SPEECH ENHANCEMENT USING SPARSE REPRESENTATION METHODS

SHEN TAK WAI

Ph.D

The Hong Kong Polytechnic University

2015

The Hong Kong Polytechnic University

Department of Electronic and Information Engineering

Speech Enhancement Using Sparse Representation Methods

SHEN Tak Wai

A thesis submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

January 2015

CERTIFICATE OF ORIGINALITY

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

_____ (Signed)

_____Shen Tak Wai_____ (Name of student)

Dedication

This thesis is dedicated to the memory of my beloved father,

Ming-On Shen.

My mother, Fung-Kau Au, my loving wife, Mandy Yeung and Chun-Lok Shen, my son.

Abstract

In this thesis, the problem of speech enhancement is investigated. In consistent with the traditional frequency domain speech enhancement algorithms, we investigated the estimation methods of some important parameters in speech enhancement, such as the speech periodogram, *a-priori* Signal-to-Noise Ratio (SNR) and Speech Presence Probability (SPP). In this study, we emphasize on making use of the sparse representation of speech signals to improve the estimation. To achieve this, the wavelet denoising technique, the cepstral analysis using expectation-maximization (EM) framework as well as the dictionary learning method based on sparse reconstruction on log-spectra have been adopted and achieved satisfactory results.

The first part of this study is related to the estimation of SPP. It is known that a reliable SPP estimator is important to many frequency domain speech enhancement algorithms. A good estimate of SPP can be obtained by having a smooth *a-posteriori* SNR function, which can be achieved by reducing the noise variance when estimating the speech power spectrum. Recently, the wavelet denoising with multitaper spectrum (MTS) estimation technique was suggested for such purpose. However, traditional approaches directly make use of the wavelet shrinkage denoiser which has not been fully optimized for denoising the MTS of noisy speech signals. In this study, we propose a two-stage wavelet denoising algorithm for estimating the speech power spectrum. First, we apply the wavelet transform to the periodogram of a noisy speech signal. Using the resulting wavelet coefficients, an oracle is developed to indicate the approximate locations of the noise floor in the periodogram. Second, we make use of the oracle developed in stage 1 to selectively remove the wavelet coefficients of the noise floor in the log MTS of the noisy speech. The remaining wavelet coefficients are then used to reconstruct a denoised MTS and in turn generate a smooth *a-posteriori* SNR

function. To adapt to the enhanced *a-posteriori* SNR function, we further propose a new method to estimate the generalized likelihood ratio (GLR), which is an essential parameter for SPP estimation. Simulation results show that the new SPP estimator outperforms the traditional approaches and enables an improvement in both the quality and intelligibility of the enhanced speeches.

While the wavelet transform can sparsely describe the sudden changes in a speech power spectrum, it misses the periodic nature of speech signals which is an important feature in speech enhancement. For the second part of this study, a new speech enhancement method based on the sparsity of speeches in the cepstral domain is investigated. It is known that voiced speeches have a quasi-periodic nature that allows them to be compactly represented in the cepstral domain. It is a distinctive feature compared with noises. Recently, the temporal cepstrum smoothing (TCS) algorithm was proposed and was shown to be effective for speech enhancement in non-stationary noise environments. However, the missing of an automatic parameter updating mechanism limits its adaptability to noisy speeches with abrupt changes in SNR across time frames or frequency components. In this part, an improved speech enhancement algorithm based on a novel EM framework is proposed. The new algorithm starts with the traditional TCS method which gives the initial guess of the periodogram of the clean speech. It is then applied to an L_1 norm regularizer in the M-step of the EM framework to estimate the true power spectrum of the original speech. It in turn enables the estimation of the *a-priori* SNR and is used in the E-step, which is indeed an MMSE-LSA gain function, to refine the estimation of the clean speech periodogram. The M-step and E-step iterate alternately until converged. A notable improvement of the proposed algorithm over the traditional TCS method is its adaptability to the changes (even abrupt changes) in SNR of the noisy speech. Performance of the proposed algorithm is evaluated using standard measures based on a large set of speech and noise signals. Evaluation results show that a significant

improvement is achieved compared to conventional approaches.

