Speech Enhancement with Adaptive Thresholding and Kalman Filtering

Mengjiao Zhao

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Electrical Engineering) at

Concordia University

Montréal, Québec, Canada

September 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mengjiao Zhao**

Entitled: **Speech Enhancement with Adaptive Thresholding and Kalman Filtering**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. R. Raut

_____ External Examiner
Dr. Wen Fang Xie

_____ Examiner
Dr. Hassan Rivaz

_____ Supervisor
Dr. Wei-Ping Zhu

Approved by _____
Dr. W. E. Lynch, Chair
Department of Electrical and Computer Engineering

_____ 2017

_____
Dr. Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Speech Enhancement with Adaptive Thresholding and Kalman Filtering

Mengjiao Zhao

Speech enhancement has been extensively studied for many years and various speech enhancement methods have been developed during the past decades. One of the objectives of speech enhancement is to provide high-quality speech communication in the presence of background noise and concurrent interference signals. In the process of speech communication, the clean speech signal is inevitably corrupted by acoustic noise from the surrounding environment, transmission media, communication equipment, electrical noise, other speakers, and other sources of interference. These disturbances can significantly degrade the quality and intelligibility of the received speech signal. Therefore, it is of great interest to develop efficient speech enhancement techniques to recover the original speech from the noisy observation. In recent years, various techniques have been developed to tackle this problem, which can be classified into single channel and multi-channel enhancement approaches. Since single channel enhancement is easy to implement, it has been a significant field of research and various approaches have been developed. For example, spectral subtraction and Wiener filtering, are among the earliest single channel methods, which are based on estimation of the power spectrum of stationary noise. However, when the noise is non-stationary, or there exists music noise and ambient speech noise, the enhancement performance would degrade considerably. To overcome this disadvantage, this thesis focuses on single channel speech enhancement under adverse noise environment, especially the non-stationary noise environment.

Recently, wavelet transform based methods have been widely used to reduce the undesired background noise. On the other hand, the Kalman filter (KF) methods offer competitive denoising results, especially in non-stationary environment. It has been used as a popular and powerful tool for

speech enhancement during the past decades. In this regard, a single channel wavelet thresholding based Kalman filter (KF) algorithm is proposed for speech enhancement in this thesis. The wavelet packet (WP) transform is first applied to the noise corrupted speech on a frame-by-frame basis, which decomposes each frame into a number of subbands. A voice activity detector (VAD) is then designed to detect the voiced/unvoiced frames of the subband speech. Based on the VAD result, an adaptive thresholding scheme is applied to each subband speech followed by the WP based reconstruction to obtain the pre-enhanced speech. To achieve a further level of enhancement, an iterative Kalman filter (IKF) is used to process the pre-enhanced speech.

The proposed adaptive thresholding iterative Kalman filtering (AT-IKF) method is evaluated and compared with some existing methods under various noise conditions in terms of segmental SNR and perceptual evaluation of speech quality (PESQ) as two well-known performance indexes. Firstly, we compare the proposed adaptive thresholding (AT) scheme with three other thresholding schemes: the non-linear universal thresholding (U-T), the non-linear wavelet packet transform thresholding (WPT-T) and the non-linear SURE thresholding (SURE-T). The experimental results show that the proposed AT scheme can significantly improve the segmental SNR and PESQ for all input SNRs compared with the other existing thresholding schemes. Secondly, extensive computer simulations are conducted to evaluate the proposed AT-IKF as opposed to the AT and the IKF as standalone speech enhancement methods. It is shown that the AT-IKF method still performs the best. Lastly, the proposed ATIKF method is compared with three representative and popular methods: the improved spectral subtraction based speech enhancement algorithm (ISS), the improved Wiener filter based method (IWF) and the representative subband Kalman filter based algorithm (SIKF). Experimental results demonstrate the effectiveness of the proposed method as compared to some previous works both in terms of segmental SNR and PESQ.

Acknowledgments

I would like to express my sincerest gratitude to my supervisor Dr. Weiping Zhu for his earnest help and patience on my thesis research. From the choice of the subject, the implementation of the project until the final completion of the paper, Dr. Zhu has always given me patient guidance and support. Every point of the results I achieved has embodied the professor's effort. I am deeply influenced and inspired by Dr. Zhu's open field of vision, rigorous scholarship attitude and the work style of excellence. I would like to express my greatest thanks and respect to my Professor Dr. Zhu.

I am indebted to Dr. Benoit Champagne of Mcgill University who has kindly helped me for algorithmic development and troubleshooting through the regular CRD project meetings. My special thanks go to Microsemi technical staff for their valuable feedbacks and interactions that broadened my horizon and helped me to understand the industrial focuses and practical techniques.

I would also like to express my gratitude to all the brothers and sisters of the laboratory for their support and assistance in carrying out the relevant work in patience. Thanks to my colleagues, Mr. Sujan Kumar Roy, Mr. Xinrui Pu, Mr. Amir Amini, Mr. Mahdi Parchami for providing enthusiastic help in the experimental process.

Thanks to my family for years of my support and understanding! They are my favourite people. It is their silent dedication that motivates and supports me for my study career.

Contents

List of Figures

List of Abbreviations

AR	Auto-Regressive
AT	Adaptive Thresholding
AT-IKF	Adaptive Thresholding Iterative Kalman Filtering
BWT	Bionic Wavelet Transform
CWT	Continuous Wavelet Transform
DWT	Discrete Wavelet Transform
IKF	Iterative Kalman Filter
ISS	Improved Spectral Subtraction
IWF	Improved Wiener Filter
KF	Kalman Filter
LPC	Linear Prediction Coefficient
MAD	Median Absolute Deviation
PESQ	Perceptual Evaluation of Speech Quality
SIKF	Subband Iterative Kalman Filter
SNR	Signal Noise Ratio
SURE-T	SURE Thresholding

U-T	Universal Thresholding
VAD	Voice Activity Detection
WP	Wavelet Packet
WPT	Wavelet Packet Transform
WPT-T	Wavelet Packet Transform Thresholding
WT	Wavelet Transform

Chapter 1

Introduction

1.1 Overview of Speech Enhancement

Speech signal processing is an important and popular discipline that is concerned with the processing of speech signals using digital signal processing techniques. It is to study the speech sound process, the statistical characteristics of speech signals, speech recognition, speech machine synthesis and other processing technologies. One important purpose of speech signal processing is to estimate speech parameters which reflect the characteristics of speech signals in order to efficiently transmit or store speech information. Thanks to the explosive development of speech processing techniques, different application fields have been addressed: speech enhancement, speech coding, speech recognition and speech synthesis.

One of the most important branches of speech processing is speech enhancement. It has been extensively studied for many years. One of the objectives of speech enhancement is to provide high-quality speech communication in the presence of background noise and concurrent interference signals [1]. In the process of speech communication, the clean speech signal is inevitably corrupted by acoustic noise from the surrounding environment, transmission media, communication equipments, electrical noise, other speakers, and other sources of interference. These disturbances significantly degrade the quality and intelligibility of the received speech signal. Therefore, it is of great interest to develop an efficient speech enhancement technique to recover the original speech from the noisy observation.

With the growing need for the development of the speech and audio systems, speech enhancement has been widely used as a front end tool for mobile phones, VoIP (voice over internet protocol), teleconferencing systems, and speech recognition[2]. Speech enhancement is also vital to hearing aid devices, which help the hearing impaired by amplifying ambient audio signals[3], [4].

As to the speech enhancement performance evaluation, it is worth-mentioning two perceptual criteria: quality and intelligibility. Since the quality reflects individual preferences of listeners, it is a subjective performance evaluation metric. The intelligibility is an objective measure since it offers the percentage of words which could be correctly identified by listeners. Based on these two criteria, the main challenge in designing effective speech enhancement algorithms is to suppress noise without introducing any perceptible distortion in the signal [5]. Speech enhancement not only involves the traditional signal processing theory, such as signal detection and wavelet estimation, but also closely relates to the characteristics of speech and human ear perception. In addition, in practical environment, noises are generated from different sources and presented in different forms, which leads to the development of a wide range of speech enhancement technologies. Therefore, it is necessary to consider the speech characteristics, human ear perception characteristics and noise characteristics in order to choose appropriate speech enhancement methods according to the specific application situations.

1.1.1 Speech characteristics

The speech signal is a non-stationary and time-varying random process. Its production process is closely related to the movement of the vocal organs[6]. Considering that the adaptive changing rate in the process of human vocal organs has a certain limit and is far more smaller than the change rate of the speech signal, the speech signal in physical and spectral characteristics can be considered stable in a short period of time (e.g. $10 - 30ms$). Thus, the analysis methods of the stationary random process and the stationary characteristics in the short-term spectrum can be applied for speech enhancement.

Speech signal is divided into unvoiced and voiced categories. Both have obvious differences in the speech mechanism and characteristics. For example, the vibration of vocal folds produces periodic or quasi-periodic excitations to the vocal tract to generate voiced speech whereas pure

transient and/or turbulent noises are aperiodic excitations to the vocal tract to generate unvoiced speech [7]. The voiced speech in the time domain has periodic and strong amplitude. Most of its energy is concentrated in the low frequency range and its frequency presents the pitch frequency. In contrast, the unvoiced speech has no obvious time domain and frequency domain features. The waveform of unvoiced speech is similar to white noise and has a weaker amplitude. Knowing about the speech characteristics is of crucial importance for speech analysis and processing, such as in voice activity detection, where a preliminary acoustic segmentation of speech is conducted based on the characteristics of speech versus noise.

1.1.2 Human ear perception

One important measurement of speech enhancement performance is the subjective feeling of the human ear, so it is worth analyzing the human ear perception feature to be used in the speech enhancement. There are some useful facts that can be applied when developing speech enhancement methods[8]:

- The perception of the human ear is mainly through the amplitude of the spectral component of the speech signal, which is not sensitive to the phase of each speech component.
- The sensibility of the human ear to the spectral component strength is a binary function of frequency and energy spectrum. The loudness is proportional to the logarithm of the spectral amplitude.
- The human ear has a masking effect, that is, the strong signal has a damaging effect on the weak signal. The degree of masking is a binary function of the speech amplitude and frequency.

1.1.3 Noise characteristics

Generally speaking, the noise can be additive or non-additive. For non-additive noise, some can be transformed into additive ones. For example, multiplicative noise can be converted to additive noise by homomorphic transformation. Additive noise is usually divided into periodic noise, impulse noise, broadband noise and simultaneous voice interference, etc [9]. Periodic noise is mainly from the periodic operation of the machine or engine, and electrical disturbances can also cause periodic noise. The feature of periodic noise is that, in the frequency spectrum, there are many

time-varying, discrete narrow-spectrum peaks, which overlap the speech signal. Based on its characteristic, it is necessary to adopt the adaptive filtering method to automatically identify and distinguish the periodic noise components. Impulsive noise is usually derived from explosion, discharge and sudden interference. Its character is that the waveform has narrow pulses which are better to be eliminated in the time domain. Broadband noise comes from a variety of sources, including wind, respiratory noise and general random noise sources. Quantization noise is usually treated as white noise, but it can also be regarded as broadband noise. Since the broadband noisy speech signal is completely overlapped in the time domain and the frequency domain, it is very difficult to eliminate the broadband noise. Simultaneous speech interference is the speech interference caused by multiple words that are overlapped together in transmission or recording. Its character is that different tones have pitch differences and we can consider the speech separation technologies to process the noisy speech.

1.2 Literature Review

The goal of speech enhancement is to improve the quality of noisy speech, which is normally accomplished by reducing the ambient noise while minimizing the speech distortion. Speech enhancement can be divided into single channel and multi-channel approaches according to the number of microphones used in speech acquisition. In this thesis, we focus on single-channel speech enhancement approaches, where the speech signal is generated from a single microphone. Our goal is to give an overview of the general ideas and principles behind the most successful single-channel speech enhancement systems. Consider an additive noise model as given by

$$y(k) = s(k) + v(k), \tag{1}$$

where $y(k)$ represents the observed noisy signal, $s(k)$ is the k^{th} sample of the clean speech and $v(k)$ is the noise samples. We can also see that a speech enhancement method is often transformed into an estimation problem of speech signal, namely, given an observation $y(k)$ of the noisy process, find an estimate of the target realization $s(k)$[10]. Although there are many speech enhancement systems available in time, frequency and wavelet domains, we will discuss, in the following subsections, the

existing popular approaches, which are closely related to the research in the thesis.

1.2.1 Spectral subtraction based techniques for speech enhancement

Spectral subtraction is a popular and effective speech enhancement method for noise reduction. It was first proposed in [11] and considered as one of the earliest speech enhancement methods. The basic principle is that we obtain an estimate of the clean signal spectrum by subtracting an estimate of the noise spectrum from the noisy speech spectrum on a frame by frame basis. The reason why the method is popular is that this algorithm is computationally simple as only a single forward and an inverse Fourier transform are involved. The basic equation is described as

$$\hat{S}(w_n) = Y(w_n) - V(w_n), \tag{2}$$

where $w_n = 2\pi n/L$, $n = 0, 1, ..., L-1$ and L is the frame length. $\hat{S}(w_n)$ represents the estimated clean speech frame spectrum, $Y(w_n)$ is the noisy speech frame spectrum and $V(w_n)$ denotes the estimated noise frame spectrum. Then the power spectrum of the estimated speech is presented as

$$|\hat{S}(w_n)|^2 = |Y(w_n)|^2 - |\hat{V}(w_n)|^2, \tag{3}$$

where $|\hat{V}(w_n)|^2$ denotes the estimate of noise power from non-speech frames. Combined with the phrase of the noisy speech signal, the frequency domain enhanced speech $|\hat{S}(w_n)|$ can be expressed as

$$\hat{S}(w_n) = [|Y(w_n)|^2 - |\hat{V}(w_n)|^2]^{\frac{1}{2}} e^{j\arg[Y(w_n)]}, \tag{4}$$

The time-domain enhanced speech $\hat{s}(k)$ is obtained by applying the inverse discrete Fourier transform to the estimated signal spectrum $\hat{S}(w_n)$.

Even though the spectral subtraction algorithm performs noise reduction effectively, it is highly dependent on the accuracy of noise spectrum estimation. If the noisy speech spectrum is over subtracted, which directly causes the speech distortion. Moreover, when we transform the estimated speech from frequency domain to time domain, small and isolated peaks happen in the spectrums,

which cause frequency change fast between frames. This phenomenon results in another type of noise, usually called "musical noise" [11].

To address the issues mentioned above, in [12], a standard spectral subtractor combined with constrained high order notch filter was applied to reduce music noise efficiently, but the speech distortion problem can not be solved. In [13], a psychoacoustical model was applied to adjust the over-subtraction values to render the residual noise inaudible. However, it is more concentrated on the speech recognition part instead of speech quality improvement. Kamath and Loizou proposed in [14] a multi-band spectral subtraction method which divided the speech spectrum into some contiguous frequency bands. This method reduced the speech distortions while obtaining a speech with high quality. In [15], an improved spectral subtraction algorithm was proposed for speech enhancement under non-stationary noise environment. In this method, smoothed spectrums were applied to estimate the speech and noisy spectrums with auto-regressive model, and construct speech and noise codebooks. The speech and noise entries for the codebooks were obtained from Log-spectral minimization, to perform the spectral subtraction algorithm.

From the above discussion, it is observed that the improved spectral subtraction based techniques for speech enhancement can easily and conveniently be implemented, but they have a few limitations.

The noise spectrum estimator always has errors which inevitably influence the performance of spectral subtraction based techniques. Moreover, musical noise is always introduced and can not be removed completely. Last, the implement of noise estimation, such as that used in the VAD scheme, restricts the update of the estimation within the period of speech absence [15].

1.2.2 Wiener filtering based techniques for speech enhancement

As one of the most fundamental approaches, Wiener filter was first introduced by Lim and Oppenheim in [16]. The basic principle is to obtain an estimate of the clean signal by minimizing the Mean Square Error between the estimated signal $\hat{s}(k)$ and the clean signal $s(k)$. The frequency

domain solution to this optimization problem gives the following filter transfer function [16]

$$H(w) = \frac{|Y(w)|^2 - |V(w)|^2}{|Y(w)|^2}, \tag{5}$$

Then, the transfer function $H(w)$ is multiplied with the noisy speech spectrum $Y(w)$ to attenuate the noise frequency components appropriatly, yielding the enhanced speech spectrum as given by

$$\hat{S}(w) = H(w)Y(w). \tag{6}$$

Traditional Wiener filtering algorithm is the linear estimator for pure signal, optimal in the mean square sense [17]. Speech signal needs to be stationary to realise accurate estimation. However, realistic speech signals do not meet the stationarity requirement in practical environment. Thus, numerous wiener filtering based techniques have been developed as stated in the following paragraph.

In [18], a perceptual modified Wiener filter with a fast noise estimation was applied for speech enhancement. In this method, a smoothing parameter relying on the estimated subband SNR was updated adaptively for fast noise estimation.Then the noise estimate was applied to the perceptual modified Wiener filter that was first proposed in [19]. In [20], complex (linear prediction coefficient) LPC speech analysis was proposed for a Wiener filter based speech enhancement algorithm. The proposed method performed better especially in lower input SNRs since the complex LPC speech analysis can estimate the spectrum more accurately. The disadvantage of this method is that some background noise was introduced in the enhanced speech. Another method using an improved iterative Wiener filtering algorithm was proposed in [21]. By applying the VAD technology with SNR tracking algorithm, the accuracy of noise power spectrum estimation can be improved. However, the number of iterations is not precisely specified, which may increase the computational cost and cause the algorithm divergence.

From the above improved Wiener filter based techniques for speech enhancement, it is observed that the noise power spectrum estimation plays a significant role on the performance of Wiener filter based techniques.

1.2.3 Wavelet thresholding based techniques for speech enhancement

Wavelet theory provides a unified framework under which a number of techniques have been developed independently during the past decades for various signal processing applications. These applications include speech and image denoising, compression, detection and pattern recognition. Among various approaches, wavelet thresholding based methods have been widely used to reduce the undesirable background noise.

In [22], the discrete wavelet transform has been applied to obtain the wavelet coefficients of the noisy speech signal, then a semisoft threshold function was used to remove noise components. However, the main problem of this method is that the detection criterion of the voiced/unvoiced segments of the speech signals is not precise enough. As the wavelet band energy is the only detection criterion used therein, the enhancement performance drastically decreases if the segments are classified incorrectly. The authors of [23] proposed a method based on the time adaption of wavelet thresholds, in which the Teager energy operator was introduced for the time dependence, which is capable of extracting the signal energy from the perspective of mechanical and physical considerations [24] – [25]. However, this method has an over - thresholding problem in speech denoisng and enhancement applications. The authors of [26] presented a bark-scaled wavelet packet decomposition combined with a soft-decision gain modification and a ”magnitude” decision-directed estimation technique. Due to the perceptual specialization, this method achieved higher intelligibility and quality of enhanced speech as compared to the speech enhancement systems based on uniform spectral decompositions [13], [27].

In [28], a new noise estimation method was proposed based on spectral entropy using histogram of intensity. The evaluation shows that it performs well for the speech especially with strong noise. However, choice of a good threshold value and thresholding scheme for reduction of noise in the wavelet domain is a challenging subject. Jong Kwan Lee and Chang D. Yoo suggested in [29]that different thresholds be applied depending whether the speech frame is voiced or unvoiced according to voice activity detection (VAD). The VAD criterion used to determine voiced and noise frames is to compare the frame energy with certain thresholds. Namely, a voiced frame is detected if the measurement exceeds a certain threshold. Otherwise, the frame is decided as noise. In practise,

however, since the authors used the energy estimate as the only criterion, not all voice or noise frames could be identified correctly by such an energy-based decision.

