) 3, 14, 66, 67, 78

[10] P.C. Loizou, A.L., Hu, Y.: Subspace algorithms for noise reduction in cochlear implants. The Journal of the Acoustical Society of America **118**(5), 2791–2793 (2005) 3, 14, 67, 79

[11] G. Kim, Y.H. Y. Lu, Loizou, P.C.: An algorithm that improves speech intelligibility in noise for normal-hearing listeners. The Journal of the Acoustical Society of America **126(3)**, 1486–1494 (2009) 3, 67

[12] E.W. Healy, Y.W. S.E. Yoho, Wang, D.L.: An algorithm to improve speech recognition in noise for hearing-impaired listeners. The Journal of the Acoustical Society of America **134(4)**, 3029–3038 (2013) 3, 67

[13] ITU-T, Coding of Speech at 8 kbit/s Using Conjugate Structure Algebraic Code - Excited Linear Prediction (CS-ACELP). Annex B: A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommend. V.70, International Telecommunication Union, Geneva, 1996 6, 34

[14] ITU-T, Coding of Speech at 8 kbit/s Using Conjugate Structure Algebraic Code - Excited Linear Prediction (CS-ACELP). , International Telecommunication Union, Geneva, 1996 6

[15] Benyassine, S.E. A., Su, H.Y.: ITU 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 Comm. Mag., **35**(9), 64–73 (1997) 6

[16] ETSI, Digital Cellular Telecommunications System (Phase 2+), Voice Activity Detector (VAD) for Adaptive Multi Rate (AMR) Speech Traffic Channels, General Description, 1999 8, 34

[17] Sohn, J., Sung, W.: A voice activity detector employing soft decision based noise spectrum adaptation. In: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference On, vol. 1, pp. 365–3681 (1998). doi:10.1109/ICASSP.1998.674443 11

[18] Sohn, J., Kim, N.S., Sung, W.: A statistical model-based voice activity detection. Signal Processing Letters, IEEE **6**(1), 1–3 (1999). doi:10.1109/97.736233 11, 12, 34

[19] Ghosh, P.K., Tsiartas, A., Narayanan, S.: Robust Voice Activity Detection Using Long-Term Signal Variability. Audio, Speech, and Language Processing, IEEE Transactions on **19**(3), 600–613 (2011) 12, 23, 28, 30, 32, 34

[20] Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. Acoustics, Speech and Signal Processing, IEEE Transactions on **27**(2), 113–120 (1979) 14, 66

[21] Jensen, S.H., Hansen, P.C., Hansen, S.D., Sorensen, J.A.: Reduction of broad-band noise in speech by truncated qsvd. Speech and Audio Processing, IEEE Transactions on **3**(6), 439–448 (1995) 14

[22] Hu, Y., Loizou, P.C.: A subspace approach for enhancing speech corrupted by colored noise. In: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference On, vol. 1, pp. 573–576 (2002) 14

[23] Ephraim, Y., Van Trees, H.L.: A signal subspace approach for speech enhancement. Speech and Audio Processing, IEEE Transactions on **3**(4), 251–266 (1995) 14

[24] Wiener, N.: Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications. Cabridge, MA: MIT Press,, ??? (1949) 15

[25] Hu, Y., Loizou, P.C.: Speech enhancement based on wavelet thresholding the multitaper spectrum. Speech and Audio Processing, IEEE Transactions on **12**(1), 59–67 (2004) 16, 66, 67, 69, 78

[26] Hu, Y., Loizou, P.C.: Incorporating a psychoacoustical model in frequency domain speech enhancement. Signal Processing Letters, IEEE **11**(2), 270–273 (2004). doi:10.1109/LSP.2003.821714 16

[27] Scalart, P., Filho, J.V.: Speech enhancement based on a priori signal to noise estimation. In: Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference On, vol. 2, pp. 629–632 (1996) 16, 66, 67, 72, 73

[28] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V.: TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia (1993) 16, 26

[29] Varga, A., Steeneken, H.J.M.: 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 Communication **12**(3), 247–251 (1993) 16, 30, 32

[30] Hu, Y., Loizou, P.C.: Subjective comparison and evaluation of speech enhancement algorithms. Speech Communication **49**(7), 588–601 (2007) 17, 74

[31] IEEE Recommnded Pratice for Speech Quality Measurements. Audio and Electroacoustics, IEEE Transactions on **17**(3), 225–246 (1969) 17, 74

[32] Hirsch, H., Pearce, D.: The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. ISCA ITRW ASR2000 (Sept. 2000) 17, 75

[33] Itoh, K., Mizushima, M.: Environmental noise reduction based on speech/non-speech identification for hearing aids. In: Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference On, vol. 1, pp. 419–422 (1997). doi:10.1109/ICASSP.1997.599662 22

[34] Freeman, D.K., Cosier, G., Southcott, C.B., Boyd, I.: The voice activity detector for the pan-european digital cellular mobile telephone service. In: Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference On, vol. 1, pp. 369–372 (1989) 22, 33