The above shows that obtaining the sparse representation of speeches is one of the keys for designing an efficient speech enhancement algorithm. One obvious question then arises if the ceptrum is the best representation of speeches as far as the sparsity is concerned. To answer this question, we further investigate a new sparse representation based speech enhancement algorithm with the transform kernel trained based on the dictionary learning method. It is known that the dictionary learning method allows the design of a transform kernel with the emphasis of sparsity in the transform domain. When applying to speech enhancement, it allows a speech to be represented by very few significant transform coefficients. In practice, the overcomplete dictionary of the clean speech signal is trained by an extended K-SVD algorithm in the log power spectra domain. The batch LARS with Coherence Criterion (LARC) method is used to reconstruct the log power spectra of the clean speech. And a new stopping criterion is proposed for the iterative speech enhancement process in order to adapt to various background noise environment. In addition, a modified two-step noise reduction with MMSE-LSA filtering is applied which solves the bias problem of the estimated a *priori* SNR. A notable improvement of the proposed algorithm over the traditional speech enhancement method is its adaptability to the changes in SNR of the noisy speech. Performance of the proposed algorithm is evaluated using standard measures based on a large set of speech and noise signals. Evaluation results show that a significant improvement is achieved compared to the traditional approaches especially when the noises are not totally random but have certain structure in the frequency domain.

List of Publications

International Journal Papers

- D.P.K. Lun, T.W. Shen, T.C. Hsung, D.K.C. Ho," Wavelet based speech presence probability estimator for speech enhancement", *Digital Signal Processing*, Vol.22, Issue 6, pp. 116, 2012.
- Daniel P.K. Lun, T.W. Shen and K.C. Ho, "A novel expectation-maximization framework for speech enhancement in non-stationary noise environments", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 22, Issue 2, pp.335-346, Feb 2014.

International Conference / Symposium Papers

- T.W. Shen, D.P.K. Lun and T.C. Hsung, "Speech Enhancement Using Harmonic Regeneration With Improved Wavelet Based A-Priori Signal To Noise Ratio Estimator" *Proceedings, 2010 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS 2010),* Cheng Du, China, Dec 2010, pp. 225-228.
- D.P.K. Lun, T.W. Shen, T.C. Hsung, D.K.C. Ho," Improved speech presence probability estimation based on wavelet denoising", *Proceedings, IEEE International Symposium on Circuits and Systems (ISCAS'2012),* Seoul, Korea, May 2012, pp.1018-1021.
- T.W. Shen and D.P.K. Lun," Speech Enhancement Based on L_1 Regularization in the Cepstral Domain", *Proceedings, IEEE International Symposium on Circuits and Systems (ISCAS'2014)*, Melbourne, Australia, June 2014, pp. 121-124.

Recent Submissions

- T.W. Shen and D.P.K. Lun, "A Speech Enhancement Method Based on Sparse Reconstruction on Log-Spectra", under preparation.

Acknowledgements

I would like to express my profound gratitude to my chief supervisor, Dr. Daniel P. K. Lun for his continual support, encouragement, supervision and valuable suggestions throughout this research work. Dr. Lun provided the perfect balanced structure and freedom which allowed me to derive the most from my research experience.

I am greatly indebted to Prof. K. C. Ho, Dr. T. C. Hsung, Dr. Y.H. Chan, Dr. M.W. Mak and Prof. W.C. Siu. Their expert knowledge and friendly encouragement have helped me resolved many difficult problems during my study.

This work would not have been possible without the support and assistance of many individuals. I would like to take this opportunity to thank all members of the Centre for Signal Processing in The Hong Kong Polytechnic University. My interactions and discussions with the group members have helped to formulate and develop this research work. It has been a wonderful time to me in these years working with them.

Meanwhile, I am glad to express my gratitude to the Department of Electronic and Information Engineering (EIE) and The Hong Kong Polytechnic University (PolyU) for providing me a comfortable working environment and for their financial support to my research work. The work described in this research studies is partially supported by the Centre for Signal Processing, Department of EIE, PolyU and a grant from the Research Grants Council of the HKSAR, China. I acknowledge the research studentships provided by the parties stated above.