The authors in [30] presented an improved wavelet-based speech enhancement method using the perceptual wavelet packet decomposition combined with the Teager energy operator. This method efficiently avoided the over thresholding of speech signal especially when the speech was corrupted by slight noises. The Bionic wavelet transform was introduced in [31] and the simulation results indicated that the enhancement quality is inferior to that of Ephraim Malah filtering and iterative Wiener filtering, but superior to the perceptually scaled wavelet method.

From the above observations on the various wavelet based thresholding methods for speech enhancement, we can conclude that in practice the wavelet based thresholding method is influenced by three basic but important components, namely, the choice of wavelet transform, choice of thresholding schemes, and the selection of thresholding values. All these components need to be carefully designed in order to achieve good speech enhancement results.

1.2.4 Kalman filter based techniques for speech enhancement

The Kalman filter (KF) is an optimal estimator that estimates the state of a system by minimizing the mean squared error, which operates through a prediction and correction mechanism. The KF is a common sensor fusion algorithm [32] and provides a better estimate than any single measurement does, since it uses the system's dynamics model and measurements to form an estimate of the state of system. The KF was first proposed by R. Kalman in [33] to predict the unknown states of a dynamic system. Since then, many variations of the KF have been proposed and they become well known in signal processing for their efficient implentation. With respect to the KF for speech enhancement, it mainly involves two steps: (1) the estimation of linear prediction coefficients (LPCs), the noise variance and the parameters of the speech model; and (2) speech signal retrieval using Kalman filtering approach. Overall, the Kalman filter is popular for offering good results in practice and it is convenient for online real time processing as well.

In 1987, Paliwal and Basu first applied the Kalman filter to speech enhancement [34], in which speech was modelled as autoregressive process and represented in the state-space domain. They developed a mathematical formulation for Kalman filtering and compared the speech enhancement

results with the Wiener filtering method. However, this method performed well only in white Gaussian noise rather than in other noise environments. Gibson in [35] proposed Kalman filter algorithms with scalar and vector outputs, under white and colored noise assumption for speech enhancement.This method provided SNR increases and improved speech quality and intelligibility in colored noise environment. A disadvantage of the Gibson's Kalman filtering algorithm is that the estimation of model parameters was not discussed.

In [36], an iterative-batch and sequential algorithm with some extensions and modifications was proposed under the support of extensive experimental research using different kinds of actual noise sources. The experimental results show that the output SNR was improved, at the same time, the intelligibility was preserved. However, the linear prediction coefficients (LPCs) and the noise variance are estimated directly from the noisy speech, which decreases the accuracy of the iterative KF approach to a certain degree. The Kalman filter based on priori estimation about the model parameters, called expectation-maiximization, was introduced in [37], where a speech model was proposed which can satisfactorily describe voiced, and unvoiced speeches and silence. A mathematically equivalent algorithm was also proposed to lower the computational cost of KF processing.

In [38], a Kalman filter based method which avoids the explicit estimation of noise variances was proposed, with a lower computational requirement but reduced SNR gain. Thus, a balance between the computational cost and SNR gain needs to be achieved based on different noise environments. A perceptual Kalman filter for speech enhancement in clolored noise was introduced in [39]. In this context, a total masking threshold considering frequency domain simultaneous masking effects and time domian forward masking effects was applied as a post-filter to a Kalman filtered signal for further enhancement in a perceptual sense. This method avoids the frequency domain complexity and makes it suitable to estimate the state-space vector in time domain. To increase the accuracy of noise variance estimation, Saha et al. in [40] applied the sensitivity and the robustness metrics, as well as the interpretations of these metrics, to give a compromised value for noise variance estimation.

In [41], another IKF-based approach is proposed along with a subband processing, where the noisy speech is decomposed into a number of subbands followed by Kalman Filtering (KF) on each subband. This method, however, demands a large computational resource for the implementation

of KF at all subbands. More recently, a subband IKF method with partial reconstruction is proposed in [42], where the noisy speech is first decomposed into a set of subbands, and then a partial reconstruction scheme is used to reconstruct the subbands into high-frequency and low-frequency subband speeches. In this context, the IKF is employed only in the high-frequency subband. Since the low-frequency subband is not filtered by the IKF, this method offers limited enhancement performance for noisy speeches which contain non-negligible noises in low-frequency region.

Based on the above literature review, it is observed that the estimation accuracy of speech model parameters is very important for the performance of the KF processing. Moreover, decreasing the computation cost for the KF processing is also a key topic for speech enhancement.

1.3 Motivation of the Thesis

Based on the review of the aforementioned different kinds of speech enhancement algorithms, it is known that improving the quality and intelligibility of noisy speech is a really important and interesting topic, which strongly motivates the research work of this thesis for further speech improvement from different aspects as articulated below.

- In view of the speech enhancement system, wavelet thresholding based methods have been widely used to reduce the undesirable background noise. However, there are some inevitable defects in the above basic wavelet thresholding methods. Firstly, in real environment most kinds of noise are non-stationary noise, which cause bad speech quality if we apply constant thresholds on it. Secondly, some unvoiced speech segments contain noise-like speech components, which decrease the speech quality of the enhanced speech. In order to address these issues, the development of thresholding rules and the selection of parameters have to be explored further, which brings motivation for more research on the wavelet thresholding based speech enhancement approaches.

- Concerning the wavelet thresholding based methods, many existing methods require the estimation of statistical noise information. Hence, the accuracy of the estimated statistical noise information also decides the performance of the speech enhancement system, which is considered as a major motivation of the research.

- In real and practical environment, most of the speech is non-stationary. However, the spectral subtraction and Wiener based methods always assume that each analysis frame of speech is stationary, which is too ideal. Since the wavelet based Kalman filter methods offer good results, especially in non-stationary environment, due to its optimality and good estimation structure, it has been used as a popular and powerful tool for speech enhancement in the past decades. In this regard, an extensive study based on the Kalman Filter is required and necessary for speech enhancement.

- As well known, the estimation accuracy of LPCs and other state parameters plays a significant role on the performance of Kalman filter based speech enhancement. As stated in Sec. 1.2.4, many kinds of algorithms were provided to increase the accuracy of parameter estimation throughly. This criterion reminds us to combine the wavelet thresholding methods and the Kalman filter together to develop novel speech enhancement methods and therefore serves as another major motivation of this research.

1.4 Objective of the Thesis

The main objective of this thesis is to develop an adaptive wavelet packet thresholding (WPT) method, a pre-enhancement system, with iterative KF being capable of reducing adverse environment noises. In this context, two basic stages are proposed with the details summarized as follows:

- With regards to an improved thresholding scheme, the objective is to obtain a pre-enhanced speech signal to be processed by the IKF. The whole process is performed on a frame-by-frame basis. The noisy speech is first decomposed into a number of subbands with the wavelet packet (WP) transform, which offers a wide range of possibilities and maintains the nature of speech samples in wavelet analysis. The VAD is then applied to each subband frame to determine whether the frame is voice or noise. In contrast to most existing works where only a single parameter is employed for voice/noise frame detection, our method makes use of two measurements in the VAD process: i) frame energy and ii) spectral flatness. A VAD based adaptive thresholding scheme is then proposed for speech enhancement in accordance with each subband frame activity. All in all, the main objective of the adaptive thresholding method is to reduce the noise of non-speech frames.

- In order to reduce the noise of voice frames and enhance parameter estimation accuracy in

different noise environments, an iterative KF based speech enhancement method is applied on a frame-by-frame basis. The whole process contains two loops of iterations, called inner and outer loops for each frame. The inner loop, iteration includes a prediction step and a measurement update step. In the prediction step, the IKF predicts the state vector and parameter covariance by using the previous samples of the sate-space model. In the measurement update step, the Kalman Gain and state vectors are updated. The outer loop contains several iterations, where the LPCs and other state-space model parameters are re-estimated from the same processed speech frame. The iterative procedure stops when the KF converges or the pre-set maximum number of iterations is exhausted, giving the further enhanced result of the same speech frame as compared to the pre-enhanced speech frame.

1.5 Organization of the Thesis

The rest of the thesis is organized as follows:

Chapter 2: This chapter first describes some popular kinds of wavelet transforms for speech enhancement methods. Then some existing thresholding methods and the thresholding values selection schemes are explained. Last, we compare the performances of four wavelet thresholding algorithms in different types of noise environments: (1) the universal threshold (U-T); (2) the WPT threshold (WPT-T); (3) the SURE threshold (SURE-T); and (4) the proposed adaptive threshold (AT).

Chapter 3: This chapter presents the adaptive WP thresholding IKF method for speech enhancement in details. Our proposed approach consists of two successive stages. In the first stage, the WP transform is first applied to the noise corrupted speech frames, which decomposes each frame into a number of subbands. For each subband, a voice activity detector (VAD) is designed to detect the voiced/unvoiced frames of the subband speech. Based on different voice activity, an adaptive thresholding scheme is then utilized to each subband speech followed by the WP based reconstruction to obtain the pre-enhanced speech. In the second stage, the reconstructed and pre-enhanced full-band speech is processed by the IKF for further speech enhancement. Lastly, comparative study of the proposed method and other existing methods is also undertaken.

Chapter 4: This chapter provides extensive computer simulation results and discussions of the proposed method as compared with some other existing competitive methods under different noise environments. The simulation setup, speech database and evaluation parameters are also described. To objectively evaluate the performance of these methods, we adopt the segmental signal-to-noise ratio (SNR_{seg}) and perceptual evaluation of speech quality (PESQ) as the criteria for the performance evaluation.

Chapter 5: This chapter mainly highlights the contributions of this research and suggests some research directions for the future study in the area of speech enahncement.

Chapter 2

Wavelet Analysis and Thresholding Techniques

2.1 Introduction

Wavelet transform (WT) has attracted much attention in the field of signal processing over the past decades. It has been widely used in signal processing applications since the WT offers new and powerful approaches for signal analysis, enhancement and compression etc. In the context of speech enhancement, the WT is first used for speech analysis which is then followed by the thresholding of the wavelet coefficients [43]. The key to wavelet denoising is to minimize the wavelet coefficients corresponding to noise and at the same time to keep the wavelet coefficients corresponding to the clean speech signal. This requires the selection of appropriate thresholds for different frequency components or at different wavelet scales. In general, if the threshold is too small, a portion of the noise wavelet coefficients will not be set to zero, and the denoised signal will retain some of the noise. If the threshold selected is too large, then part of the useful signal will be filtered out. Therefore, the performance of wavelet denoising method mainly depends on the design of the thresholding function and the regulation rules.

In this chapter, we first describe some popular wavelet transforms in Section. 2.2. Then, in section. 2.3 some existing thresholding methods and the threshold value choosing methods are explained. In Section. 2.4, comparisons and evaluations are provided for the choosing of the best

wavelet thresholding method.

2.2 Wavelet Analysis for Speech Enhancement

Wavelet transform, as a powerful time-frequency analysis tool, has been widely applied for speech denoising, due to its higher resolution in the frequency domain as compared with the Fourier transform and Short-time Fourier transform. The main theory of the wavelet transform is an infinite set of various transforms, depending on the merit function used for its computation. This section briefly reviews some typical wavelet transforms that have been applied to signal denoising, including continuous wavelet transform (CWT) [22], a modified CWT or bionic wavelet transform (BWT) [44], the discrete wavelet transform (DWT) [45], and the wavelet packet transform (WPT) [46].

2.2.1 Continuous wavelet transform

Given a time-varying signal $y(t)$, CWT can construct a time-frequency representation, the inner products of the signal and a family of "wavelets", that offers good time and frequency localization. For the wavelet family, its wavelet "prototype" $\psi_{a,b}(t)$ is defined as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}), \tag{7}$$

where $\psi_{a,b}(t)$ is also commonly called the mother wavelet, a band-pass function. The factor $\frac{1}{\sqrt{a}}$ is used to ensure energy preservation [47], [48], [49]. Different ways of discretizing time-scale parameters (b, a) yield different types of wavelet transform.

The CWT was originally developed in [47] which is denoted as

$$CWT(a, b, y(t)) = \int_{-\infty}^{\infty} y(t)\psi^*(\frac{t-b}{a})dt, \tag{8}$$

where the asterisk denotes the complex conjugate of ψ.

2.2.2 Bionic wavelet transform

The BWT is developed based on the the dynamic control mechanism of cochlear in the speech signal processing. The cochlea is located inside the inner ear and is capable of resolving the spectral information of speech. When the sound enters into the cochlea, it causes the vibration of the cochlea basement membrane. Different frequencies of the sound lead to the maximum displacements in different locations of the basement membrane. Thus, the cochlea played a role in the separation of the frequency spectrums. In order to simulate the physiological function of the cochlea, adaptive parameters $T(a,\tau)$ were introduced in the CWT, which is regarded as the scaling of the cochlear filter bank quality factor Q_{eq}.The adaptation factor $T(a,\tau)$ [31] for each scale is updated as:

$$T(a,\tau) = \frac{1}{(1 - G_1 \frac{C_s}{C_s + |X_{bwt}(a,\tau)|})(1 + G_2 |\frac{\partial}{\partial \tau} X_{bwt}(a,\tau)|)}, \tag{9}$$

where G_1 and G_2 are active factors, a is the number of scales, C_s represents non-linear saturation effects in the cochlear model, and X_{bwt} is the BWT coefficients [44].

Based on [50], X_{bwt} can be expressed by the CWT coefficients multiplied by a constant $K(a,\tau)$, which is a function of the adaption factor $T(a,\tau)$. $K(a,\tau)$ and X_{bwt} are given respectively, by

$$K(a,\tau) = \frac{\sqrt{\pi}}{2} \frac{T_0}{\sqrt{1 + T(a,\tau)^2}}, \tag{10}$$

$$X_{bwt}(a,\tau) = K(a,\tau) X_{cwt}(a,\tau). \tag{11}$$

2.2.3 Discrete wavelet transform

Since the time domain signal often contains redundant information, the discrete wavelet transform (DWT) can be used to reduce the redundancy (compression), which operates at different scales according to different applications. DWT is obtained by discretizing the scale and displacement of continuous wavelet transform in accordance with the power of 2. Figure. 2.1 shows a 3-level DWT tree for a 1-D input signal.

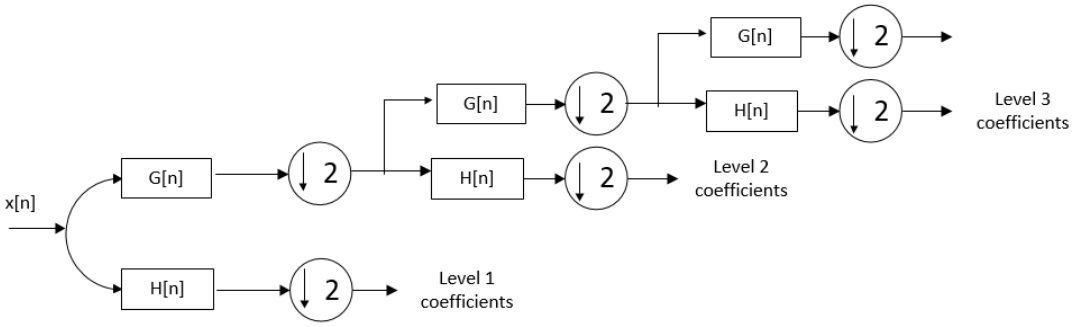


Figure 2.1: A 3-level discrete wavelet decomposition tree

In the above figure, $x[n]$ denotes the discrete input signal, $G[n]$ is the low pass filter, $H[n]$ is the high pass filter and $\downarrow 2$ is the downsampling operator. For each level, only low pass filter ($G[n]$) is decomposed and high pass filter ($H[n]$) is retained.

For many signals, low-frequency components are very important, which often contain the characteristics of the signal. High-frequency components show the signal details or differences. If we ignore the high-frequency components, the speech would sound different, but we can still understand what the content is; if we omit enough low-frequency components, however, we would hear meaningless sound. All in all, the significance of DWT is that it can decompose the signal on different scales, and the choice of different scales depends on different applications.

2.2.4 Wavelet packet transform

Since the resolution of DWT is poor in the high - frequency region and the noise typically distributes in high frequency, this potential drawback can affect the application of the wavelet - based speech enhancement techniques. In this case, the wavelet packet decomposition is developed, which adopts redundant basis functions and offers an arbitrary time-frequency resolution. The wavelet packet decomposition can adaptively select the corresponding frequency band and signal spectrum phase according to the signal characteristics and analysis requirements.

Example: A blasting vibration signal is taken as an example to explain how the WPT works. The original signal is shown in Figure. 2.2. The sampling frequency of the data is 1024 Hz and the Nyquist frequency is $512Hz$.

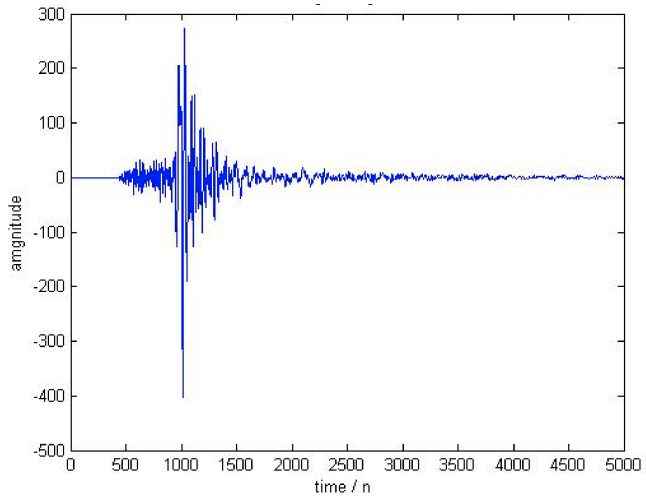


Figure 2.2: Original signal

The decomposition tree is shown in the picture of Figure. 2.3. It is decomposed into 3 levels by using Wavelet $dB5$ and each subband is denoted as a node. According to the sampling frequency of the signal, each decomposition node frequency range can be obtained. Since the signal is 3-level decomposed, $2^3 = 8$ subbands are received and each band bandwidth of $512/8 = 64Hz$. Therefore, the the node $(3, 0)$ represents a frequency range of $0 - 64Hz$, the frequency range of the node $(3, 1)$ means $65 - 128Hz$.....It is not difficult to obtain all the spectral characteristics for subband speeches as shown in Fig. 2.4 where f_m denotes the frequency $512Hz$.

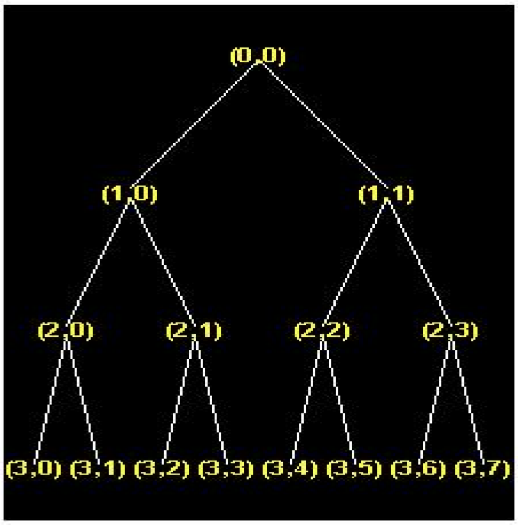


Figure 2.3: A 3-level wavelet packet decomposition tree

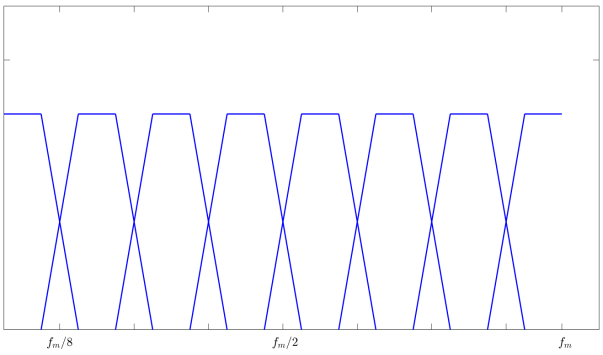


Figure 2.4: Spectral components of 3-level wavelet packet decomposition

After decomposition, the WPT coefficient for each node is shown in Fig. 2.5. It can be seen that the energy of the original signal is concentrated in the first two low frequency subbands, that is, $0 - 64HZ$ and $65 - 128HZ$.