[35] Faubel, F., Georges, M., Kumatani, K., Bruhn, A., Klakow, D.: Improving hands-free speech recognition in a car through audio-visual voice activity detection. In: Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop On, pp. 70–75 (2011). doi:10.1109/HSCMA.2011.5942412 22

[36] Syed, W.Q., Wu, H.-C.: Speech Waveform Compression Using Robust Adaptive Voice Activity Detection for Nonstationary Noise in Multimedia Communications. In: Global Telecommunications Conference, 2007. GLOBECOM 07. IEEE, pp. 3096–3101 (2007) 22

[37] Kondoz, A.M., Evans, B.G.: A high quality voice coder with integrated echo canceller and voice activity detector for VSAT systems. In: Satellite Communications-ECSC-3, 1993., 3rd European Conference On, pp. 196–200 (1993) 22

[38] Benyassine, A., Shlomot, E., Su, H.-Y., Yuen, E.: A robust low complexity voice activity detection algorithm for speech communication systems. In: Speech Coding For Telecommunications Proceeding, 1997 IEEE Workshop On, pp. 97–98 (1997) 22

[39] Rabiner, L.R., Sambur, M.R.: An algorithm for determining the endpoints of isolated utterances. Bell System Technical Journal **54**(2), 297–315 (1975) 22

[40] Davis, A., Nordholm, S., Togneri, R.: Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. Audio, Speech, and Language Processing, IEEE Transactions on **14**(2), 412–424 (2006) 22, 27, 32, 33

[41] Shuyin, Z., Ying, G., Buhong, W.: Auto-Correlation Property of Speech and its Application in Voice Activity Detection. In: Education Technology and Computer Science, 2009. ETCS '09. First International Workshop On, vol. 3, pp. 265–268 (2009) 22

[42] Marzinzik, M., Kollmeier, B.: Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. Speech and Audio Processing, IEEE Transactions on **10**(2), 109–118 (2002) 22

[43] Nemer, E., Goubran, R., Mahmoud, S.: Robust voice activity detection using higher-order statistics in the LPC residual domain. Speech and Audio Processing, IEEE Transactions on **9**(3), 217–231 (2001) 22

[44] Ramirez, J., Segura, J.C., Benitez, C., Torre, A., Rubio, A.: Efficient voice activity detection algorithms using long-term speech information. Speech Communication **42**(3–4), 271–287 (2004) 22, 23, 34

[45] Lee, B., Hasegawa-Johnson, M.: Minimum Mean Squared Error A posteriori Estimation of High Variance Vehicular Noise. In: in Proc. Biennial on DSP for In-Vehicle and Mobile Systems (2007) 22

[46] Khoa, P.C.: Noise Robust Voice Activity Detection. Master's thesis, NangYang Technological University (2012) 22

[47] Madhu, N.: Note on measures for spectral flatness. Electronics Letters **45**(23), 1195–1196 (2009) 24

[48] Renevey, P., Drygajlo, A.: Entropy based voice activity detection in very noisy conditions. In: Proc. Eurospeech, pp. 1887–1890 (2001) 27

[49] Manolakis, D.G., Ingle, V.K., Kogon, S.M.: Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing. Artech House signal processing library, pp. 212–237. Artech House, ??? (2005) 27, 30

[50] Bies, D.: Engineering Noise Control: Theory and Practice. Taylor and Francis, ??? (2003) 28

[51] Prasad, R.V., Sangwan, A., Jamadagni, H.S., M.C, C., Sah, R., Gaurav, V.: Comparison of Voice Activity Detection Algorithms for VoIP. In: Proceedings of the Seventh International Symposium on Computers and Communications (ISCC'02). ISCC'02, pp. 530–535. IEEE Computer Society, Washington, DC, USA (2002) 31

[52] Beritelli, F., Casale, S., Ruggeri, G.: Performance evaluation and comparison of itu-t/etsi voice activity detectors. In: Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference On, vol. 3, pp. 1425–1428 (2001) 32

[53] Beritelli, F., Casale, S., Ruggeri, G.: A psychoacoustic auditory model to evaluate the performance of a voice activity detector. In: Signal Processing Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference On, vol. 2, pp. 807–810 (2000) 32

[54] Digital Cellular Telecommunications System (Phase 2+), Adaptive Multi Rate (AMR) Speech, ANSI-C Code for AMR Speech Codec, 1998 34

[55] ITU, Coding of Speech at 8 kbit/s Using Conjugate Structure Algebraic Code - Excited Linear Prediction. Annex I: Reference Fixed-Point Implementation for Integrating G.729 CS-ACELP Speech Coding Main Body with Annexes B, D and E, International Telecommunication Union, 2000 34

[56] Schroeder, M.R.: Apparatus for suppressing noise and distortion in communication signals. U.S. Patent 3180936 (Apri., 27 1965) 66

[57] Chen, J., Benesty, J., Huang, Y.: New insights into the noise reduction wiener filter. Audio, Speech, and Language Processing, IEEE Transactions on **14(4)**, 1218–1234 (2006) 66, 72