I am, as ever, indebted to my family for their love, encouragement and support throughout my life. Without their understanding and patience, it is impossible for me to complete this research study.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AR	Autoregressive
LARSC	Batch LARS with Coherence Criterion method
CD	Cepstral distance measures
CS	Compressed sensing
DFT	Discrete Fourier Transformation
DWT	Discrete Wavelet Transform
EbayesThresh	Empirical Bayes Thresholding
EM	Expectation-maximization
FFT	Fast Fourier transform
FOCUSS	Focal underdetermined system solver
GGD	Generalized Gaussian distribution
GLR	Generalized likelihood ratio
GP	Gradient pursuits
GAD	Greedy adaptive dictionary learning algorithm
HRNR	Harmonic Regeneration Noise Reduction Algorithm
HMMs	Hidden Markov models
ICA	Independent component analysis
IDFT	Inverse discrete Fourier transform
IDWT	Inverse DWT
IS	Itakura-Saito measures
K-SVD	K-singular value decomposition
LASSO	Least absolute shrinkage and selection operator
LARS	Least angle regression
LPC	Linear predictive coding
LSA	Log spectral amplitude
LLR	Log-likelihood ratio
MP	Matching pursuit
MAP	Maximum *a-posterior*
ML	Maximum-likelihood

MOS	Mean Opinion Score
MSE	Mean square error
MOD	Method of Optimal Directions
MMSE	Minimum Mean Square Error estimator
MTS	Multitaper spectrum
NMF	Non-negative matrix factorization
OMP	Orthogonal MP
OSA	Obstructive sleep apnea
PDF	Probability density function
PESQ	Perceptual Evaluation of Speech Quality
PSD	Power spectral density
PCM	Pulse-code modulation
segSNR	Segmental signal-to-noise ratio
STSA	Short-term spectral amplitude
STFT	Short-Time Fourier transforms
SNR	Signal-to-Noise Ratio
Sthresh	Simple Thresholding
SVD	Singular value decomposition
SPP	Speech Presence Probability
SURE	Stein's unbiased risk estimate
TCS	Temporal cepstrum smoothing
TSNR	Two-step noise reduction
VAD	Voice activity detection
WSS	Weighted-slope spectral distance

Chapter 1 Introduction

Speech enhancement is a challenging problem due to the diversity of noise sources and their effects in different applications [1]. Over the last three decades a substantial increase is noted in the use of speech-processing devices in cellular phones, digital hearing aids, and various human-to-machine speech-processing applications. Originally most of these applications assumed the acquired speech signals were noise-free. It was soon proven to be not the case. There is thus a great demand to make these applications to work robustly under noisy conditions as well. It becomes an extremely challenging task for the speech-processing devices, particularly when considering the large variety of noisy environments. Consequently, speech enhancement methods were developed to improve the robustness of these speech-processing devices. The term speech enhancement in fact refers to a large group of methods for improving the quality of speech signals. Some examples of speech enhancement methods include noise reduction, bandwidth extension, acoustic echo control (dereverberation), packet loss concealment, etc. In this study, we focus on the speech enhancement methods for noise reduction.

Speech is a highly non-stationary signal with specific properties [1][2][3]. However, over a sufficiently short period of time (10 - 32 ms), its spectral characteristics are fairly stationary. This allows speech processing algorithms to operate on a frame-by-frame basis with duration of each frame ranging from 10 to 32ms. Speech bandwidth varies approximately from 50Hz to 8kHz. Speeches can be classified into two categories, i.e. voiced speeches and unvoiced speeches. In many practical working environments, the background noise can be considered as additive and statistically independent with the speech signal. For other noises such as the convolutional noise, multiplicative noise or signal-dependant quantization noise in pulse-code

modulation (PCM), they can also be modeled as an additive process after some conversion [4].

Speech enhancement algorithms could be classified as single-microphone or multi-microphone based. Single-microphone speech enhancement algorithms estimate the clean speech signal using a realization of the noisy speech that is obtained using one microphone only. Multi-microphone speech enhancement algorithms, on the other hand, use more than one microphone and can as such also exploit spatial properties, and as a result, their performance is in general better than single-microphone speech enhancement algorithms. However, multi-microphone techniques often require higher cost and impose more constraints on the system (e.g. distance between the microphones). In this thesis, we focus on single-microphone speech enhancement algorithms. This is the most difficult situation, because the speech and noise are in the same channel. In addition, single-microphone methods can be extended and used in a multi-microphone system or combined with multi-microphone algorithms as a post-processor to get a better noise reduction performance.

Fig. 1.1 illustrates a typical single-microphone speech enhancement process. The clean speech signal is denoted as x while the additive noise is denoted as n. The speech enhancement algorithm attempts to suppress noise without distorting speech and produces the enhanced speech signal $\hat{x}$. Speech enhancement algorithms try to reduce the impact of background noise on the speech signal. They improve the quality of the speech to be used in recording systems, telephone systems, hearing aids devices and the communications between human beings. Besides they can improve the accuracy of the machine automatic speaker verification or speech recognition processes.