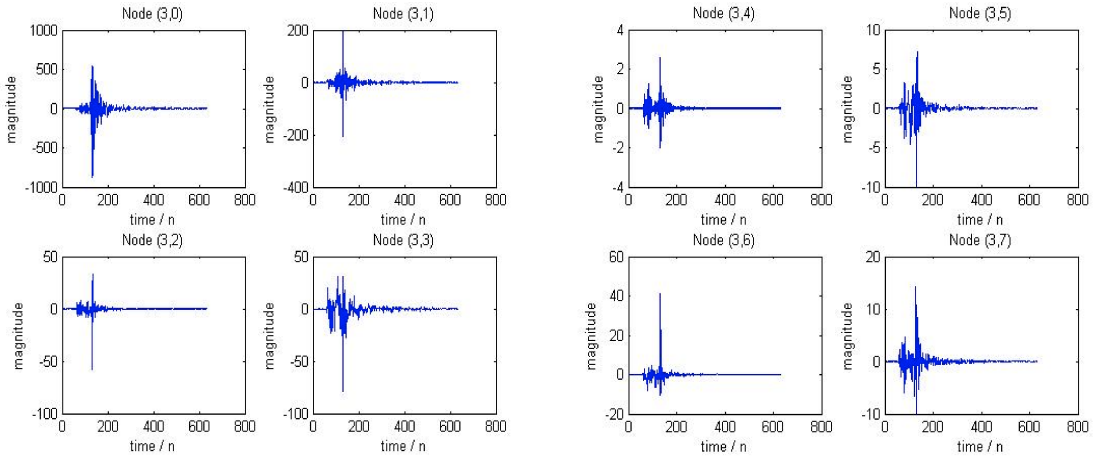


Figure 2.5: WPT subbands from node (3,0) to (3,7)

2.2.5 Comparison of wavelet transforms

Based on several wavelet transforms introduced above, the CWT and the DWT mainly differ in how they discretize the scale parameter. The CWT typically uses exponential scales with a base smaller than 2. However, the DWT always uses exponential scales with a base equal to 2. With the DWT or WPT, we can always get the same number of coefficients as that of the original signal. Since many of the coefficients may be close to zero in value, we can just throw away those coefficients, which still maintain a high-quality signal approximation. However, with the CWT, the

computational resources required is much larger than the DWT. If we process an N-length signal with I scales, we will get an N-by-I matrix, which increases the computational cost for speech processing. Since our application is to obtain the sparsest possible signal representation for speech denoising, the DWT and WPT methods are preferred in terms of the speech enhancement.

With respect to the DWT and WPT algorithms, the WPT offers full band decomposition in wavelet analysis, which provides more flexibility as to what scales need to be processed for speech enhancement. Compared with the DWT, at each level j in Fig. 2.3, the low subband and high subband are both divided into two subbands. However, in Fig. 2.1 we can see that at each level only the high subband is divided. The WPT is an attractive alternative since it divides the frequency axis into finer intervals. Considering all the conditions, the WPT is chosen for the wavelet analysis as the first step in the proposed speech enhancement in this thesis.

2.3 Wavelet based Thresholding Algorithms for Speech Enhancement

2.3.1 Hard thresholding

In hard thresholding, all coefficients below a pre-defined threshold value are set to zero. The hard thresholding function is defined as [51]

$$\bar{y}(k) = \begin{cases} \widetilde{y}(k), & |\widetilde{y}(k)| \geqslant \hat{T} \\ 0, & Others \end{cases} \tag{12}$$

An example for illustration of hard thresholding operation is shown in Fig. 2.6

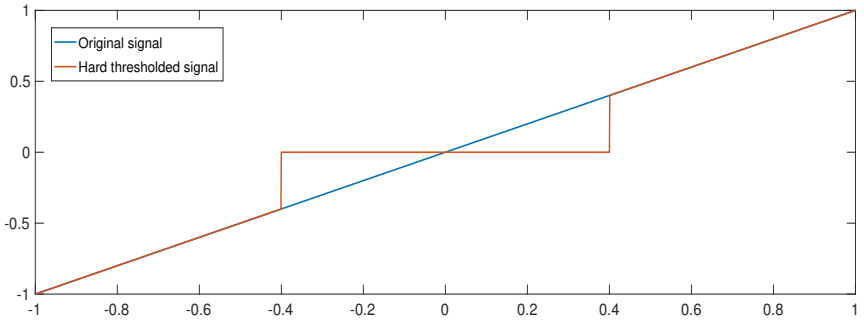


Figure 2.6: Hard threshold mapping function

In Fig. 2.6, the threshold value $\hat{T}$ is chosen to be 0.4. In general, Donoho's hard thresholding estimation method [51] preserves some of the sharp characteristics of the original signal, but the hard threshold function has worse smoothness and is discontinuous at $\hat{T}$and $-\hat{T}$. This discontinuity causes the reconstructed signal to suffer from the pseudo-Gibbs phenomenon, and many undesirable oscillations may occur, making the thresholded signal lose the smoothness of the original signal. In order to address these issues, soft thresholding technique has been proposed, which will be briefly described in the next subsection.

2.3.2 Soft thresholding

Soft thresholding has become a very popular tool in computer vision and machine learning. The soft thresholding function is defined as

$$\bar{y}(k) = \begin{cases} \widetilde{y}(k) - \hat{T}, & \widetilde{y}(k) \geqslant \hat{T} \\ 0, & -\hat{T} \leqslant |\widetilde{y}(k)| \leqslant \hat{T} \\ \widetilde{y}(k) + \hat{T}, & \widetilde{y}(k) \leqslant -\hat{T} \end{cases} \tag{13}$$

An example for illustration of soft thresholding operation is shown in Fig. 2.7

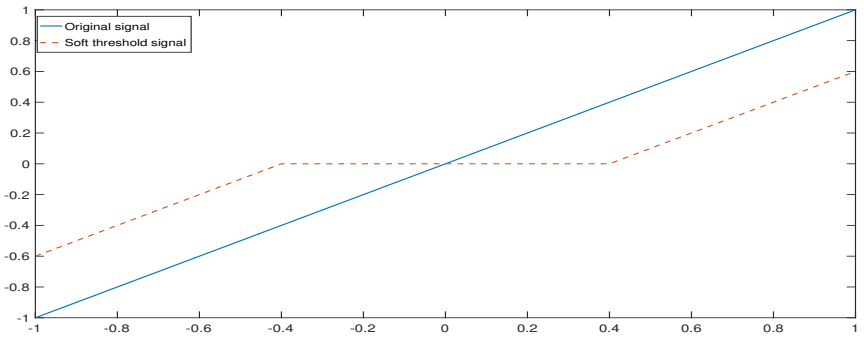


Figure 2.7: Soft threshold mapping function

In Fig. 2.7, the threshold value is also chosen to be 0.4. The soft-threshold method retains the good overall continuity of the wavelet coefficients, which makes the estimated signal to not produce additional oscillations. However, when the sample value is larger than the threshold value $\hat{T}$, there

is always a constant deviation between the estimated and actual values, and the derivative of the soft threshold function is not continuous, which directly affects the degree of approximation between the reconstructed signal and the real signal. Thus the soft threshold has certain limitations. To avoid an abrupt value change, a non-linear thresholding method has been presented as shown below.

2.3.3 Non-linear thresholding

In non-linear thresholding, a smooth function is used to map the original function to the new one. This mapping function is described by

$$\bar{y}(k) = \begin{cases} \widetilde{y}(k), & |\widetilde{y}(k)| \geqslant \hat{T} \\ sgn(k)\frac{|k|^3}{\hat{T}^2}, & |\widetilde{y}(k)| < \hat{T} \end{cases} \tag{14}$$

An example for illustration of non-linear thresholding operation is shown in Fig. 2.8, where the threshold value is also set to be $\hat{T} = 0.4$.

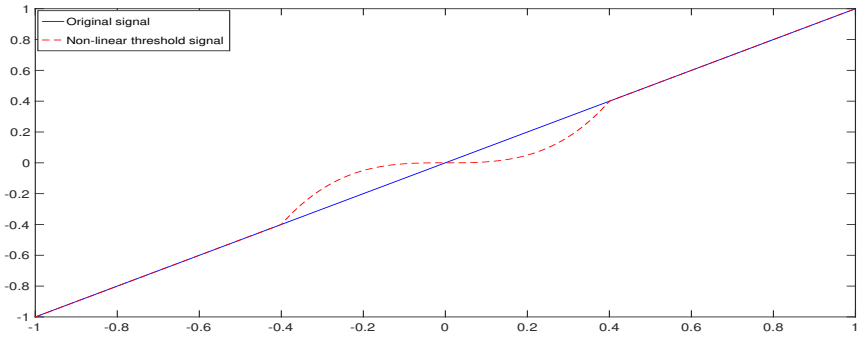


Figure 2.8: Non-linear threshold mapping function

The speech signal $\tilde{y}(k)$ is compared sample by sample with $\hat{T}_n$ to surppress the values to some extent or remain the values, yielding a modified signal $\bar{y}(k)$, where $sgn(k)\frac{|k|^3}{\hat{T}^2}$ denotes a non-linear function employed to avoid the musical noise, and in general it gives smaller values than $\widetilde{y}(k)$ when the sample value is below the threshold $\hat{T}$.

2.3.4 Threshold value selection

For any of these approaches, the threshold parameter may be either a fixed value or a level-dependent value that is a function of the wavelet decomposition level. One of the key elements for successful wavelet denoising is the selection of the threshold value. In this subsection, the choosing of the threshold value is discussed. The proper value for the threshold can be determined in many ways as will be shown in the following.

(1) A universal threshold for the fast wavelet transform is proposed in [52]. The shrinkage rule is near optimal in the minimax sense and the universal threshold is defined as

$$\hat{T} = \sigma\sqrt{2ln(N)}, \tag{15}$$

where $\sigma = \frac{\text{MAD}}{0.6745}$[51], N is the length of noisy speech y and σ the noise level. Here, MAD is the median of the absolute value of the wavelet's coefficients of the first scale.

(2) For the WPT, the threshold is improved and determined as

$$\hat{T} = \sigma\sqrt{2\log(N\log_2 N)}. \tag{16}$$

(3) The so-called SURE method is derived from minimizing Stein's unbiased risk estimator [53], which uses the following thresholding value

$$\hat{T} = \underset{0\leq\hat{T}\leq\sigma\sqrt{2\log N}}{\arg\min}\{\sigma^2 N + \sum_{(n=1)L}^{N}[y[k],\hat{T}^2) - 2\sigma^2 I(|y[k]|\leq\hat{T}]\}, \tag{17}$$

(4) Proposed adaptive thresholding method: In the above basic wavelet thresholding methods, some inevitable defects exist for the speech corrupted by real-life noise. Firstly, in real environment most kinds of noises are non-stationary noise, which cause bad speech quality if we apply basic thresholds for denoising. Secondly, we need to consider the voiced and unvoiced speech separately since some unvoiced speech segments also contain noise like speech components. From the perspective of practical applications, the time-adaptation of the threshold, which takes the time behaviour of the noisy signal into account, constitutes an interesting approach. Suppose $y_n(k)$ is

processed by using the subband VAD scheme along with adaptive thresholding in the WP domain on a frame by frame basis, where n is the frame index. First, the noise and voice frames are detected a VAD method. Based on each subband frame activity, the estimated noise variance (σ^2_{v,n_l}) and the frame-dependent threshold (T) are updated as

$$\sigma^2_{v,n_l} = \lambda\sigma^2_{v,n_l-1} + (1-\lambda)\sum_{(n_l-1)L}^{n_lL} \tilde{y}^2_{n_l}(k), \tag{18}$$

$$T_{n_l}=\lambda T_{n_l-1}+(1-\lambda)\frac{\text{MAD}_{n_l}}{0.6745}\sqrt{2\log_{10}(L\log_2 L)}, \tag{19}$$

where n_l is the index of the noise only frame detected and $\sum_{(n_l-1)L}^{n_lL} \tilde{y}^2_{n_l}(k)$ denotes the power of the newly detected noise frame.

$(\text{MAD}_{n_l}/0.6745)\sqrt{2\log_{10}(L\log_2 L)}$ is given based on [54]. MAD_{n_l} is the the median absolute value of the n_l-th detected noise frame. Scalar λ is a scaling factor which affects the estimation accuracy. It is noted that n_l is the index of the detected noise frame, meaning that the values of the estimated noise variance (σ^2_{v,n_l}) and the frame-dependent threshold (T) can only be updated when the incoming frame is a noise frame based on the VAD result. Segmental SNR is also updated for every subband frame along with the corresponding noise variance $\sigma^2_{v,n}$, which is presented as

$$SNR_{seg,n} = 10\log_{10}\sum_{(n-1)L}^{nL} \tilde{y}^2_n(k)/\sigma^2_{v,n}. \tag{20}$$

It is noted that the segmental SNR reflects noise portion in speech frames, which is a good feature to decide the threshold value for each subband frame.Then an adaptive threshold $\hat{T}_n$ based on each frame segmental SNR and the basic threshold value is defined as

$$\hat{T}_n = \begin{cases} T_n + T_n e^{-SNR_{seg,n}/\tau}, & SNR_{seg,n} \geqslant 0 \\ 2T_n, & SNR_{seg,n} < 0 \end{cases} \tag{21}$$

The nonlinear function $e^{-SNR_{seg,n}/\tau}$ is used to gradually suppress the value of $\hat{T}_n$ as $SNR_{seg,n}$ increases, where τ is a pre-determined constant which is set to $\tau = 2$ in our work. In this case, if subband frame $SNR_{seg,n}$ is equal to or higher than 10dB, which means the subband frame has

relatively more voice portion, then, $e^{-SNR_{seg}/\tau} = e^{-5}$ is almost equal to zero and $\hat{T}_n$ is taken as T_n.

2.4 Performance Evaluation of the Wavelet Thresholding Schemes

From the discussion in the previous section on different thresholding schemes, it is obvious that the thresholded signal with non-linear thresholding is most close to the original signal as compared to the other thresholding methods. As such, it is expected that using the non-linear function would cause less time-frequency discontinuities in the enhanced speech spectrum than the hard-thresholding method does. Therefore, non-linear thresholding method is applied to speech enhancement in this thesis.

With regards to the selection of the threshold value, let us evaluate first the performance of the proposed thresholding method. Suppose 20 speech utterances, including 10 male and 10 female speakers, are generated from TSP database[55], and white noise, non-stationary noise and babble noise taken from NOISEX-92[56] are added to the clean speech signal with different input SNRs ranging from $-10dB$ to $10dB$. The speech is sampled at 16kHZ and the frame size is set as 512 samples. A 3-level WP decomposition tree with Daubechies 1 wavelet is applied on the noisy speech. To objectively evaluate the performance of these methods, we take the segmental signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) as the criteria for the performance evaluation. The definitions of these evaluation methods are given below:

- *Segmental SNR*: Since classical SNR does not correlate well with speech quality for a wide range of distortions. One variation, i.e., segmental signal-to-noise ratio (SNR), has been proposed as objective measure. It is defined as the average of the measurements of SNR over short frames and computed using both distorted and undistorted (clean) speech samples. The frame length is normally set between 15 and $20ms$. The segmental SNR is defined as

$$SNR_{seg,n} = 10\log_{10} \sum_{k=(n-1)L}^{nL} \frac{\tilde{y}_n^2(k)}{(\hat{y}_n(k) - \tilde{y}_n(k))^2} \tag{22}$$

- *PESQ*: The Perceptual Evaluation of Speech Quality (PESQ) [57] is one of the most sophisticated

and accurate estimation methods to test the speech quality. Its value is based on the comparison with the traditional MOS (Mean Opinion Score) method in which a number of listeners are used to rate the voice quality, and it ranges from -0.5 (bad) to 4.5 (excellent). The block digram of the PESQ measurement is shown in Fig. 2.9 [58]

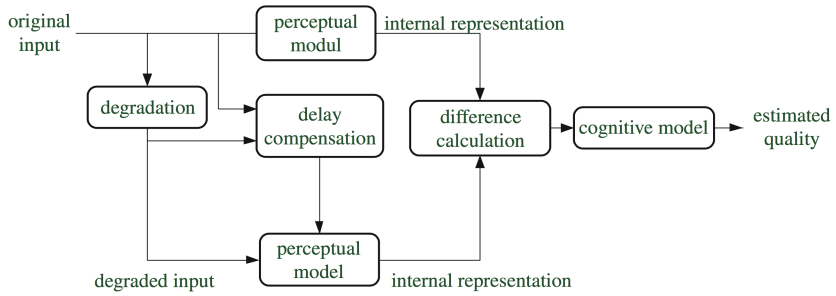


Figure 2.9: Diagram of the PESQ system

In our experiment, we compare the performances of four wavelet thresholding algorithms in three types of noise environments: (1) the non-linear universal threshold (U-T); (2) the non-linear WPT threshold (WPT-T); (3) the non-linear SURE threshold (SURE-T); (4) the proposed non-linear adaptive threshold (AT).

Firstly, we compare the performances in terms of segmental SNR which mainly reflects the ability for speech denoising. The simulation results are given respectively in Fig. 2.10, Fig. 2.11 and Fig. 2.12.