[58] M. Berouti, R.S., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79., vol. 4, pp. 208–211 (1979) 66

[59] H. Gustafsson, S.E.N., Claesson, I.: Spectral subtraction using reduced delay convolution and adaptive averaging. Speech and Audio Processing, IEEE Transactions on **9**(8), 799–807 (2001) 66

[60] Kamath, S., Loizou, P.C.: A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference On, vol. 4, p. 4164 (2002) 66

[61] Jabloun, F., Champagne, B.: Incorporating the human hearing properties in the signal subspace approach for speech enhancement. Speech and Audio Processing, IEEE Transactions on **11**(6), 700–708 (2003) 66

[62] Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. Acoustics, Speech and Signal Processing, IEEE Transactions on **32**(6), 1109–1121 (1984) 66

[63] Loizou, P.C.: Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum. Speech and Audio Processing, IEEE Transactions on **13**(5), 857–869 (2005) 66

[64] Lim, J.S., Oppenheim, A.V.: Enhancement and bandwidth compression of noisy speech. Proceedings of the IEEE **67**(12), 1586–1604 (1979) 66

[65] Lim, J.S., Oppenheim, A.V.: All-pole modeling of degraded speech. Acoustics, Speech and Signal Processing, IEEE Transactions on **26**(3), 197–210 (1978) 66

[66] Thomson, D.J.: Spectrum estimation and harmonic analysis. Proceedings of the IEEE **70**(9), 1055–1096 (1982) 67, 68

[67] Kay, S.M.: Modern Spectral Estimation. NJ: Prentice-Hall, Englewood Cliffs (1988) 68

[68] Riedel, K.S., Sidorenko, A.: Minimum bias multiple taper spectral estimation. Signal Processing, IEEE Transactions on **43**(1), 188–195 (1995) 68

[69] Moulin, P.: Wavelet thresholding techniques for power spectrum estimation. Signal Processing, IEEE Transactions on **42**(11), 3126–3136 (1994) 68

[70] A.T. Walden, D.B.P., McCoy, E.J.: Spectrum estimation by wavelet thresholding of multitaper estimators. Signal Processing, IEEE Transactions on **46**(12), 3153–3165 (1998) 68, 69

[71] Cristan, A.C., Walden, A.T.: Multitaper power spectrum estimation and thresholding: wavelet packets versus wavelets. Signal Processing, IEEE Transactions on **50**(12), 2976–2986 (2002) 68

[72] Wahba, G.: Automatic smoothing of the log periodogram. Journal of the American Statistical Association **75**(369), 122–132 (1980) 68

[73] Percival, D.B.: Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques. MA: Cambridge Univ. Press, Cambridge (1993) 68

[74] Bartlett, M.S., Kendall, D.G.: The statistical analysis of variance-heterogeneity and the logarithmic transformation. Supplement to the Journal of the Royal Statistical Society **8**(1), 128–138 (1946) 69

[75] S. Quackenbush, T.B., Clements: Objective Measures of Speech Quality. NJ: Prentice-Hall, Englewood Cliffs (1988) 69, 75

[76] Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994) 69

[77] Donoho, D.L.: De-noising by soft-thresholding. Information Theory, IEEE Transactions on **41**(3), 613–627 (1995) 69

[78] Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association **90**(432), 1200–1224 (1995) 69

[79] Johnstone, I.M., Silverman, B.W.: Wavelet threshold estimators for data with correlated noise. Journal of the Royal Statistical Society. Series B (Methodological) **59**(2), 319–351 (1997) 69

[80] Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **11**(7), 674–693 (1989) 70, 73

[81] J. Ma, Y.H., Loizou, P.C.: Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. The Journal of the Acoustical Society of America **125(5)**, 3387–3405 (2009) 71, 75

[82] Hu, Y., Loizou, P.C.: Evaluation of objective quality measures for speech enhancement. Audio, Speech, and Language Processing, IEEE Transactions on **16**(1), 229–238 (2008) 71, 75

[83] Rangachari, S., Loizou, P.C.: A noise-estimation algorithm for highly non-stationary environments. Speech Communication **48**(2), 220–231 (2006) 73

[84] Hu, Y., Loizou, P.C.: Evaluation of objective measures for speech enhancement. In: Proc. of INTERSPEECH (2006) 74, 75, 77, 79, 86, 93

[85] Ma, J., Loizou, P.C.: SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech. Speech Commun. **53(3)**, 340–354 (Mar., 2011) 74, 75, 77, 78, 86, 93

[86] P.862, I.-T.: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union (2000) 75

[87] Klatt, D.: Prediction of perceived phonetic distance from critical-band spectra: A first step. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82., vol. 7, pp. 1278–1281 (1982) 76

[88] ITU-T P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms. ITU-T Recommendation P.835 (2003) 76

[89] ANSI, 1997: Methods for calculation of the speech intelligibility index. Technical Report S3.5-1997, American National Standards Institute (1997) 78