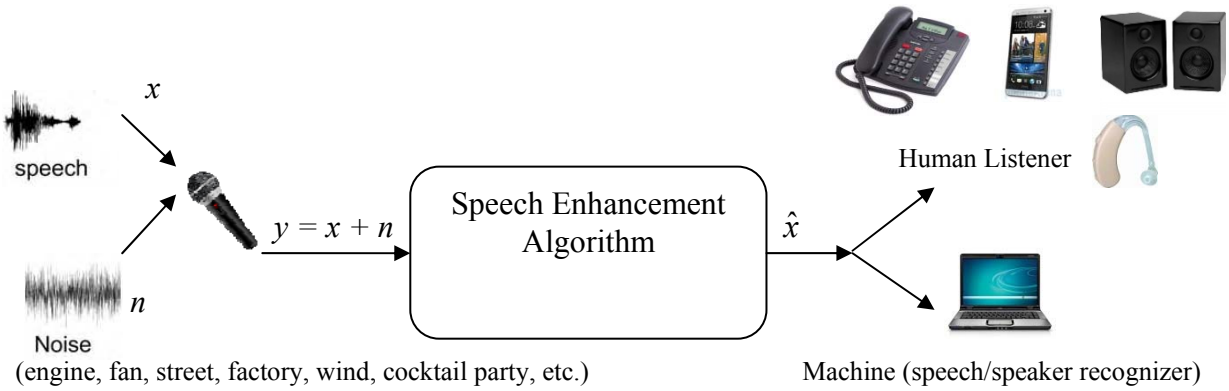


Fig. 1.1 – Typical single-microphone speech enhancement process

Single-microphone speech enhancement algorithms can be generally classified as parametric and nonparametric. Parametric techniques model the speech signal as a stochastic autoregressive (AR) process embedded in Gaussian noises. The speech enhancement algorithm then consists of estimating the speech AR parameters and applying a filter (e.g. Wiener or Kalman filter) to the noisy signal, where the optimal filters are designed based on the estimated AR parameters [5][6]. Non-parametric techniques do not estimate the speech parameters and require a noise fingerprint in the transform domain. It will be used during speech-and–noise periods to obtain an estimate of the clean speech signal. Well-known non-parametric methods include the signal subspace decomposition methods and methods based on processing in the Discrete Fourier Transform (DFT) domain (frequency domain). In this study, we focus mainly on the non-parametric speech enhancement methods.

The subspace algorithms are designed based on the principle that clean speech signals are often confined to a subspace of the noisy Enclidean space. As a result, given a method of decomposing the vector space of a noisy signal into a subspace that is occupied mainly by the clean signal and a subspace occupied mainly by the noise signal, the clean signal simply could be estimated by nulling the component of the noisy vector residing in the "noise subspace". Dendrinos et al. [5] initially proposed the use of the singular value decomposition (SVD) on a data matrix containing time-domain amplitude values. And then, Ephraim and

Van Trees [8] proposed using the eigenvalue decomposition method on the signal covariance matrix. Extensions of the subspace based enhancement methods have been proposed to work for colored noise [9] and take perceptual aspects into account [10].

Frequency domain techniques require the application of the DFT to the noisy speech before filtering, followed by the inverse discrete Fourier transform (IDFT). By transforming the noisy speech to the frequency domain, noises can be better distinguished from speeches and removed. For example, the energy of voiced speech is concentrated at certain frequencies, but the energy of white noise is uniformly spread throughout the entire frequency spectrum. The famous spectral subtraction algorithm in [10] was extended to the Fourier domain by Boll [12] and became a very popular method. The major advantage of the spectral subtraction algorithms is their simplicity. They are based on the principle that, as the noise contamination process is additive, the noise spectrum can be estimated/updated when speech is not present and subtracted from the noisy speech. Although the spectral subtraction methods are popularly used, it is well known that they can generate musical noise which can be rather annoying to human listeners. Another important class of Fourier domain speech enhancement algorithms was initiated by MacAulay and Malpass [13]. They proposed a maximum-likelihood (ML) approach for estimating the Fourier transform coefficients (spectrum) of the clean speech. Their works were followed by Ephraim and Malah [14] who proposed a Minimum Mean Square Error (MMSE) estimator of the magnitude spectrum. These estimators are a function of the distributional parameters of the variance of the noise and speech DFT coefficients [14]-[17]. While the MMSE estimator was proposed 30 years ago, it still attracts much attention in the field recently [18] since it does not only reduce the noise power but also reduce the level of musical noise as compared to the spectral subtraction methods. Other works that adopt different statistical filters such as the Wiener filters can also be found in [4], [5] and [19]. Also some variants have been proposed which take human

perception into account [20]-[21].