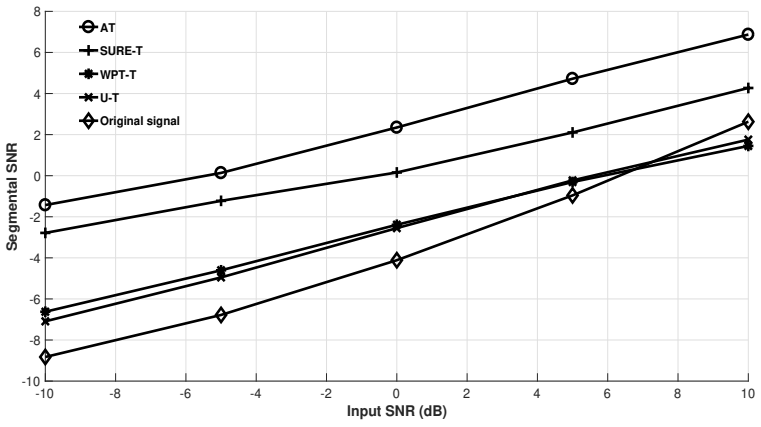


Figure 2.10: Segmental SNR of enhanced speech in white noise at -10, -5, 0, 5, 10dB input SNR levels

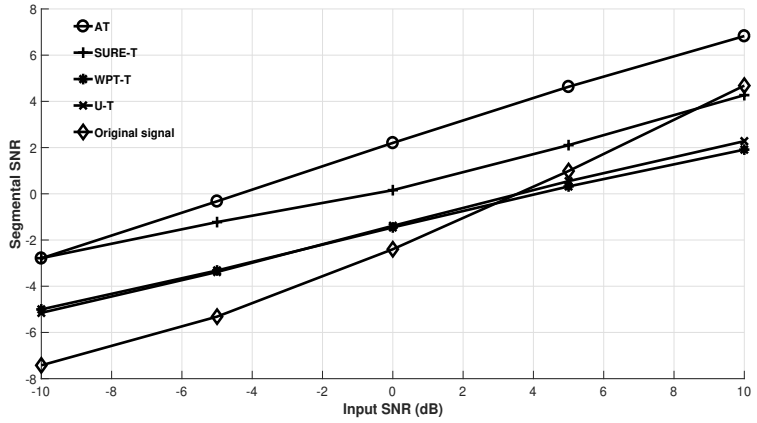


Figure 2.11: Segmental SNR of enhanced speech in non-stationary noise at -10, -5, 0, 5, 10dB input SNR levels

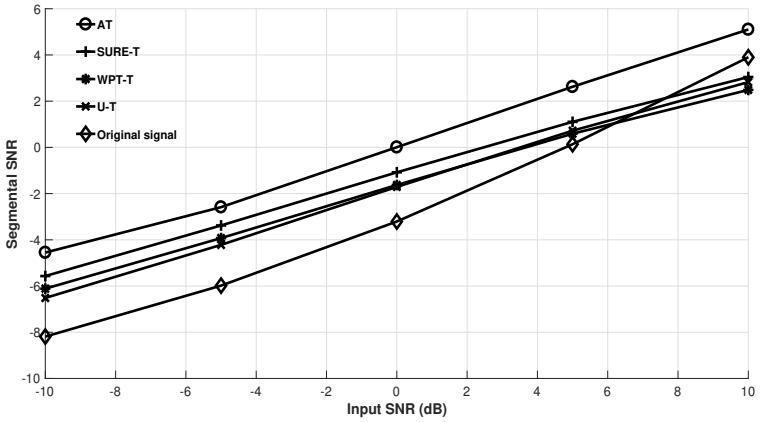


Figure 2.12: Segmental SNR of enhanced speech in babble noise at -10, -5, 0, 5, 10dB input SNR levels

The experimental results show that the proposed thresholding scheme (AT) consistently outperforms the existing methods in terms of segmental SNR for all noise types. It is observed that AT performs well even in lower input SNRs for different kinds of noise environment. For white noise corrupted speech, the segmental SNR improvement for the proposed scheme is around 6dB, which is at least 1dB higher than the SURE-T scheme. In the non-stationary noise case, it is noted that the proposed AT performs better especially in higher SNR region, while other three schemes degrade the speech quality. In babble noise environment, the three existing methods perform worse at higher input SNRs and the improvement of segmental SNRs is less than 3dB. However, the proposed AT

improves the segmental SNR by 3.5dB on the average. Through a large number of simulation experiments, we have found the proposed scheme can efficiently enhance the speech for a wide range of input SNRs in different kinds of noise environment compared with other existing methods.

Secondly, we evaluate the performances in terms of PESQ which mainly measures the ability of speech enhancement system in terms of speech quality. The simulation results are given respectively in Fig. 2.13, Fig. 2.14 and Fig. 2.15.

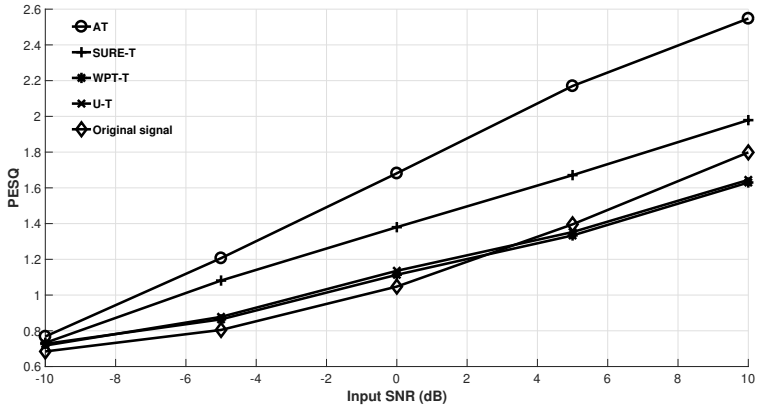


Figure 2.13: PESQ of enhanced speech under white noise at -10, -5, 0, 5, 10dB input SNR levels

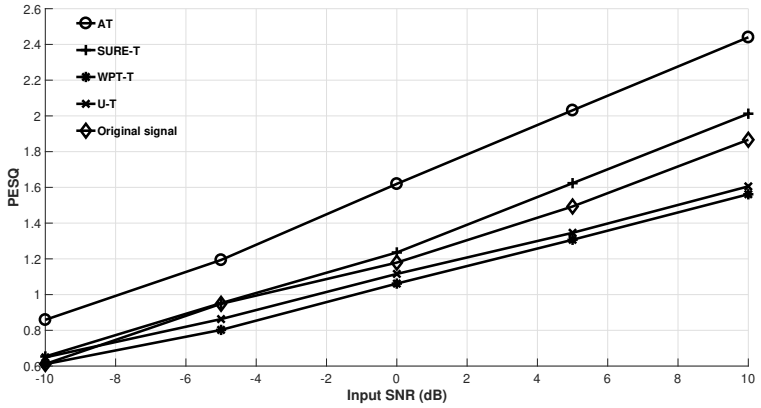


Figure 2.14: PESQ of enhanced speech under non-stationary noise at -10, -5, 0, 5, 10dB input SNR levels

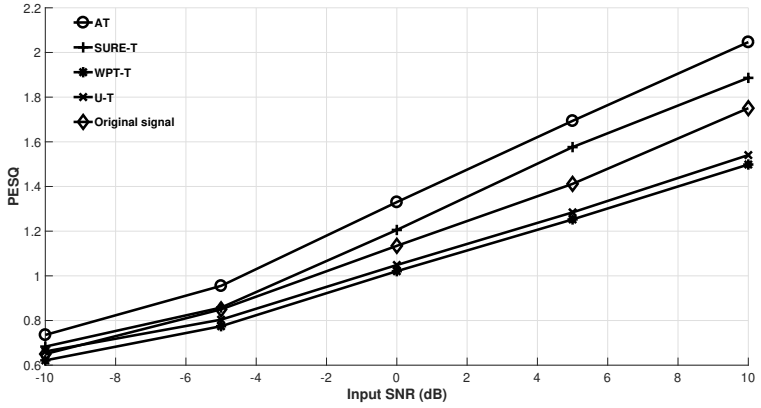


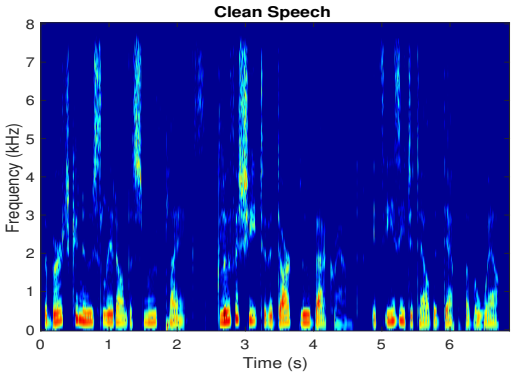
Figure 2.15: PESQ of enhanced speech under babble noise at -10, -5, 0, 5, 10dB input SNR levels

The experimental results show that the proposed method can significantly improve the PESQ for all input SNRs compared with other existing methods. In the white noise case, the PESQ value for all the methods has not been improved considerably at lower SNRs. However, when the input SNR increases, the proposed method (AT) improves the PESQ considerably. For the non-stationary noise, the proposed method offers a stationary PESQ improvement along with the increase of input SNRs. For the babble noise case, it is observed that the PESQ improvement is small when the input SNR is -10dB. But after -5dB, the PESQ improvement is 0.4dB on the average.

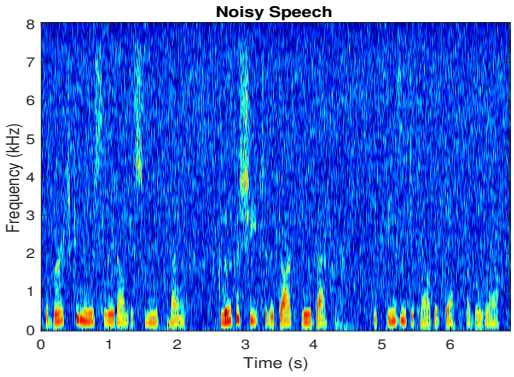
Despite the above objective evaluations, the spectrograms have also been investigated for comparison of different methods in terms of their performance improvements. For performing these experiments, 30 speech sentences are taken from the TSP database. The main goal of this simulation study is to show the advantage of the proposed method (AT). The spectrograms for the clean, noisy and enhanced speeches by using the above 4 methods in the presence of white noise and non-stationary noises at 5dB input SNR are shown below in detail.

Firstly, the white noise is added to the clean speech. The spectrograms of the clean speech, noisy speech and all the enhanced speeches generated by the four algorithms are illustrated in Fig. 2.16. As we can see from Fig. 2.16c and Fig. 2.16d, lots of noise, especially in high frequency region, could not be removed by the U-T and WPT-T methods. Fig. 2.16e shows a better result in denoising but some speech in high frequency have also been removed, degrading the speech quality. However, the proposed method (AT) as seen in Fig. 2.16f has significantly reduced the high frequency noise

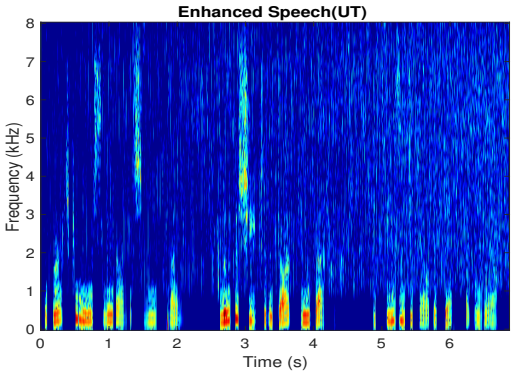
and the patterns of the characteristics are much clearer.



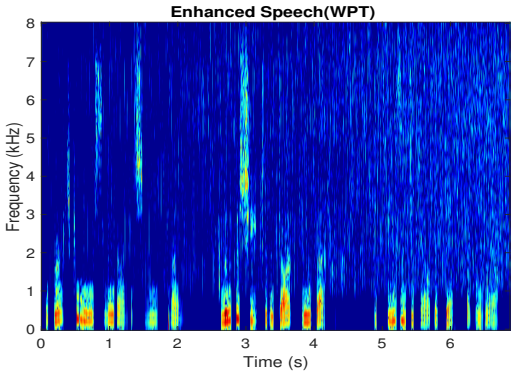
(a) Clean speech spectrum



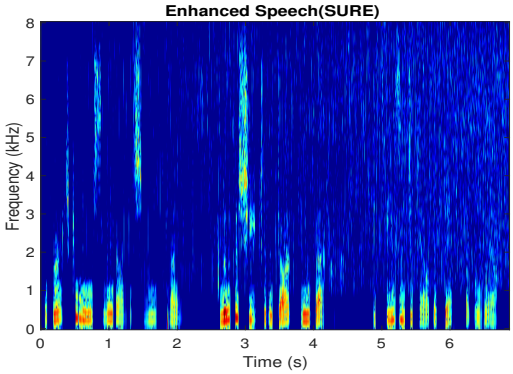
(b) Noisy speech spectrum



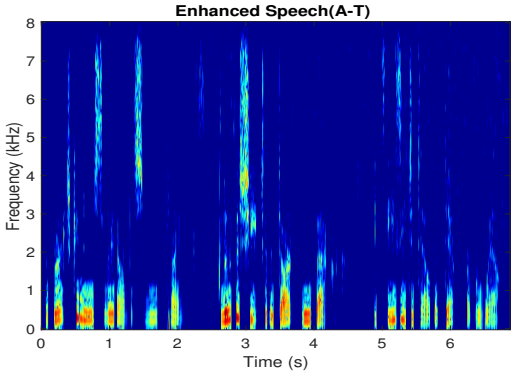
(c) Enhanced speech spectrum using the UT



(d) Enhanced speech spectrum using the WPT



(e) Enhanced speech spectrum using the SURE-T

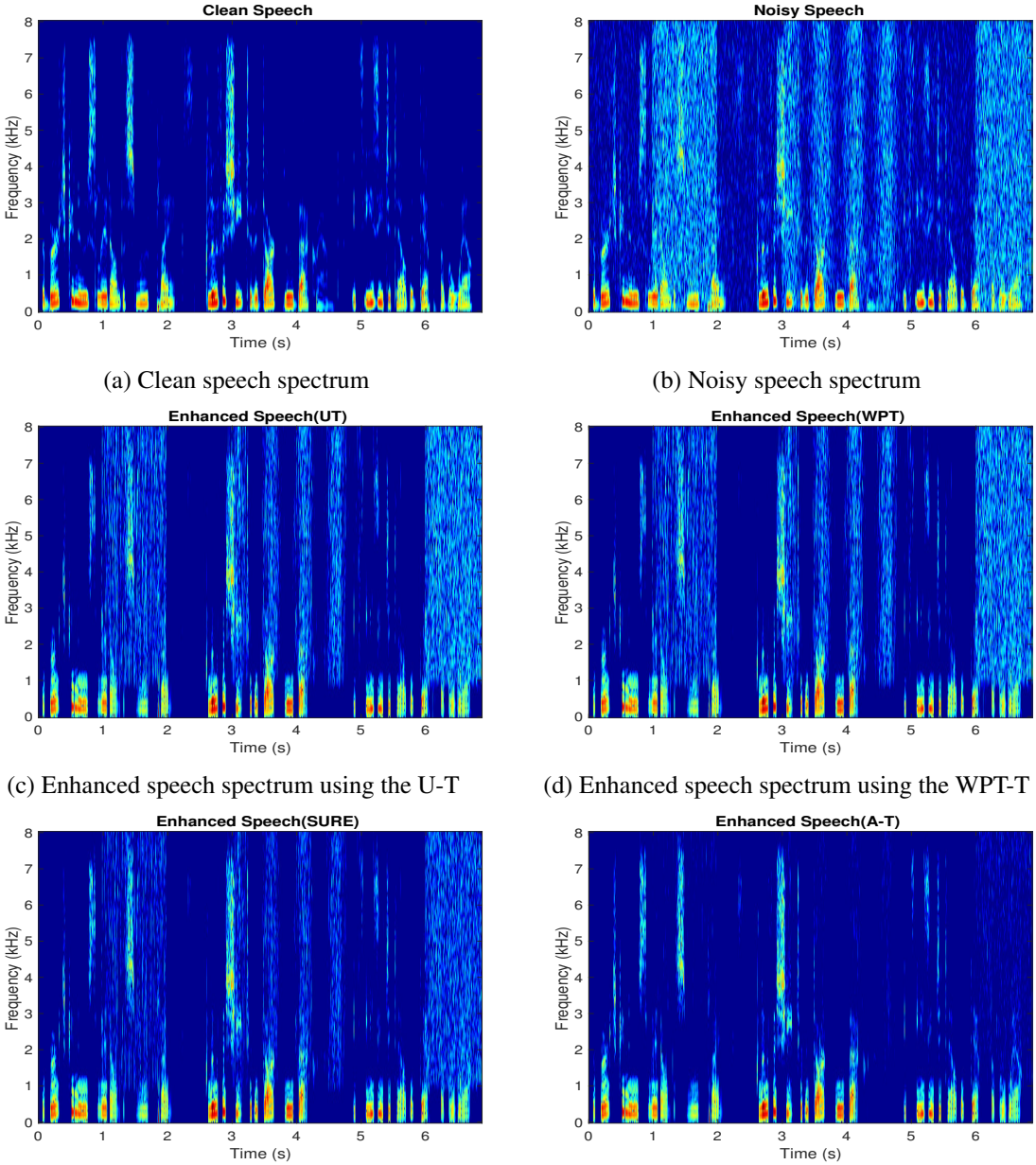


(f) Enhanced speech spectrum using the AT

Figure 2.16: Speech spectrums under white noise

Secondly, the non-stationary noise is added to the clean speech. The spectrograms of the clean speech, noisy speech and all the enhanced speeches generated by the four algorithms are illustrated in Fig. 2.17. Among the existing methods, it is observed that the SURE method in Fig. 2.17e performs better than the UT and WPT methods generally. During the $1-2$ and $6-7$ seconds of the

speech, it is obvious that only the proposed method (AT) in Fig. 2.17f successfully removes most of the noise and retains the high quality of speech in the high frequency region.



(a) Clean speech spectrum

(b) Noisy speech spectrum

(c) Enhanced speech spectrum using the U-T

(d) Enhanced speech spectrum using the WPT-T

(e) Enhanced speech spectrum using the SURE-T

(f) Enhanced speech spectrum using the AT

Figure 2.17: Speech spectrums under the non-stationary noise

2.5 Conclusion

In this chapter, we have discussed different kinds of wavelet transforms, several thresholding schemes and the selection of the threshold values. First, we introduced the basic wavelet transforms: CWT and DWT, based on which, BWT and WPT are presented as modifications to further improve the basic wavelet transforms. Since the WPT requires less computational resources compared with the BWT, it is selected as wavelet analysis for our speech enhancement.

In the second part of this chapter, hard thresholding, soft thresholding and non-linear thresholding with different threshold values are introduced. A wavelet packet based non-linear thresholding algorithm is proposed. In order to compare different thresholding schemes, an extensive simulation study has been conducted for the evaluation of the proposed methods in the presence of three different noises for a wide range of input SNRs. The performances have been evaluated and compared in terms of spectrograms and two objective measurements: segmental SNRs and PESQ of the enhanced speech. The experimental results have revealed that the non-linear thresholding with adaptive threshold value achieves the best performance in terms of all the performance metrics. In addition, among several existing methods, the non-linear thresholding with SURE threshold performs better than the non-linear thresholding with soft threshold and that with hard threshold.

Chapter 3

Proposed Speech Enhancement Algorithm

3.1 Introduction

This chapter gives a detailed description of the proposed speech enhancement method [59]. The input noisy speech $y(k)$ is first segmented into frames $y_n(k)$, where n is the frame index. The subsequent processing is then carried out on a frame by frame basis. The proposed approach consists of two successive stages. In the first stage, the WP transform is first applied to the noise corrupted speech frame, which decomposes each frame into a number of subbands. For each subband, a voice activity detector (VAD) is designed to detect the voiced/unvoiced parts of the speech. Based on different voice activity, an adaptive thresholding scheme is then utilized to each subband speech to obtain the pre-enhanced speech. In the second stage, the reconstructed and pre-enhanced full-band speech is processed by the IKF for further speech enhancement. The details of the proposed method are presented in the following sections.

3.2 Speech Subband Decomposition with Wavelet Packet Transform

In this section, the input noisy speech $y_n(k)$ is first decomposed into equal bandwidth subband speeches through the wavelet packet transform (WPT). The wavelet filter-bank in general is an

array of band-pass filters that separates the input signal into multiple components, each carrying a single frequency subband of the original signal [60]. It is worth mentioning that the number of decomposition channels (subbands) in WP analysis is usually a power of two, which can easily be implemented by several levels of decomposition, each level creating twice subbands. In this thesis, we adopt a 3-level WP decomposition, yielding a total of 8 subbands. The structure of the wavelet decomposition is shown in Fig. 3.1, where a three-level decomposition is performed. Clearly each decomposition creates two equal subbands, called low-frequency and high-frequency components denoted by L and H respectively. Each subband signal can further be decomposed into another two subbands.

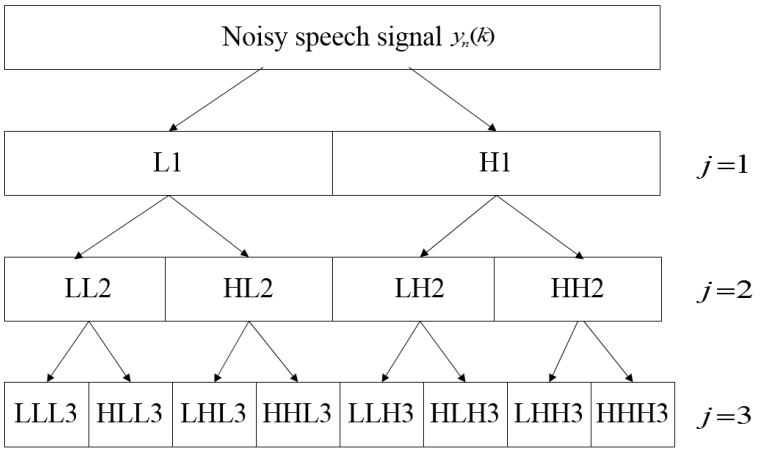


Figure 3.1: 3-level wavelet packet decomposition structure

In general, each subband frequency interval can be described by

$$[\frac{if_s}{2^{j+1}}, \frac{(i+1)f_s}{2^{j+1}}), i = 0, 1, 2, ..., 2^j - 1, \tag{23}$$

where f_s is the sampling frequency, i is the subband index and j is the wavelet decomposition level.

Each decomposition subband is described as $\widetilde{y}_n^{(i)}(k)$. After processing all the subbands by using the proposed speech enhancement scheme, a modified subband speech $\bar{y}_n^{(i)}(k)$ is yielded. The WP reconstruction is adopted in order to reconstruct a full-band speech signal $\hat{y}_n(k)$. The block-diagram of wavelet packet decomposition and wavelet packet reconstruction is shown in Fig. 3.2.

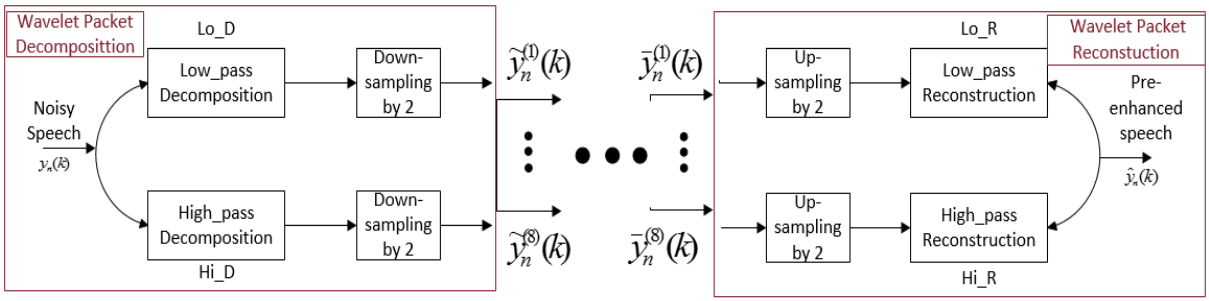


Figure 3.2: Wavelet packet decomposition and reconstruction

3.3 VAD - Based Adaptive Thresholding Scheme

A VAD - based adaptive thresholding scheme is then applied to each subband $\tilde{y}_n^{(i)}(k)$. As each subband $\tilde{y}_n^{(i)}(k)$ goes through the same VAD based thresholding scheme, we will drop the subband index i in the following discussions. Fig. 3.3 shows the flowchart of the VAD based adaptive thresholding approach.

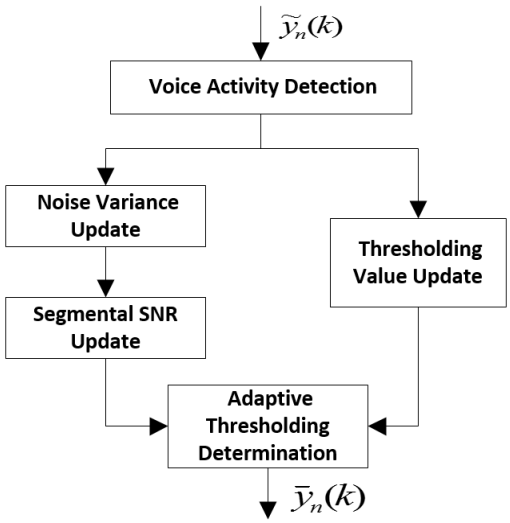


Figure 3.3: Flowchart of VAD based adaptive thresholding

Based on the above flowchart, each subband frame $\tilde{y}_n^{(i)}(k)$ is first going through the VAD. If the subband frame is determined as a noise frame, the segmental SNR and thresholding value will be updated, then an appropriate thresholding scheme is applied to each noisy speech sample. Based on its corresponding thresholding value, some speech samples are suppressed and the other speech samples are retained.

3.3.1 Voice activity detection

The main idea of the VAD scheme is to extract the measured features from the input noisy speech, and compare the feature values with thresholds, which are computed from noise-only frames. To design a good VAD algorithm, the following aspects should be taken into account:

- An appropriate decision rule should be employed to decide whether the noisy speech frames are voiced frames and noise frames.
- The VAD scheme should adapt to noisy speeches under the non-stationary noise environment.
- The computational cost should be considered for the implementation of the VAD scheme, especially for real-time applications.

The basic VAD design is shown in Fig. 3.4. A voiced frame is detected if the measured values exceed the thresholds. Otherwise, the input speech frame is considered as a noise frame. When the VAD is performed, a voice frame is flagged as VAD = 1, while a noise frame is marked as VAD = 0.

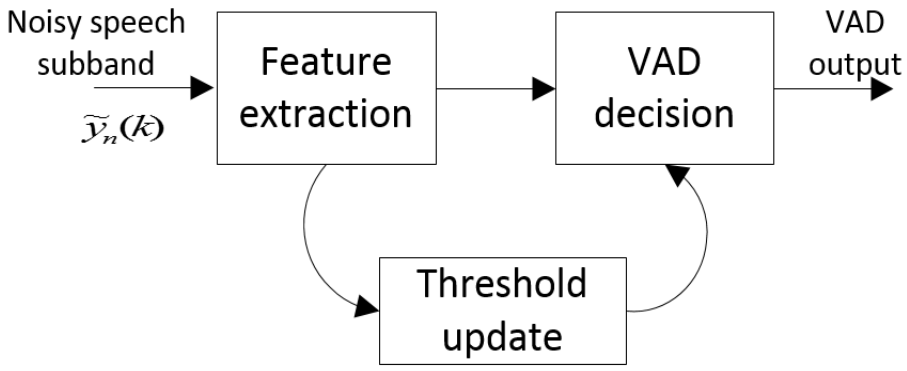


Figure 3.4: A block diagram of a basic VAD design

In our VAD scheme, we adopt the energy E_n as the first feature for each frame. As the frame energy ignores the frequency features, the energy based feature is not sufficient for input speech that has lower SNRs, but contains rich frequency information. Therefore, we would like to take frequency based features into account. This feature is referred to as the spectral flatness (F). The two measured features are defined as

- *Frame energy (E_n)*:

$$E_n = \sum_{k=(n-1)L}^{nL} |\tilde{y}_n(k)|^2, \tag{24}$$

where L is the number of samples in a frame, and n denotes the frame index.

- *Spectral flatness (F_n)*:

$$F_n = 10\log_{10}(\frac{m_a}{m_g}), \tag{25}$$

where m_a and m_g, respectively, denote the arithmetic and geometric means of the noisy speech spectrum, which are calculated as

$$m_a = \frac{1}{L} \sum_{k=(n-1)L}^{nL} S(w_k), \tag{26}$$

$$g_a = \sqrt[L]{\prod_{k=(n-1)L}^{nL} S(w_k)}, \tag{27}$$

where $S(w_k)$ is denoted as the short-time spectrum and can be estimated by applying the Welch-Bartlett method, namely, by averaging the spectral estimates of a certain number (M) of consecutive frames. The expressions are

$$S(k, w_k) = \frac{1}{M} \sum_{k=(n-M+1)}^{n} |X(k, w_k)|^2, \tag{28}$$

where $X(k, w_k)$ is the short-time Fourier transform coefficient at frequency w_k of the kth frame [61]. Spectral flatness indicates the width and noisiness of the spectral power. A low spectral flatness indicates that the spectral power is concentrated in a relatively small number of bands, which behaves more like voice frames. However, a high spectral flatness shows that the spectrum power is more uniform in different frequency bands, and appears relatively flat and smooth, which appears more likely as noise.

For the proposed VAD algorithm, we first consider the threshold initialization. For each decomposed subband, we calculate the two features according to (24) and (25) for the first N frames, then the minimum value of each feature among these frames is taken as the initial thresholding value as denoted by $E_{T,0}$ and $F_{T,0}$ respectively,

$$E_{T,0} = min\{E_0, E_1, ..., E_N\}, \tag{29}$$

$$F_{T,0} = min\{F_0, F_1, ..., F_N\}, \tag{30}$$

The VAD process starts with calculating the two features for frame n ($n \geq 1$) obtained from (24) and (25), which results in E_n and F_n. Both feature values will start to compare with the initial thresholding values $E_{T,0}$ and $F_{T,0}$, respectively. As suggested in [62], if the two feature values exceed the thresholds $E_{T,0}$ and $F_{T,0}$ respectively, the speech frame n is marked as a voice frame and the two thresholding values are not updated. Otherwise, frame n is marked as a noise frame, and the two thresholding values are then updated as

$$E_{T,n_l} = 40 \log(\frac{(n_l - 1)E_{T,n_l-1} + E_{n_l}}{n_l}) + E_{T,0}, \tag{31}$$

$$F_{T,n_l} = \alpha\, F_{T,n_l-1} + (1 - \alpha)F_{n_l} + F_{T,0}, \tag{32}$$

where n_l is the index of the noise only frame detected and α is the exponential smoothing factor.

The proposed complete VAD algorithm is shown in Fig. 3.5.

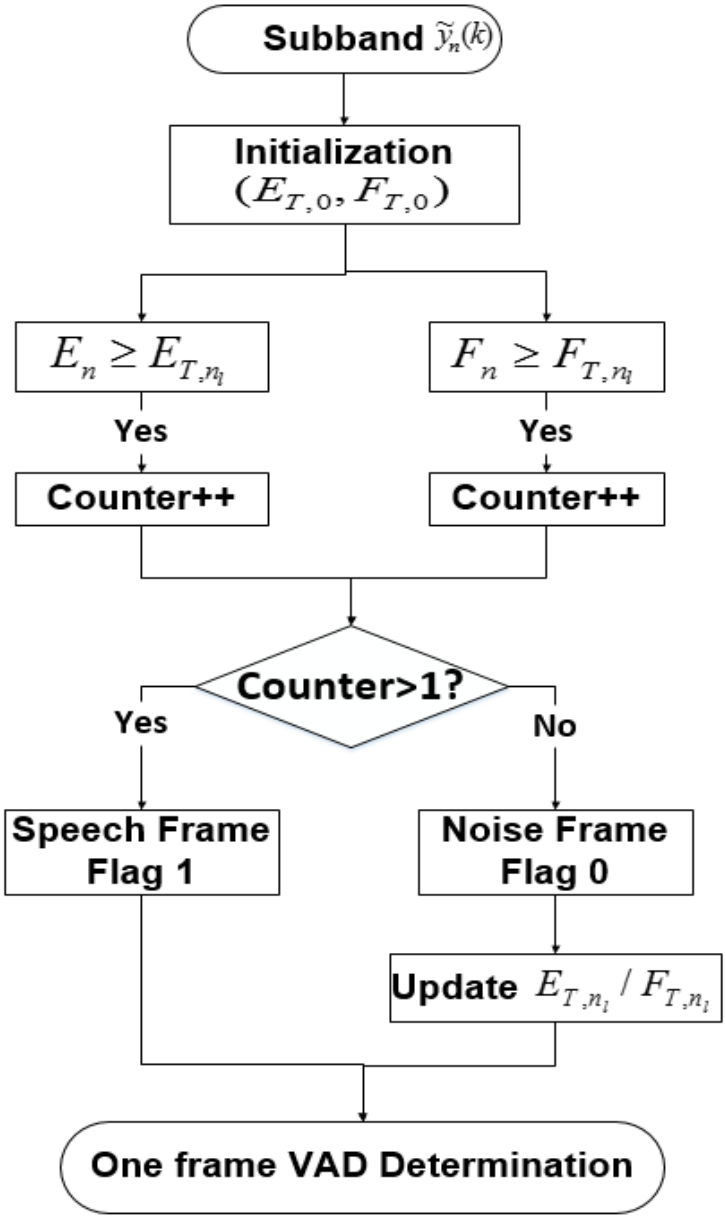


Figure 3.5: A flowchart of proposed VAD algorithm

Fig. 3.6 shows an example of the proposed VAD results. The detected noisy speech has 10 frames and each frame length is $L = 64$. As we can see, the frames 1, 2, 6 and 7 are marked as noise frames, while frames $3-5$ and $8-10$ are detected as voice frames.

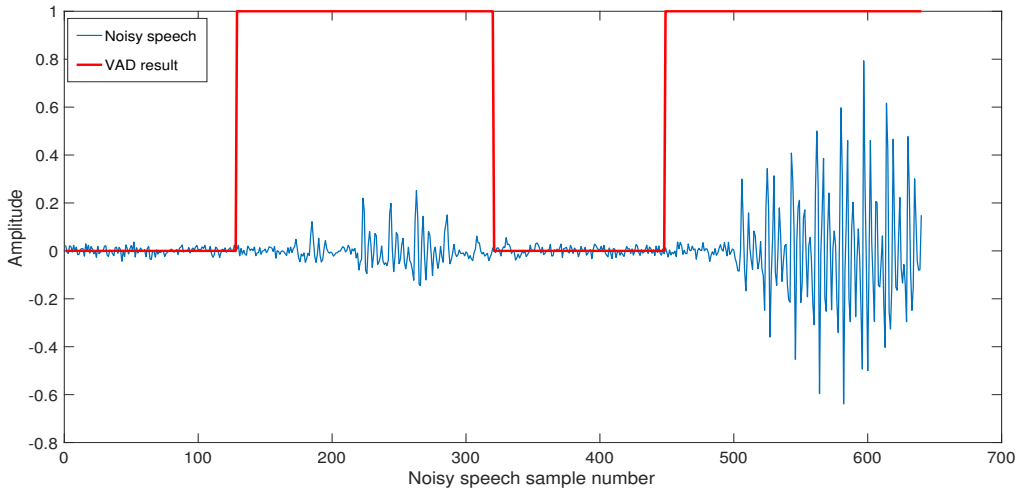


Figure 3.6: Performance of VAD algorithm

3.3.2 Adaptive thresholding

Following the VAD step, the noise and voice frames are detected.When a noise frame is detected, the estimated noise variance (σ^2_{v,n_l}) and the frame-dependent threshold value (T) are updated as

$$\sigma^2_{v,n_l} = \lambda\sigma^2_{v,n_l-1} + (1-\lambda)\sum_{k=(n_l-1)L}^{n_lL} \widetilde{y}^2_{n_l}(k), \tag{33}$$

$$T_{n_l}=\lambda T_{n_l-1}+(1-\lambda)\frac{\text{MAD}_{n_l}}{0.6745}\sqrt{2\log_{10}(L\log_2 L)}, \tag{34}$$

where n_l is the index of the detected noise frame and $\sum_{(n_l-1)L}^{n_lL} \tilde{y}^2_{n_l}(k)$ denotes the power of the newly detected noise frame.

The term $(\text{MAD}_{n_l}/0.6745)\sqrt{2\log_{10}(L\log_2 L)}$ is achieved based on [54], where MAD_{n_l} is a robust measure of the variability of the n_l-th detected noise frame, which is defined as

$$\text{MAD}_{n_l} = median(|\widetilde{y}_{n_l} - median(\widetilde{y}_{n_l})|). \tag{35}$$

The parameter λ in (33) and (34) is a scaling factor which affects the estimation accuracy. It is noted that n_l is the index of the detected noise frame, which means the values of estimated noise variance (σ^2_{v,n_l}) and the frame-dependent threshold (T) will be updated when the coming frame is a noise frame based on the VAD result, which makes the thresholding value adaptive for each frame.

Segmental SNR for each subband frame is also updated along with the corresponding noise variance $\sigma^2_{v,n}$, which is presented as

$$SNR_{seg,n} = 10\log_{10} \sum_{k=(n-1)L}^{nL} \tilde{y}_n^2(k)/\sigma^2_{v,n}. \tag{36}$$

It is noted that segmental SNR reflects noise portion in speech frames, which is a proper feature to decide the threshold value for each subband frame. Based on each frame's segmental SNR, an adaptive threshold $\hat{T}_n$ is proposed as

$$\hat{T}_n = \begin{cases} T_n + T_n e^{-SNR_{seg,n}/\tau}, & SNR_{seg,n} \geqslant 0 \\ 2T_n, & SNR_{seg,n} < 0 \end{cases} \tag{37}$$

where T_n is the threshold value for each subband frame including both noise and voice frames. The nonlinear function $e^{-SNR_{seg,n}/\tau}$ is used to gradually suppress the value of $\hat{T}_n$ when $SNR_{seg,n}$ increases. We found from experimentation that $\tau = 2$ is a very good choice. In this case, if subband frame $SNR_{seg,n}$ is equal to or higher than 10dB, the subband frame contains relatively more voice. Thus, $e^{(-SNR_{seg}/\tau)} = e^{-5} \approx 0$ and $\hat{T}_n$ is taken as T_n. After getting the adaptive threshold value $\hat{T}_n$, the subband $\tilde{y}_n(k)$ is compared sample by sample with $\hat{T}_n$ to either suppress the values to some extent or retain the same values, giving the modified subband $\bar{y}_n(k)$ speech as shown below

$$\bar{y}_n(k) = \begin{cases} \widetilde{y}_n(k), & |\tilde{y}_n(k)| \geqslant \hat{T}_n \\ sgn(k)\frac{|k|^3}{\tilde{y}_n(k)^2}, & |\tilde{y}_n(k)| < \hat{T}_n \end{cases} \tag{38}$$

where $sgn(k)\frac{|k|^3}{\hat{T}_n^2}$ denotes a non-linear function which is employed to avoid the musical noise. It decreases the values of WP coefficients $\widetilde{y}_n(k)$ when the sample value is smaller than $\hat{T}_n$. The inverse

WP transform is then applied to each subband $\bar{y}_n(k)$ in order to reconstruct the pre-enhanced full-band speech signal $\hat{y}(k)$.

3.4 Iterative Kalman Filter

In this section, an iterative Kalman filter (IKF) is presented as the second stage for the proposed speech enhancement technique. Consider a time-domain pre-enhanced speech $\hat{y}(k)$ which has been obtained from the first stage and can be written as

$$\hat{y}(k) = s(k) + w(k), \tag{39}$$

where $s(k)$ is the k^{th} sample of the clean speech and $w(k)$ is the noise samples. Note that the IKF is also conducted on the frame by frame basis but the frame index is omitted in the following formulation for the sake of notational simplicity.

The clean speech $s(k)$ is modeled as a P^{th} order auto-regressive process that is given by

$$s(k) = \sum_{t=1}^{P} a_t s(k-i) + u(k), \tag{40}$$

where a_t denotes the t^{th} linear prediction coefficient (LPC), $u(k)$ and $w(k)$ are uncorrelated Gaussian white noise sequences with zero mean and the variances ${\sigma_u}^2$ and ${\sigma_w}^2$, respectively. Then the noisy speech in (39) can also be modeled as the state-space model (SSM),

$$\hat{y}(k) = \boldsymbol{H}\boldsymbol{x}(k) + w(k), \tag{41}$$

$$\boldsymbol{x}(k) = \boldsymbol{F}\boldsymbol{x}(k-1) + \boldsymbol{G}u(k), \tag{42}$$

with $\boldsymbol{H} = \boldsymbol{G}^T = [\,1, \ldots, 1\,] \in \mathbb{R}^{1\times p}$, $\boldsymbol{x}(k) = [s(k-p+1), ..., s(k)] \in \mathbb{R}^{1\times p}$ and $\boldsymbol{F}$ denotes the $p \times p$

state transition matrix composed of the LPCs, namely,

$$\boldsymbol{F} = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_{p-1} & -a_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{43}$$

Since the LPCs of the clean speech are not available, one can use the Modified Yule-Walker equations [63] to obtain the estimate of LPCs, ie,

$$\begin{bmatrix} \hat{a}_p \\ \vdots \\ \hat{a}_1 \end{bmatrix} = \begin{bmatrix} r_{yy}(1) & \cdots & r_{yy}(p) \\ \vdots & \ddots & \vdots \\ r_{yy}(p+l) & \cdots & r_{yy}(2p+l-1) \end{bmatrix}^{\dagger} \begin{bmatrix} r_{yy}(p+l) \\ \vdots \\ r_{yy}(2p+l) \end{bmatrix} \tag{44}$$

where $r_{yy}(m) = E[y(s)y(s-m)]$ is the observation autocorrelation function, and $[.]^{\dagger}$denotes the pseudoinverse operator.