Recently, signal denoising approaches are often equipped with an assumption of sparsity [22], which refers to the fact that many natural signals can be described by very few non-zero coefficients in some transform domains. This also applies to speech signals. Sparse signal representations have been successfully applied to signal denoising and blind source separation [22]-[24]. For a noisy speech, the speech component can be represented by very few transform coefficients but with large magnitude. It is not the case for the noise component which has its coefficients spread over the transform domain. Hence the speech coefficients can be identified from the noise coeffcients. By removing the noise coefficients, the unwanted noise signal can be suppressed. In this thesis, our speech enhancement approach first transforms the noisy speech signal into the frequency domain, and then the sparse coding technique is applied using the signal models of speech (called dictionary). Sparse coding is able to separate a noisy speech into its structured components and to suppress any unstructured components (i.e. noise) that are incoherent to its dictionary. Finally, an enhanced speech is obtained by performing the IDFT back to the time domain. For the rest of this chapter, we state the research objective and summarize the layout of the thesis.

1.1 Present Works

In this thesis, we mainly focus on applying the sparse representation techniques to the class of frequency domain single-channel speech enhancement algorithms. Although a lot of research has been done in the field of speech enhancement, the field of single-channel speech enhancement is still very challenging and many problems remain to be solved. There are many scenarios, e.g. under low signal-to-noise ratio (SNR) or under non-stationary noise conditions where existing systems fail to give a satisfactory result. In recent years, there has been substantial development in the field of sparse representation. It has been used successfully in applications such as signal denoising, restoration, and reconstruction, etc. By representing speeches in a sparse manner, speech energy can be concentrated into a few transform coefficients, which will be highly distinctive from that of noise. Besides, many advanced regularization methods were developed in recent years that emphasize on promoting the sparsity of the signal. Some of them can also be applied to improve the existing speech enhancement methods. The major objective of this study is to investigate clean speech estimators based on the sparse representation techniques in the frequency domain. In particular, we find that the wavelet transform, cepstral transform and dictionary learing methods are useful to obtain the sparse representation of speeches and hence facilitate the design of speech enhancement algorithms. More specifically, the present work can be divided into the following three different parts.

1.1.1 Wavelet Based Speech Presence Probability Estimator

As mentioned above, a reliable SPP estimator can significantly improve their performance to many frequency domain speech enhancement algorithms. It is known that a good estimate of SPP can be obtained by having a smooth *a-posteriori* SNR function [49],

which can be achieved by reducing the noise variance when estimating the speech power spectrum. Recently, the wavelet denoising with multitaper spectrum (MTS) estimation technique was suggested for such purpose. However, traditional approaches directly make use of the wavelet shrinkage denoiser which has not been fully optimized for denoising the MTS of noisy speech signals. In this part of study, we firstly propose a two-stage wavelet denoising algorithm for estimating the speech power spectrum. In the first stage, we apply the wavelet transform to the periodogram of a noisy speech signal. Using the resulting wavelet coefficients, an oracle is developed to indicate the approximate locations of the noise floor in the periodogram. In the second stage, we make use of the oracle developed in stage 1 to selectively remove the wavelet coefficients of the noise floor in the log MTS of the noisy speech. The wavelet coefficients that remained are then used to reconstruct a denoised MTS and in turn generate a smooth *a-posteriori* SNR function. To adapt to the enhanced *a-posteriori* SNR function, we further propose a new method to estimate the generalized likelihood ratio (GLR), which is an essential parameter for SPP estimation. Simulation results show that the new SPP estimator outperforms the traditional approaches and enables an improvement in both the quality and intelligibility of the enhanced speeches.