The IKF process contains two loops of iterations, called inner and outer loops, for each frame. For inner loop, the operation principle includes a prediction step and a measurement update step. In the prediction step, the IKF predicts the state vector and parameter covariance by using the previous samples of the sate-space model. The estimate of clean speech $\hat{\boldsymbol{x}}$ and the posteriori estimation error covariance $\boldsymbol{P}(k|k)$ are predicted from time step $(k-1)$ to step k (the status are $\hat{\boldsymbol{x}}(k|k-1)/\boldsymbol{P}(k|k-1)$),

$$\hat{\boldsymbol{x}}(k|k-1) \quad = \boldsymbol{F}\hat{\boldsymbol{x}}(k-1|k-1), \tag{45}$$

$$\boldsymbol{P}(k|k-1) \quad = \boldsymbol{F}\boldsymbol{P}(k-1|k-1)\boldsymbol{F}^T + {\sigma_u}^2\boldsymbol{G}\boldsymbol{G}^T. \tag{46}$$

In the measurement update step, the Kalman gain and state vectors are updated as

$$\boldsymbol{K}(k) \quad = \boldsymbol{P}(k|k-1)\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}(k|k-1)\boldsymbol{H}^T + {\sigma_w}^2)^{-1}, \tag{47}$$

$$\hat{\boldsymbol{x}}(k|k) \quad = \hat{\boldsymbol{x}}(k|k-1) + \boldsymbol{K}(k)(y(k) - \boldsymbol{H}\hat{\boldsymbol{x}}(k|k-1)), \tag{48}$$

$$\boldsymbol{P}(k|k) \quad = (\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{H})\boldsymbol{P}(k|k-1), \tag{49}$$

where $\boldsymbol{I}$ denotes the identity matrix.

In the KF process, the Kalman gain $\boldsymbol{K}(k)$ is chosen to minimize the a posteriori error covariance $\boldsymbol{P}(k|k)$. The state space model parameters of the KF are updated sample-by-sample through an iterative procedure. The additive noise components are reduced significantly when the inner loop is completed for one entire frame. After one frame iteration, the estimated speech frame is achieved as

$$\breve{y}(k) = \boldsymbol{H}\hat{\boldsymbol{x}}(k|k). \tag{50}$$

It is observed from equation (48) that when $\boldsymbol{K}(k)$ decreases, the priori state estimate is trusted more and the noisy measurement becomes less important.

For the outer loop iteration, the LPCs and other state-space model parameters are re-estimated from the same processed speech frame. The number of outer loop iterations is usually set to be 2 or 3. The iterative procedure stops when the pre-set maximum number of iterations is exhausted, giving the further enhanced result of the same speech frame with respect to the input pre-enhanced speech. At the end of processing all speech frames, the ultimate enhanced speech $\breve{y}(k)$ is obtained. The flowchart of iterative Kalman Filter processing is shown in Fig. 3.7.

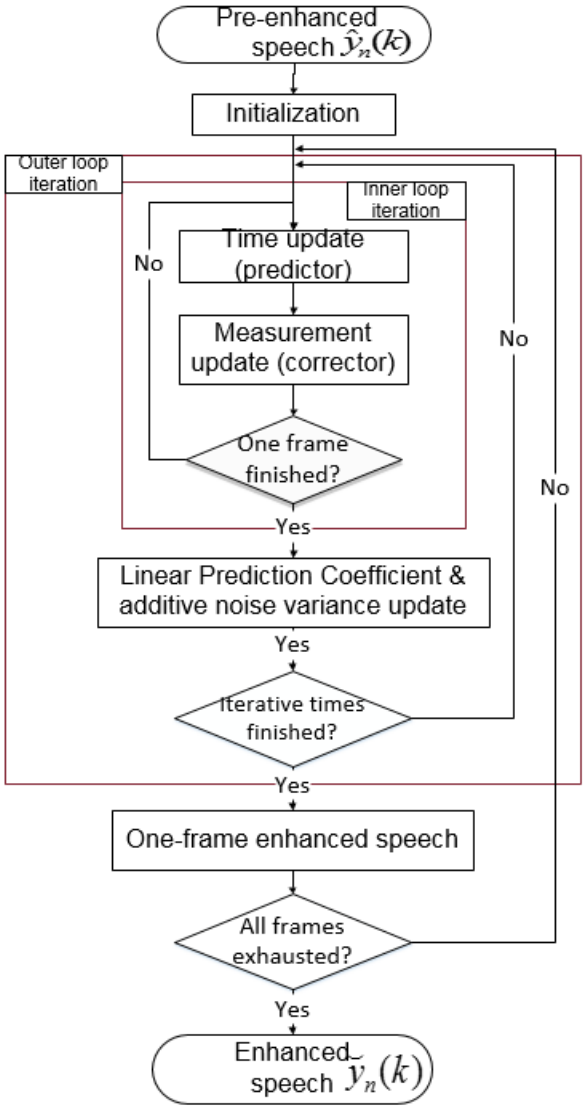


Figure 3.7: Flowchart of Iterative Kalman filter algorithm

It is deserved to point out that the adaptive wavelet thresholding and the IKF each can serve as a stand alone method. They can also be combined as one improved speech enhancement method. The overall speech enhancement algorithm based on adaptive thresholding with IKF is summarized in the following Algorithm.

THE PROPOSED SPEECH ENHANCEMENT ALGORITHM

Stage I: Noisy speech pre-enhancement

P1. *WP transform*: Decompose noisy speech y into 8 levels of frequency subbands.

P2. *VAD based Adaptive Thresholding*: Each subband $\tilde{y}(k)$ is processed through the VAD based adaptive WP thresholding algorithm, getting 8 modified subbands. Each modified subband is defined as $\bar{y}(k)$.

- VAD: Apply (24) and (25) as two features to detect voiced and noise frames of each subband based on features adaptation (31) and (32).
- Thresholding: Adjust each sample of subband based on adaptive threshold following equations (33) – (38).

P3. *Inverse WP transform*: Reconstruct a full-band speech signal $\hat{y}(k)$.

Stage II: Iterative Kalman Filter

I1. *Inner loop iteration*

- Prediction step: Apply (45) and (46) to predict the state vector and parameter variance.
- Measurement update step: Update the Kalman Gain and state vectors following the equations (47) – (49).

I2. *Outer loop iteration*: Re-estimate the LPCs and other state-space model parameters for 2 or 3 times. After processing all the speech frames, the ultimate enhanced speech $Y(k)$ is obtained.

3.5 Performance Evaluation of the Proposed Methods

In this section, we carry out a simulation study to evaluate the performance of the proposed adaptive thresholding method and its combination with iterative Kalman filter as an improved method. We also make comparison with the existing IKF method. In this simulation study, we use the same clean speech and noise database as in Chapter 2, as well as the same parameters setup. In addition, for VAD initialization, we pick the minimum frame feature value from the first $S = 10$ frames. it is noted that the effect of the initial feature value is often negligible even if the assumption of the noise frames in the beginning is not true. We found that the smoothing parameter $alpha$ and the scaling parameter λ perform reliably when they are in the range of $[0.9 \sim 0.95]$. The LPC order

considered in this simulation is set to $P = 8$. The proposed adaptive thresholding method (AT) and its combination with iterative Kalman filter (AT-IKF) are evaluated and compared with the iterative Kalman filter (IKF) method.

Fig.3.8 shows the segment SNR of the enhanced speech resulting from the three methods together with that of the original noisy speech in the white noise case. It is seen that the segmental SNR performances of AT-IKF, AT and IKF are similar. However, the PESQ performance for AT-IKF is the best as compared with the other methods as seen from Fig.3.9, especially in lower SNRs. Fig.3.10 and Fig.3.11 give the comparisons of segment SNR and PESQ in the non-stationary noise case. It is evident that the proposed AT-IKF method demonstrates a good performance gain in terms of PESQ. Fig.3.12 and Fig.3.13 depict the performances of all methods in babble noise. The proposed scheme (AT-IKF) outperforms the other methods both in terms of segmental SNR and PESQ. Overall, the proposed method consistently outperforms the existing methods in terms of segmental SNR and PESQ for all three noise types.

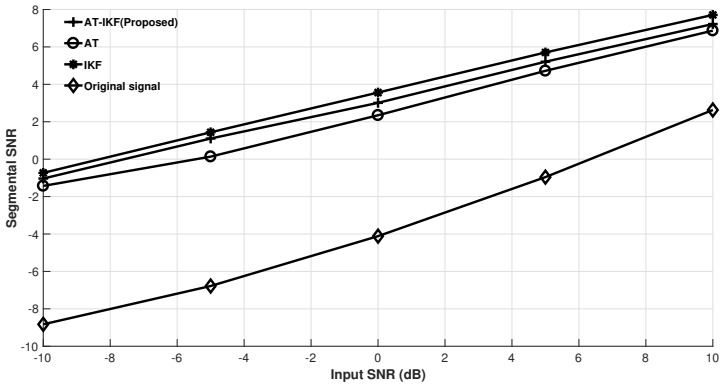


Figure 3.8: Segmental SNR performance comparison in white noise with input -10, -5, 0, 5, 10dB.

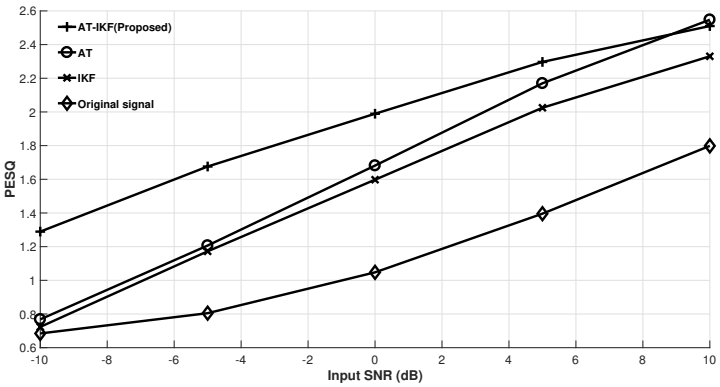


Figure 3.9: PESQ performance comparison in white noise with input -10, -5, 0, 5, 10dB.

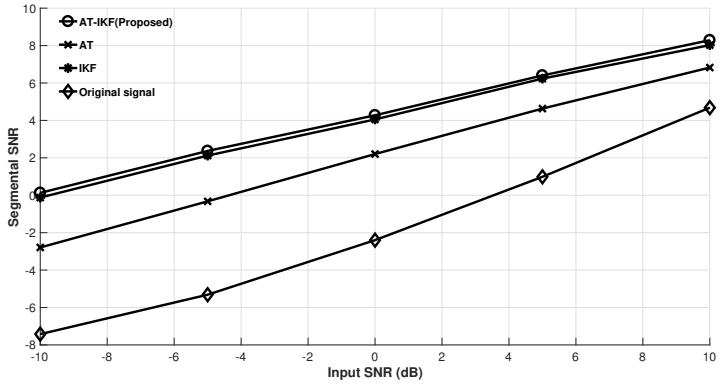


Figure 3.10: Segmental SNR performance comparison in non-stationary noise with input -10, -5, 0, 5, 10dB.

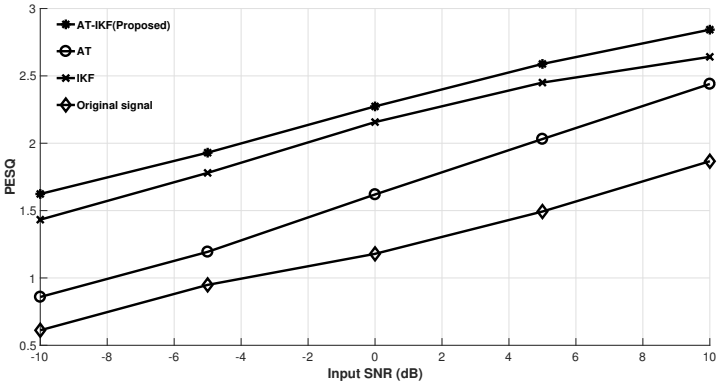


Figure 3.11: PESQ performance comparison in non-stationary noise with input -10, -5, 0, 5, 10dB.

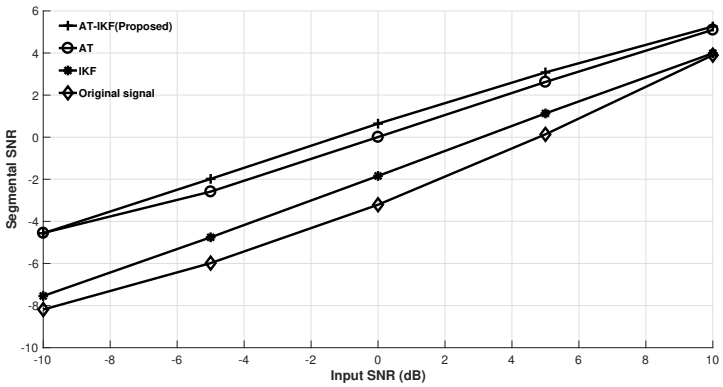


Figure 3.12: Segmental SNR performance comparison in babble noise with input -10, -5, 0, 5, 10dB.

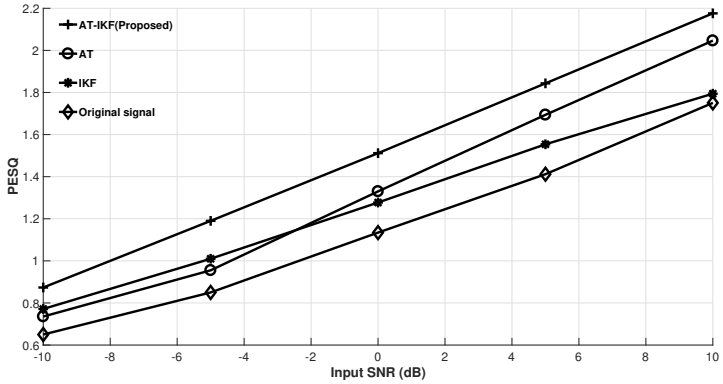


Figure 3.13: PESQ performance comparison in babble noise with input -10, -5, 0, 5, 10dB.

3.6 Conclusion

In this chapter, we have first proposed a VAD based adaptive WP thresholding scheme with the IKF. The noisy speech was first decomposed into 8 subbands. Two features have been chosen for the VAD to detect whether each subband speech frame is a voiced or noise frame. Based on the VAD results, the threshold was updated for each frame of different subbands, and then applied to adjust the speech samples in each subband frame. Through the inverse WP transform, a pre-enhanced whole- band speech has been received. The IKF was then used to further enhance the speech. As the AT scheme has made use of the subband properties of noisy speech for preliminary noise reduction and the full band speech was used as input to the IKF, the proposed method has reduced

the computational complexity compared with conventional subband KF based methods. Moreover, a pre-enhanced speech, rather than the direct noisy speech, was processed by the IKF, thus increasing the estimation accuracy of LPCs and noise variance. As verified by extensive simulations, based on segmental SNR and PESQ evaluations, the proposed method efficiently improved the speech for a wide range of input SNRs and different kinds of noise environments.

Chapter 4

Performance Evaluation and Discussion

To illustrate that the proposed adaptive thresholding plus iterative Kalman filter method outperforms other existing competitive methods, extensive computer simulations are conducted in this chapter.

4.1 Experimental Setup

In this simulation study, the female and male speech samples are generated from TSP database [55]. White noise, non-stationary noise and babble noise sequences taken from NOISEX-92 [56] are added to the clean speech signal with different input SNRs ranging from -10dB to 10dB. The speech data is sampled at the rate of 16kHZ and the size of each frame is 512 samples. A three level WP decomposition tree with Daubechies 1 wavelet is applied on the noisy speech. In addition, For VAD initialization, we pick the minimum frame feature value from the first $N = 10$ frames. This implies that we assume the beginning frames to be noise frame. Note that the effect is often negligible even if the noise frame assumption in the beginning is not satisfied. We found that the smoothing parameter α and the scaling parameter λ perform reliably when they are in the range of [0.9, ..., 0.95]. The LPC order considered in this simulation is set to $P = 8$. The proposed adaptive thresholding iterative Kalman filter (AT-IKF) method is evaluated in the time domain, where no overlapping during frame segmentation is considered and the rectangular window is fitted appropriately for each frame.

4.2 Performance Comparison between Proposed and Existing Methods

4.2.1 State of the art for comparison

To illustrate the efficiency of the proposed method, three representative and popular existing methods are evaluated and compared with the proposed adaptive thresholding iterative Kalman filter (ATIKF) method. The first comparative method is an improved spectral subtraction [ISS] based speech enhancement [15], which used smoothed spectrums to approximate the speech and noisy spectrums with auto-regressive (AR) model and construct speech codebook and noise codebook. The second comparative method is an improved Wiener filter [IWF] based speech enhancement [64], which applied the subband Wiener filter with pitch synchronous analysis. The third method for comparison is a subband iterative Kalman filter [SIKF] based speech enhancement [42], which used a partial reconstruction scheme based on consecutive mean squared error combined with the subband iterative Kalman filter.

Since Segmental SNR mainly reflects the ability of speech denoising and PESQ mainly reflects the speech quality, we first choose Segmental SNR and PESQ as two evaluation parameters in Sec. 4.2.2 to measure above mentioned speech enhancement algorithms. Sec. 4.2.3 shows the time waveforms and spectrograms of clean speech, noisy speech and enhanced speech processed by all the methods in comparison under different kinds of noise environments.

4.2.2 Segmental SNR and PESQ comparisons

Firstly, we compare the performances in terms of segmental SNR and PESQ in white noise environment. The simulation results are given respectively in Fig. 4.1 and Fig. 4.2. In white noise environment, the proposed method (AT-IKF) outperforms other existing methods especially in lower input SNRs. For PESQ, it is observed that the proposed method provides around 0.5 score improvement than the SIKF method at the input SNR of $-10dB$. In addition, at higher input SNRs, such as 5dB and 10dB, the SIKF method performs slightly better than the proposed method. However, the proposed method achieves a much better PESQ score compared with the

SIKF method. Generally speaking, among the existing methods, the iterative KF based methods perform relatively better than the spectral subtraction based method and Wiener filter based method. In terms of both Segmental SNR and PESQ values, it is clearly observed that the proposed method (AT-IKF) offers overall the best speech enhancement performance among a wide range of input SNRs.