1.1.2 Expectation-Maximization Framework with Cepstral Representation

As mentioned above, many approaches were suggested to reduce the musical noise. However, their performance can get worse significantly when the SNR is low or when the noise is non-stationary. Speech enhancement method based on the sparsity of speeches in the cepstral domain is investigated to overcome this problem. It is known that voiced speeches have a quasi-periodic nature that allows them to be compactly represented in the cepstral domain. It is a distinctive feature compared with noises. Recently, the temporal cepstrum smoothing (TCS) algorithm was proposed and was shown to be effective for speech

enhancement in non-stationary noise environments. However, the missing of an automatic parameter updating mechanism limits its adaptability to noisy speeches with abrupt changes in SNR across time frames or frequency components. In this part, an improved speech enhancement algorithm based on a novel expectation-maximization (EM) framework is proposed. The new algorithm starts with the traditional TCS method which gives the initial guess of the periodogram of the clean speech. It is then applied to an L_1 norm regularizer in the M-step of the EM framework to estimate the true power spectrum of the original speech. It in turn enables the estimation of the *a-priori* SNR and is used in the E-step, which is indeed an MMSE-LSA gain function, to refine the estimation of the clean speech periodogram. The M-step and E-step iterate alternately until converged. A notable improvement of the proposed algorithm over the traditional TCS method is its adaptability to the changes (even abrupt changes) in SNR of the noisy speech. Performance of the proposed algorithm is evaluated using standard measures based on a large set of speech and noise signals. Evaluation results show that a significant improvement is achieved compared to conventional approaches, particularly when the noise is non-stationary.

1.1.3 Sparse Reconstruction of the Log-Spectra by the Dictionary Learning Method

The above shows that obtaining the sparse representation of speeches is one of the keys for designing an efficient speech enhancement algorithm. One obvious question then arises if the ceptrum is the best representation of speeches as far as the sparsity is concerned. To answer this question, we further investigate a new sparse representation based speech enhancement algorithm with the transform kernel trained based on the dictionary learning method. It is known that the dictionary learning method allows the design of a transform kernel with the emphasis of sparsity in the transform domain. When applying to speech enhancement, it allows a speech to be represented by very few significant transform

coefficients. In practice, the overcomplete dictionary of the clean speech signal is trained by an extended K-SVD algorithm in the log power spectra domain. The batch LARS with Coherence Criterion (LARC) method is used to reconstruct the log power spectra of the clean speech. And a new stopping criterion is proposed for the iterative speech enhancement process in order to adapt to various background noise environment. In addition, a modified two-step noise reduction with MMSE-LSA filtering is applied which solves the bias problem of the estimated a *priori* SNR. A notable improvement of the proposed algorithm over the traditional speech enhancement method is its adaptability to the changes in SNR of the noisy speech. Performance of the proposed algorithm is evaluated using standard measures based on a large set of speech and noise signals. Evaluation results show that a significant improvement is achieved compared to the traditional approaches especially when the noises are not totally random but contain certain structure in the frequency domain.

1.2 Organization of the Thesis

This thesis is organized as follows. In Chapter 2, we give an overview of speech enhancement methods based on sparse representation. In Chapter 3, the wavelet denoising technique for smoothing a periodogram is investigated. A new two-stage wavelet denoising algorithm for estimating the speech power spectrum is proposed. In Chapter 4, speech enhancement method based on the sparsity of speeches in the cepstral domain is introduced. An improved speech enhancement algorithm based on a novel expectation-maximization (EM) framework is proposed. In Chapter 5, the overcomplete dictionary learning method is described. The design of a new speech enhancement method using sparse reconstruction of the log-spectra is explained. Finally, the conclusion of the whole study is drawn in Chapter 6 where possible future works are also discussed.

Chapter 2 Overview of Speech Enhancement Methods Based on Sparse Representation

In this thesis, we are concerned with the speech enhancement problem using single-microphone. In particular, we investigate transforming speech signals into their sparse representations, because they allow the most important information within a speech signal to be conveyed with only a few elementary components. In fact, the concept of sparse representation is not new to the problem of speech enhancement. Arguably all frequency domain speech enhancement algorithms can be considered to have adopted such concept. It is because by transforming a speech signal into the frequency domain, the energy of the speech will be concentrated at the low frequency part of the spectrum. The concept of sparse representation is re-vitalized actually due to the recent results on dictionary learning methods and the iterative regularization techniques that give a much better understanding of the ways to generate a sparse transform kernel and to process the signal if it is sparse in the transform domain. In this chapter, we review some of the commonly used single-channel speech enhancement algorithms in the frequency domain. Besides, some methods that make use of the sparse representations of speeches in enhancement applications are reviewed. And the dictionary learning method is also introduced. Finally, we provide an overview of the evaluation methods.