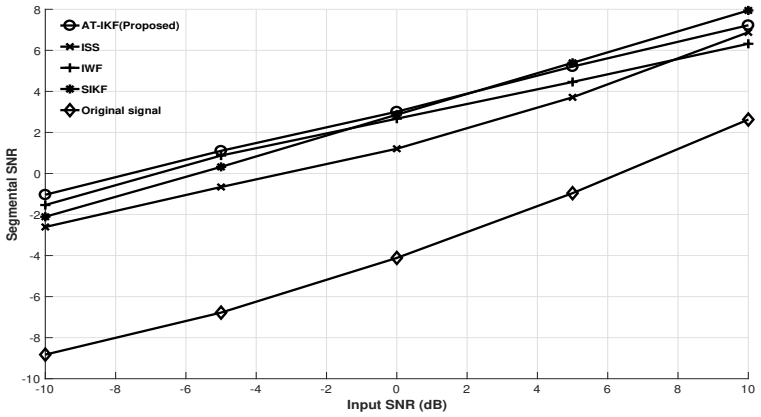


Figure 4.1: Segmental SNR results of enhanced speech in white noise case at -10, -5, 0, 5, 10dB input SNR levels

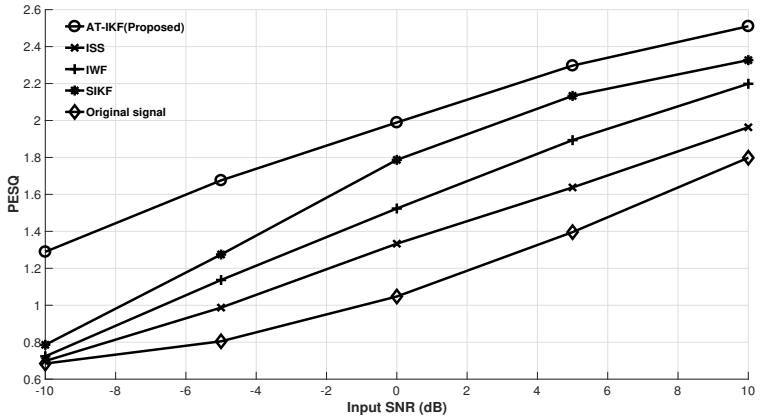


Figure 4.2: PESQ of enhanced speech in white noise case at -10, -5, 0, 5, 10dB input SNR levels

Secondly, we compare the performances in terms of segmental SNR and PESQ in non-stationary noise environment. The simulation results are given respectively in Fig. 4.3 and Fig. 4.4. In the non-stationary noise environment, it is obvious that the proposed method and the SIKF method outperform other existing methods in terms of Segmental SNR and PESQ. As to the segmental

SNR improvement, the SIKF method provides better performance with the increase of input SNR. However, it is clearly indicated that the enhanced speech of the SIKF bears more speech distortion in higher input SNRs based on the PESQ scores, which made the enhanced speech quality worse. Since the proposed method offers more speech quality improvement, it is a relatively best speech enhancement choice.

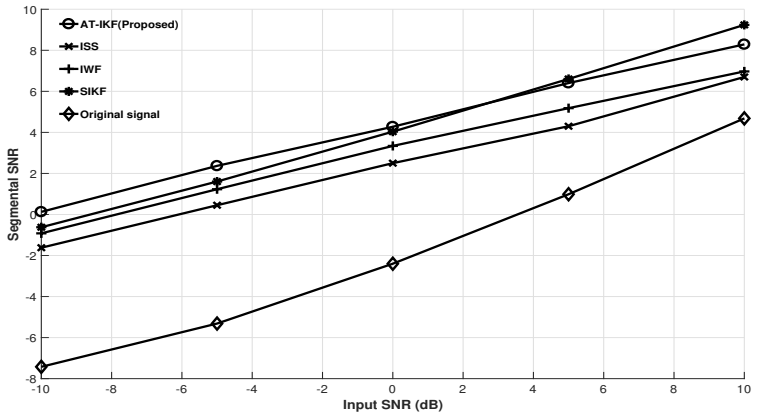


Figure 4.3: Segmental SNR results for non-stationary noise case at -10, -5, 0, 5, 10dB input SNR levels

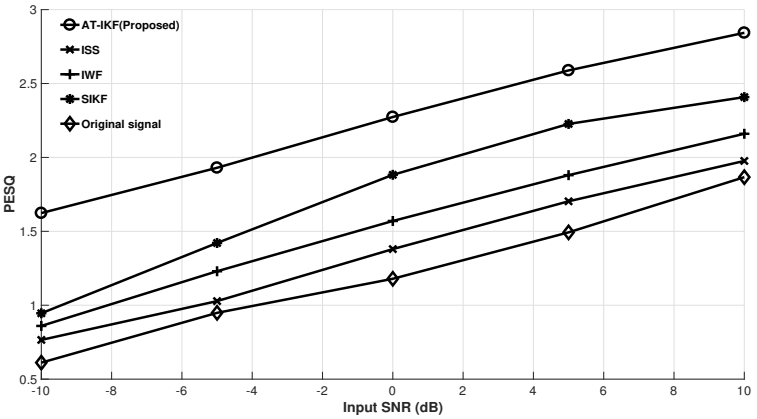


Figure 4.4: PESQ results for non-stationary noise case at -10, -5, 0, 5, 10dB input SNR levels

Thirdly, we compare the segmental SNR and PESQ performances of the four methods in babble noise. The simulation results are given respectively in Fig. 4.5 and Fig. 4.6. It is evident that the proposed method provides a significant segmental SNR and PESQ improvement as opposed the three existing methods for all the range of input SNRs. Especially for the PESQ improvement of

the proposed method, it is seen from Fig.4.6 that at 5dB or higher input SNR, the proposed method provides at least 0.4 score more than the SIKF method, which is a significant PESQ improvement. Overall, the adaptive thresholding plus iterative Kalman filter method performs the best as compared with the subband iterative Kalman Filter method, the spectral subtraction based method and the Wiener filter based method.

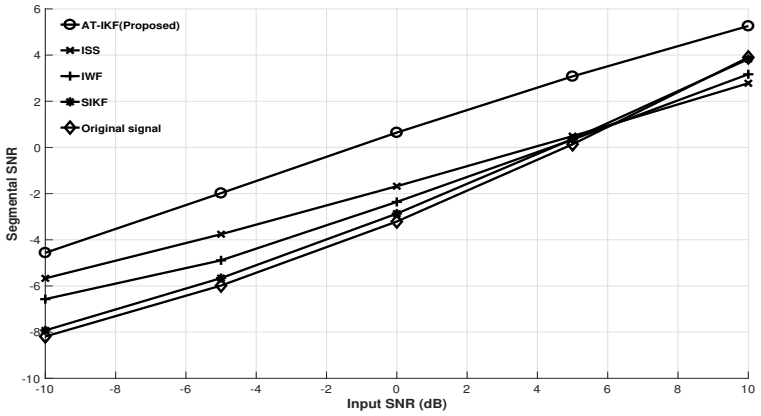


Figure 4.5: Segmental SNR of the enhanced speech in babble noise case at -10, -5, 0, 5, 10dB input SNR levels

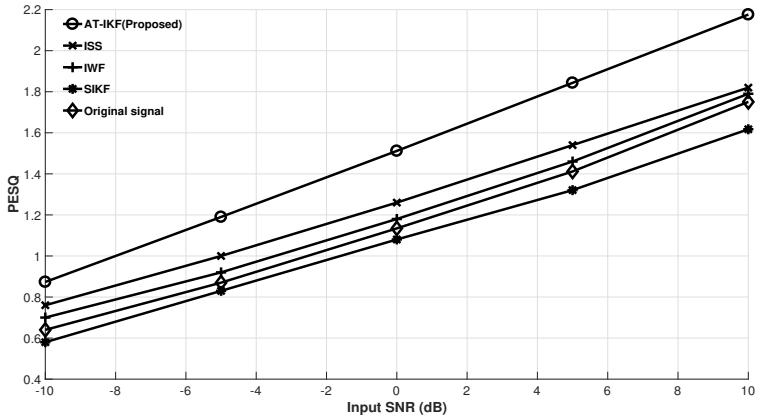


Figure 4.6: PESQ results of the enhanced speech in babble noise case at -10, -5, 0, 5, 10dB input SNR levels

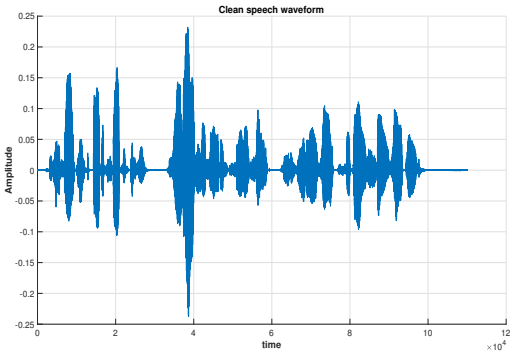
Based on the segmental SNR and PESQ performance comparisons for all the speech enhancement algorithms under three kinds of noisy cases, the simulation results show that the proposed method outperforms other existing methods. It is worth mentioning that we have also found some

other problems for the existing methods through extensive experiments. For the ISS method, the music noise is introduced, which adversely influences the enhanced speech quality. For the IWF method, a priori SNR with adaptive parameter is applied for each subband. However, it is relatively difficult to achieve a high accuracy of the pitch period estimation in noise corrupted environment, which influences the performance of noise reduction for speech enhancement. For the SIKF method, it offers limited enhancement performance for noisy speeches which contain non-negligible noises in low-frequency region. In general, the proposed method gives less speech distortions but better perceived quality, which is the best speech enhancement algorithm for all the noise cases in terms of all the evaluation metrics.

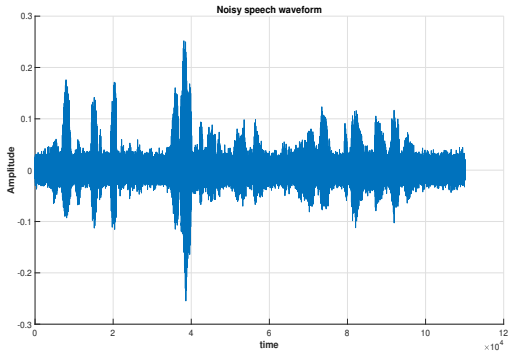
4.2.3 Speech waveforms and spectrograms comparisons

Despite the above objective evaluations, the waveforms and spectrograms have also been investigated for the comparison of different methods concerning their performance improvements. For performing these experiments, 30 speech sentences are take from the TIMIT database [65] for white noise environment and 30 speech sentences are take from the TSP database [55] for non-stationary noise environment. The main goal of this simulation study is to show that the proposed method (AT-IKF) performs the best. The spectrograms of the clean, noisy and enhanced speech by using the above four methods in the presence of white noise and non-stationary noises at 5dB input SNR are shown below respectively.

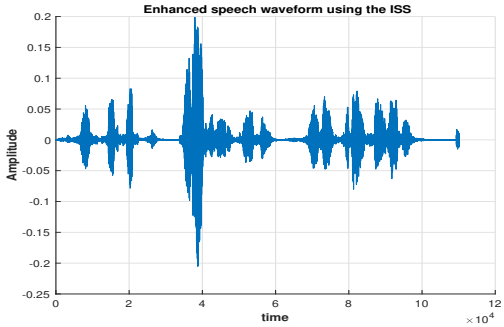
Figs. 4.7a – Fig. 4.7f show the time domain waveform performances of clean speech, noisy speech and enhanced speech by the ISS, the IWF, the SIKF and the AT-IKF methods under white noise environment with 5dB input SNR. Clearly, Fig. 4.7c and Fig. 4.7d appear to have more speech distortion, than Fig. 4.7e based on the amplitude of the enhanced speech. This reveals that the KF based methods yield better speech quality than the ISS and IWF methods. Between Fig. 4.7e and Fig. 4.7f, it is noted that, in lower amplitudes parts, the enhanced speech by using the SIKF method contains more excessive noise than the enhanced speech with the AT-IKF method. In general, the AT-IKF method outperforms the other existing methods in keeping the high amplitude speech parts and reducing the unvoiced parts, which in turn keeps the integrity of the speech.



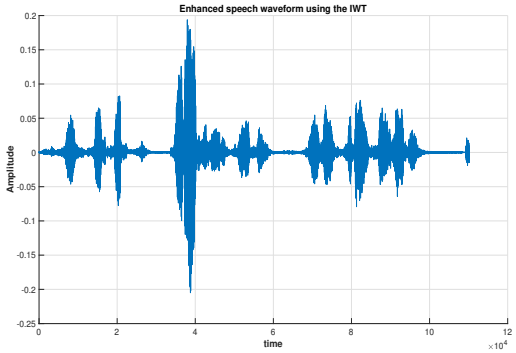
(a) Clean speech waveform



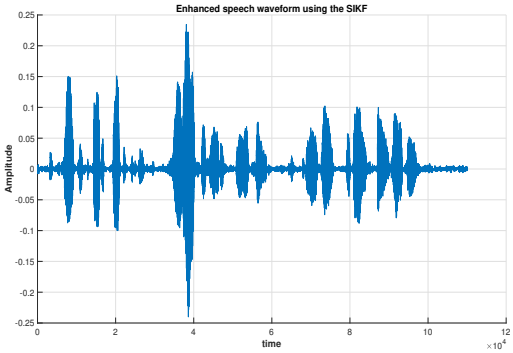
(b) Noisy speech waveform



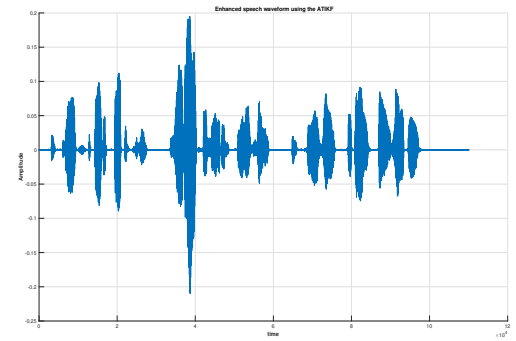
(c) Enhanced speech waveform using the ISS



(d) Enhanced speech waveform using the IWF



(e) Enhanced speech waveform using the SIKF

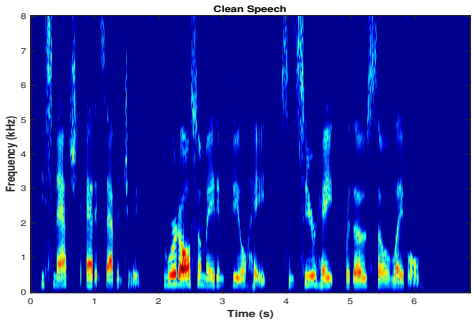


(f) Enhanced speech waveform using the AT-IKF

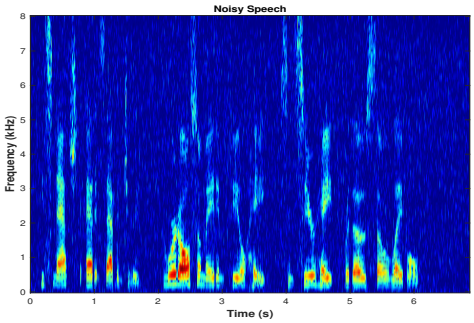
Figure 4.7: Speech waveforms under white noise

Fig. 4.8 shows the spectral performances of clean speech, noisy speech and enhanced speech by the ISS, the IWF, the SIKF and the AT-IKF methods under white noise environment with 5dB input SNR. Among all spectral figures, Fig. 4.8c and Fig. 4.8d show shat the ISS and the IWF algorithms have removed higher frequency speech components as they removed the high frequency noise, which leads to speech distortion and degraded speech quality. From Fig. 4.8e and 4.8f,

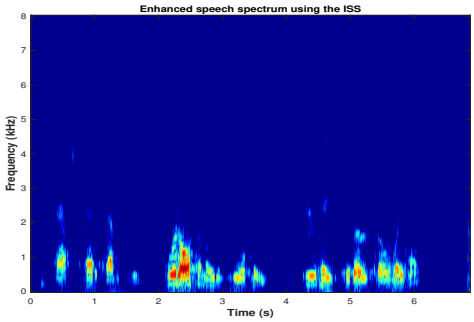
it is seen that the KF based methods perform better than the ISS and the IWF in terms of high frequency noise reduction. Between the SIKF and the AT-IKF algorithms, it is clearly observed that the AT-IKF keeps more speech portion than the SIKF algorithm and removes high frequency noise efficiently.



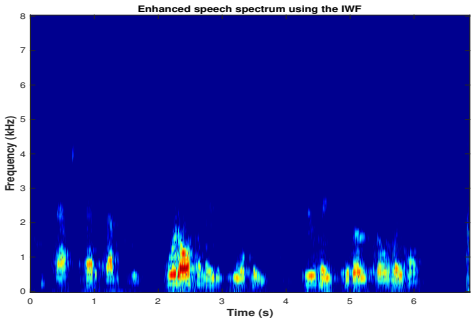
(a) Clean speech spectrum



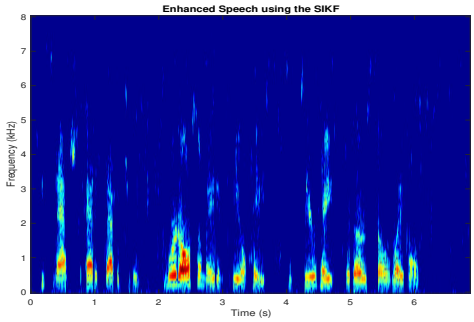
(b) Noisy speech spectrum



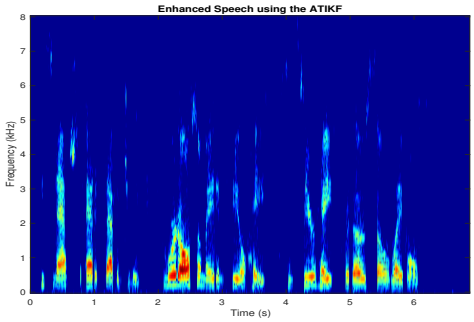
(c) Enhanced speech spectrum using the ISS



(d) Enhanced speech spectrum using the IWF



(e) Enhanced speech spectrum using the SIKF

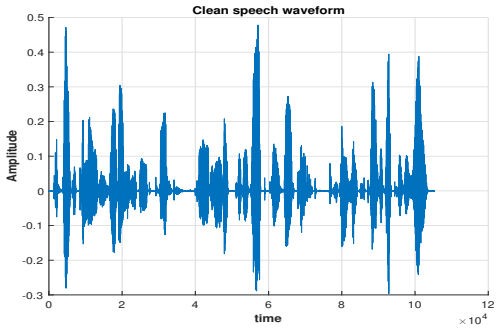


(f) Enhanced speech spectrum using the AT-IKF

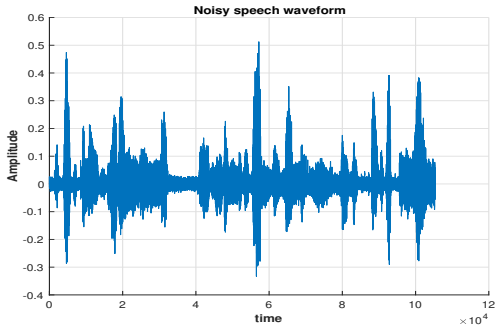
Figure 4.8: Speech spectrums under white noise

Fig. 4.9 gives the time domain waveform performances of clean speech, noisy speech and enhanced speech by the ISS, IWF, SIKF and AT-IKF methods under non-stationary noise environment with 5dB input SNR. It is clearly observed from Fig. 4.9e and Fig. 4.9f that the KF based methods

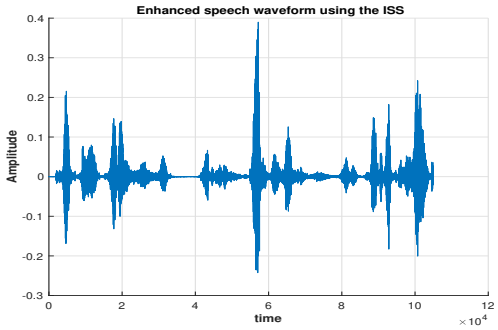
outperform the ISS and the IWF algorithms in terms of speech distortion. Compared to the AT-IKF method, the enhanced speech by the SIKF does not remove noise completely in unvoiced speech frames. Thus, the AT-IKF method outperforms the other existing methods in keeping the speech quality and reducing the noise as well.



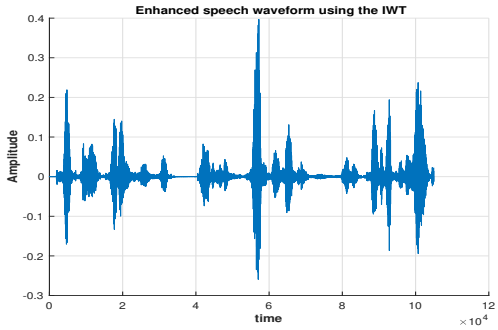
(a) Clean speech waveform



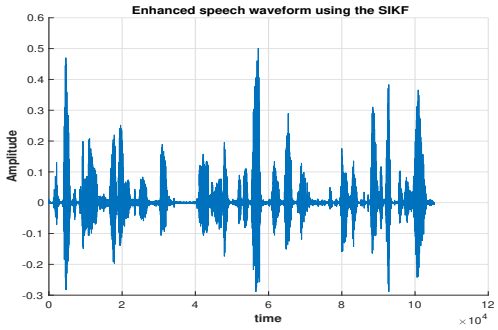
(b) Noisy speech waveform



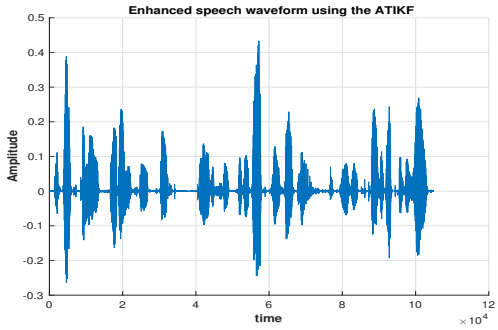
(c) Enhanced speech waveform using the ISS



(d) Enhanced speech waveform using the IWF



(e) Enhanced speech waveform using the SIKF

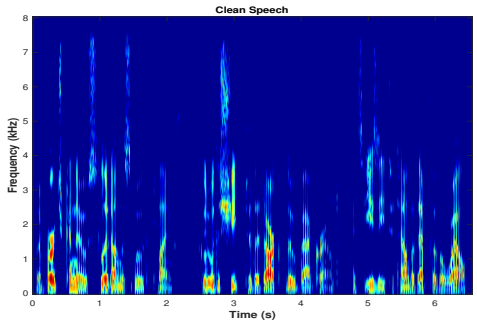


(f) Enhanced speech waveform using the AT-IKF

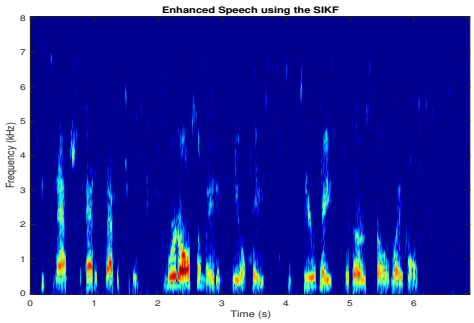
Figure 4.9: Speech waveforms under non-stationary noise

Fig. 4.10 indicates the spectrum performances of clean speech, noisy speech and enhanced speech by the ISS, IWF, SIKF and AT-IKF methods under non-stationary noise environment with 5dB input SNR. Again, among all the competitive algorithms, the AT-IKF method preserves good

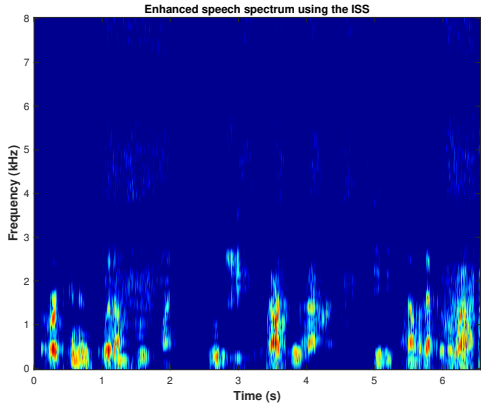
speech quality, provides the best resolution in the speech spectral peaks and gives the lowest residual noise floor in the enhanced speech.



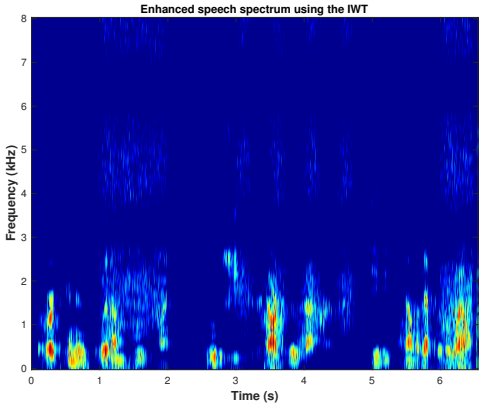
(a) Clean speech spectrum



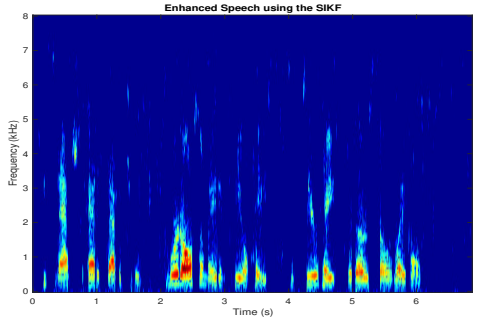
(b) Enhanced speech spectrum using the SIKF



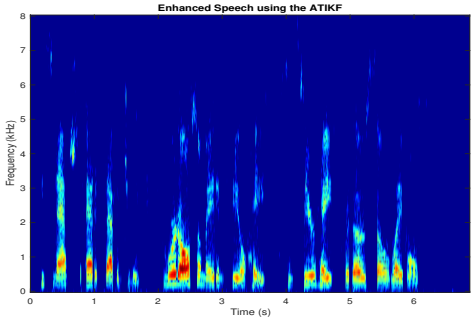
(c) Enhanced speech spectrum using the ISS



(d) Enhanced speech spectrum using the IWF



(e) Enhanced speech spectrum using the SIKF



(f) Enhanced speech spectrum using the AT-IKF

Figure 4.10: Speech spectrums under non-stationary noise

4.3 Conclusion

In this chapter, extensive computer simulations are conducted to evaluate the performances for the proposed method with comparison to other existing popular methods under three different kinds of noise environments for a wide range of input SNRs. Segmental SNR and PESQ are chosen as two metrics to objectively evaluate the quality of the enhanced speech by using different kinds of speech enhancement algorithms. The time domain waveforms and spectrograms have also been researched to show the different characteristics of the enhanced speeches processed by all comparative methods under different noise conditions.

The experimental results show that the proposed method (AT-IKF) outperforms other existing methods in terms of all the evaluation metrics under different kinds of noise environments. It is also shown that overall the KF based methods are better than the ISS and IWF methods, especially in non-stationary noise environment. Even though the SIKF method performs slightly better than the proposed method in white and non-stationary environments in terms of Segmental SNR improvement, the proposed method achieves a much better PESQ score compared with the SIKF method, which is also clearly indicated in the time domain waveforms and spectrograms. In babble noise environment, the proposed method performs the best in terms of all the evaluation metrics. Generally speaking, the proposed adaptive wavelet packet thresholding with iterative Kalman filter method gives less speech distortions and better perceived speech quality, which is the best speech enhancement algorithm for all the noise cases in terms of all the evaluation metrics.

Chapter 5

Conclusion

5.1 Summary of the Work

In this thesis, speech enhancement techniques using wavelet thresholding and Kalman filtering have been throughly studied. The objective was to design an effective method to process a noisy speech in adverse environments, without considering any a priori knowledge of the clean speech and noise information, to reduce the noise while keeping certain level of speech fidelity. To this end, a new adaptive wavelet packet thresholding method with iterative Kalman filter has been proposed for speech enhancement.

Firstly, based on different wavelet thresholding based methods, we proposed a WP thresholding scheme with time-adaptation as a pre-enhancement algorithm. As the wavelet transform is a powerful time-frequency tool which offers a high frequency resolution, it is first applied to the noise corrupted speech on a frame-by-frame basis, decomposing each frame into a number of subbands. This analysis offers a richer range of possibilities and maintains the nature of speech samples in wavelet domain. The VAD is then applied to each subband frame to determine whether the frame is a voice or noise frame. In contrast to most existing works where only the frame energy is employed for voice/noise frame detection, which is not sufficient for input speech with lower SNRs, we have taken frequency features into account. Our VAD method makes use of two measurements namely, i) frame energy and ii) spectral flatness. Following the VAD step, the noise and voice frames are detected. Based on each subband frame activity, the estimated noise variance and the frame-dependent

threshold value are updated. Segmental SNR for each subband frame is also updated along with the corresponding noise variance update. A new adaptive thresholding scheme is then designed and applied to each speech subband for enhancement in accordance with different voice activity. After performing the thresholding operation, the pre-enhanced speech is obtained via an inverse WP transform.

Secondly, to achieve a further level of enhancement, an IKF is next applied to the pre-enhanced speech for further noise reduction. The IKF also operates on a frame-by-frame basis and undergoes two loops of iterations, called inner and outer loop iterations, for each frame. The inner loop iteration includes a prediction step and a measurement update step. In the prediction step, the IKF predicts the state vector and parameter covariance by using the previous samples of the sate-space model. In the measurement update step, the Kalman gain and state vectors are updated. The outer loop, requires only several iterations within each frame. At the beginning of each outer loop iteration, the LPCs and other state-space model parameters are re-estimated from the same processed speech frame that is obtained from the inner loop iteration after all the samples of the frame are processed. The iterative procedure stops when the KF converges or the pre-set maximum number of iterations is exhausted. The IKF gives further enhanced result of the same speech frame with respect to the input pre-enhanced speech frame. At the end of processing all the speech frames, the ultimate enhanced speech is obtained.

The proposed adaptive wavelet thresholding scheme with IKF is evaluated under various noise conditions. It is confirmed through a large number of experiments that the proposed method gives much better noise reduction results than some known methods in the literature in terms of segmental SNR as well as perceptual PESQ. The time domain waveforms and spectrograms of the enhanced speech have also been investigated, showing that the proposed method leads to less speech distortion and better perceived speech quality, which is the best speech enhancement algorithm for all the noise cases in terms of all the evaluation metrics.

5.2 Future Work

Although the wavelet thresholding schemes and the Kalman filter method have been throughly studied in this thesis, there are still rooms for further improvement in the near future.

- The proposed adaptive wavelet thresholding and IKF method only focuses on single channel speech enhancement. We can develop multi-channel versions of the proposed speech enhancement method.
- In general, noise always appears in high frequency region. Following the WPT analysis, each subband has its own frequency range. It is not appropriate to deal with every subband in the same manner for speech enhancement. In the future we can pay more attention to noise characteristics to process each subband accordingly based on its frequency distribution.
- We have only discussed noise reduction in the proposed method. The room dereverberation and acoustic echo cancellation, as important environmental disturbances, also need to be considered for further speech enhancement.

Bibliography

[1] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[2] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[3] J. Benesty and Y. Huang, *Adaptive signal processing: applications to real-world problems*. Springer Science & Business Media, 2013.

[4] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay, *et al.*, *Advances in network and acoustic echo cancellation*. Springer, 2001.

[5] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[6] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[7] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 250–255, 1993.

[8] J. D. Warren, B. A. Zielinski, G. G. Green, J. P. Rauschecker, and T. D. Griffiths, "Perception of sound-source motion by the human brain," *Neuron*, vol. 34, no. 1, pp. 139–148, 2002.

[9] U. Mortensen, "Additive noise, weibull functions and the approximation of psychometric functions," *Vision Research*, vol. 42, no. 20, pp. 2371–2393, 2002.

[10] R. C. Hendriks, T. Gerkmann, and J. Jensen, "Dft-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, 2013.

[11] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[12] Y. Malca and D. Wulich, "Improved spectral subtraction for speech enhancement," in *European Signal Processing Conference, EUSIPCO 1996. 8th*, pp. 1–5, IEEE, 1996.

[13] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on speech and audio processing*, vol. 7, no. 2, pp. 126–137, 1999.

[14] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise.," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 44164–44164, 2002.

[15] L.-y. Sui, X.-w. Zhang, J.-j. Huang, and B. Zhou, "An improved spectral subtraction speech enhancement algorithm under non-stationary noise," in *Proc. of IEEE Int. Conf. on Wireless Communications and Signal Processing (WCSP)*, pp. 1–5, IEEE, 2011.

[16] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[17] B. Fan, H. Song, M. Liu, and Y. Wang, "The improvement and realization of speech enhancement algorithm based on wiener filtering," in *Proc. of IEEE Int. Conf. on Image and Signal Processing (CISP)*, pp. 1116–1120, IEEE, 2015.

[18] L. Lin, W. H. Holmes, and E. Ambikairajah, "Subband noise estimation for speech enhancement using a perceptual wiener filter," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. I–80, IEEE, 2003.

[19] L. Lin, W. H. Holmes, and E. Ambikairajah, "Speech denoising using perceptual modification of wiener filtering," *Electronics Letters*, vol. 38, no. 23, pp. 1486–1487, 2002.

[20] K. Funaki, "Speech enhancement based on iterative wiener filter using complex speech analysis," in *Signal Processing Conference, 2008 16th European*, pp. 1–5, IEEE, 2008.

[21] R. Mao, Y. Zhou, W. Yuan, and H. Liu, "An improved iterative wiener filtering algorithm for speech enhancement," in *Proc. of IEEE Int. Conf. on Computer and Communications (ICCC)*, pp. 436–440, IEEE, 2015.

[22] J. W. Seok and K. S. Bae, "Speech enhancement with reduction of noise components in the wavelet domain," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 1323–1326, IEEE, 1997.

[23] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE Signal Processing Letters*, vol. 8, no. 1, pp. 10–12, 2001.

[24] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 421–424, IEEE, 1991.

[25] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, vol. 21, no. 3, pp. 191–207, 1997.

[26] I. Cohen, "Enhancement of speech using bark-scaled wavelet packet decomposition," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[27] L. Singh and S. Sridharan, "Speech enhancement using critical band spectral subtraction," in *Fifth International Conference on Spoken Language Processing*, 1998.

[28] S. Chang, Y.-h. Kwon, S.-i. Yang, and I.-j. Kim, "Speech enhancement for non-stationary noise environment by adaptive wavelet packet," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. I–561, IEEE, 2002.

[29] J. K. Lee and C. D. Yoo, "Wavelet speech enhancement based on voiced/unvoiced decision," *WESTPAC V94 Technical Papers*, pp. 4149–4156, 2003.

[30] S.-H. Chen and J.-F. Wang, "Speech enhancement using perceptual wavelet packet decomposition and teager energy operator," in *Real World Speech Processing*, pp. 51–65, Springer, 2004.

[31] M. T. Johnson, X. Yuan, and Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding," *Speech Communication*, vol. 49, no. 2, pp. 123–133, 2007.

[32] L. Drolet, F. Michaud, and J. Côté, "Adaptable sensor fusion using multiple kalman filters," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2000.(IROS 2000)*, vol. 2, pp. 1434–1439, IEEE, 2000.

[33] R. E. Kalman *et al.*, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[34] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 12, pp. 177–180, IEEE, 1987.

[35] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on signal processing*, vol. 39, no. 8, pp. 1732–1742, 1991.

[36] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Transactions on speech and audio processing*, vol. 6, no. 4, pp. 373–385, 1998.

[37] N. B. Yoma, F. R. McInnes, and M. A. Jack, "Improving performance of spectral subtraction in speech recognition using a model for additive noise," *IEEE Transactions on speech and audio processing*, vol. 6, no. 6, pp. 579–582, 1998.

[38] M. Gabrea, E. Grivel, and M. Najun, "A single microphone kalman filter-based noise canceller," *IEEE Signal Processing Letters*, vol. 6, no. 3, pp. 55–57, 1999.

[39] N. Ma, M. Bouchard, and R. A. Goubran, "Perceptual kalman filtering for speech enhancement in colored noise," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. I–717, IEEE, 2004.

[40] M. Saha, R. Ghosh, and B. Goswami, "Robustness and sensitivity metrics for tuning the extended kalman filter," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 4, pp. 964–971, 2014.

[41] R. Ishaq, B. G. Zapirain, M. Shahid, and B. Lövström, "Subband modulator kalman filtering for single channel speech enhancement," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7442–7446, IEEE, 2013.

[42] S. K. Roy, W.-P. Zhu, and B. Champagne, "Single channel speech enhancement using subband iterative kalman filter," in *Proc. of IEEE Int. Symposium on Circuits and Systems (ISCAS)*, pp. 762–765, IEEE, 2016.

[43] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE transactions on information theory*, vol. 38, no. 2, pp. 617–643, 1992.

[44] J. Yao and Y.-T. Zhang, "Bionic wavelet transform: a new time-frequency method based on an auditory model," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 8, pp. 856–863, 2001.

[45] M. J. Shensa, "The discrete wavelet transform: wedding the a trous and mallat algorithms," *IEEE Transactions on signal processing*, vol. 40, no. 10, pp. 2464–2482, 1992.

[46] B. Walczak and D. Massart, "Noise suppression and signal compression using the wavelet packet transform," *Chemometrics and Intelligent Laboratory Systems*, vol. 36, no. 2, pp. 81–94, 1997.

[47] P. Goupillaud, A. Grossmann, and J. Morlet, "Cycle-octave and related transforms in seismic signal analysis," *Geoexploration*, vol. 23, no. 1, pp. 85–102, 1984.

[48] R. Kronland-Martinet, J. Morlet, and A. Grossmann, "Analysis of sound patterns through wavelet transforms," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 1, no. 02, pp. 273–302, 1987.

[49] J.-M. Combes, A. Grossmann, and P. Tchamitchian, *Wavelets: Time-Frequency Methods and*

Phase Space Proceedings of the International Conference, Marseille, France, December 14–18, 1987. Springer Science & Business Media, 2012.

[50] J. Yao and Y.-T. Zhang, "The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 11, pp. 1299–1309, 2002.

[51] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *biometrika*, pp. 425–455, 1994.

[52] D. L. Donoho, "Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition," *Applied and computational harmonic analysis*, vol. 2, no. 2, pp. 101–126, 1995.

[53] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, pp. 1135–1151, 1981.

[54] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on time–scale adaptation," *Speech Communication*, vol. 48, no. 12, pp. 1620–1637, 2006.

[55] P. Kabal, "Tsp speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[56] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[57] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 749–752, IEEE, 2001.

[58] K. Kondo, "Speech quality," in *Subjective Quality Measurement of Speech*, pp. 7–20, Springer, 2012.

[59] M.-J. Zhao and W.-P. Zhu, "Adaptive wavelet packet thresholding thresholding with iterative kalman filter for speech enhancement," *accepted for presentation in IEEE Global Conference on Signal and Information Processing (GlobalSIP) to be held in Montreal, Canada during November*, 2017.

[60] M. Vetterli and C. Herley, "Wavelets and filter banks: Theory and design," *IEEE transactions on signal processing*, vol. 40, no. 9, pp. 2207–2232, 1992.

[61] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 87, 2013.

[62] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," in *Proc. 17th European Signal Processing Conference*, pp. 2549–2553, IEEE, 2009.

[63] S. M. Kay, *Modern spectral estimation*. Pearson Education India, 1988.

[64] V. Sunnydayal and T. K. Kumar, "Speech enhancement using sub-band wiener filter with pitch synchronous analysis," in *Proc. of IEEE Int. Conf. on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 20–25, IEEE, 2013.

[65] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.