2.1 Speech Enhancement in the DFT Domain

Most of the known works on speech enhancement use the additive noise model to describe background noise, which is justified by the principle of superposition. In the additive noise model, the noisy speech *y* is assumed to be the sum of the clean speech *x* and the additive noise *n* as defined in the following equation:

$$y(t) = x(t) + n(t) \tag{2.1}$$

The DFT converts the time domain information of a signal into its frequency domain information. The forward transform is defined as

$$v(k) = \sum_{n=0}^{N-1} u(n) W_N^{kn}, \qquad k = 0, 1, \ldots, N-1 \tag{2.2}$$

where

$$W_N = \exp\left\{\frac{-j2\pi}{N}\right\} \tag{2.3}$$

The inverse transform is given by

$$u(k) = \frac{1}{N} \sum_{k=0}^{N-1} v(n) W_N^{-kn}, \qquad n = 0, 1, \ldots, N-1 \tag{2.4}$$

The transform coefficients are complex and can be separated into the magnitude and phase components. The DFT of a real sequence shows conjugate symmetry about *N/2* which means only half the data require to be processed.

The short-time Fourier transform (STFT) has been popularly adopted in the development of speech enhancement algorithms. By means of the STFT, the additive relationship of noise and speech in a noisy speech signal can be expressed with a time-frequency setting as follows:

$$\boldsymbol{Y(k,i) = X(k,i) + N(k,i)} \tag{2.5}$$

where i is the frame index, $Y(k,i)$, $X(k,i)$ and $N(k,i)$ represent the k-th spectral component

from the Fourier transforms of *y*, *x*, and *n* over the time frame *i*, respectively. For improved readability, the frame index *i* is dropped wherever possible. In many speech enhancement algorithms, the knowledge of the power spectrum of speech and noise is essential to their performance. Here we denote $S_y(k)$, $S_x(k)$ and $S_n(k)$ as the power spectrum of noisy speech, clean speech and input additive noise, respectively.

A speech enhancement algorithm estimates the clean short-term spectral amplitude (STSA) by removing the additive noise part. The overall structure of a typical frequency domain speech enhancement algorithm is shown in Fig. 2.1. As shown in the figure, a digitized speech (usually sampled at a rate of 8kHz or 16kHz) is windowed into overlapping frames (typically with the duration of 10-32ms, 50% or 75% overlap) in order to ensure that the speech signal satisfies the assumptions of wide sense stationary. The Hanning or Hamming window is often used. The windowed speech frame passes though the DFT stage and is then separated into the magnitude and phase components. The noisy speech magnitude is filtered while the phase is left untouched since the phase information is less importance in speech enhancement [25][26]. After filtering, the magnitude component is recombined with the original phase, and the IDFT is carried out. Finally, the overlap and add technique [27] is applied to reconstruct the enhanced speech signal.

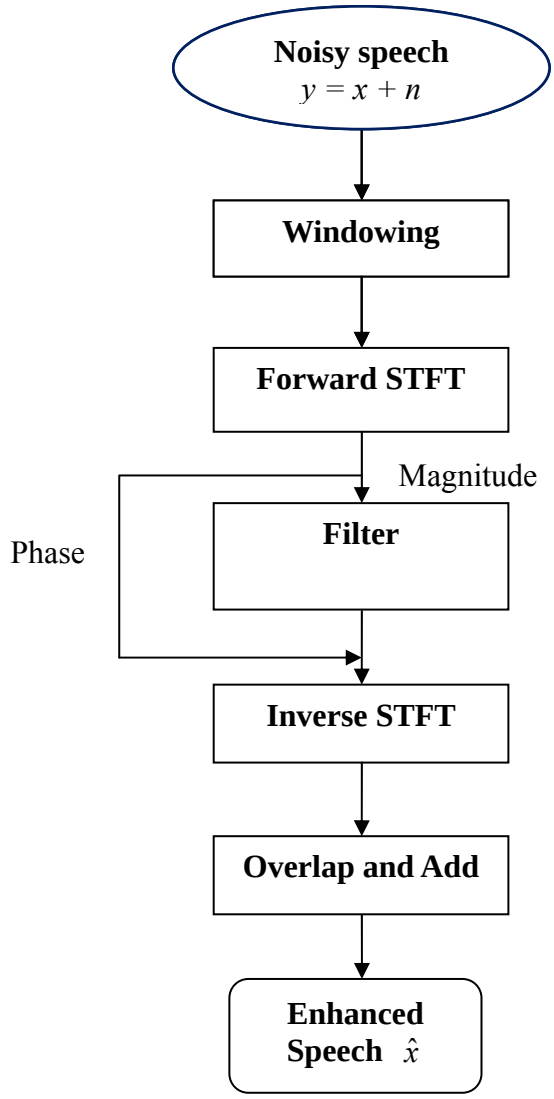


Fig. 2.1 – Block diagram of frequency domain speech enhancement algorithm

2.3 Speech Enhancement Algorithm

2.3.1 Spectral Subtraction

Magnitude spectral subtraction is one of the simplest noise reduction techniques which operate in the frequency domain [12]. Let the magnitude of the filtered speech be $|\boldsymbol{X(k)}|$. The relationship is given as follows:

$$|X(k)| = max(|Y(k)| - E(|N(k)|),0) \quad (2.6)$$

$E(|N(k)|)$ is the average noise magnitude of coefficient k. A half wave rectification process is fulfilled by the $max( )$ function to simply set the negative components to zero in order to avoid possible negative magnitudes by error in the subtraction.

The main problems with the magnitude spectral subtraction are that it does not attenuate noise sufficiently during the silence period and the residual noise has musical tone as shown in Fig. 2.2. The musical residual noise is very annoying to human listeners. Much effort has been devoted to solve this problem with some degree of success. One simple way is to set up a spectral noise floor [28] in order to mask the tonal nature of the residual noise.

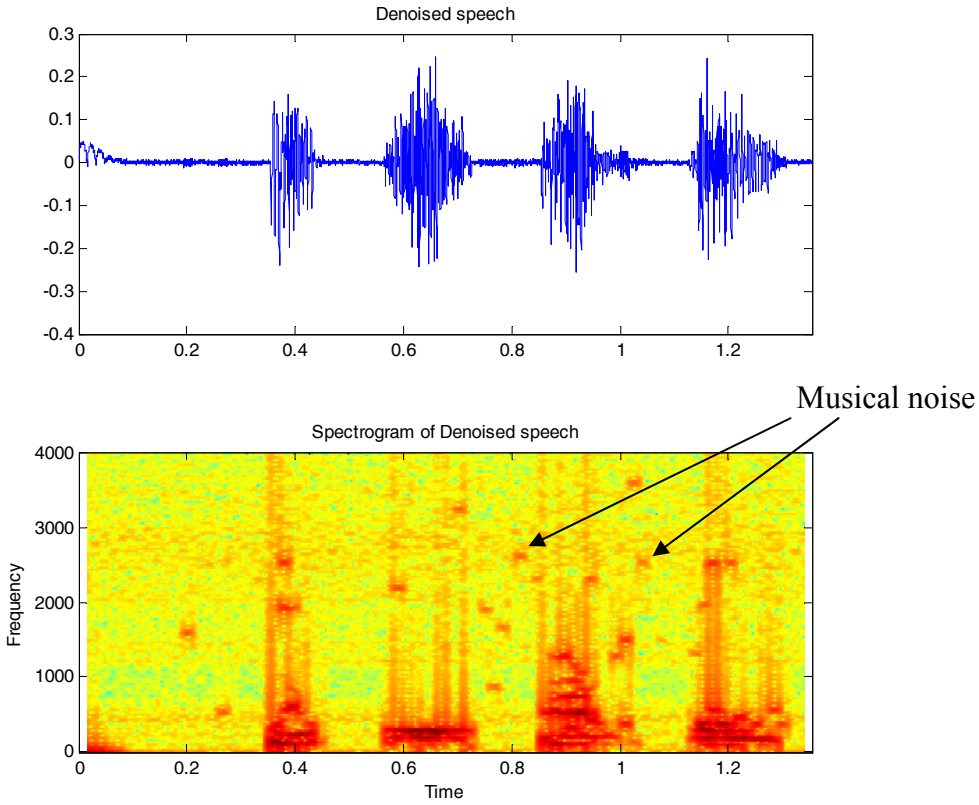


Fig. 2.2– The signal and spectrogram of denoised speech with spectral subtraction.

The isolated peaks indicated in the figure will result in musical noise.

To reduce the musical noise problem, the modified spectral power subtraction algorithm was proposed by Scalart [29] using the spectral power subtraction technique together with the *a-priori* signal to noise ratio, ξ, estimated by the decision-directed approach [14]. In that method, ξ is updated using information from the estimate in previous frame. More specifically, the estimated speech magnitude $\left|\hat{X}(k)\right|$ is related to the noisy speech magnitude $\left|Y(k)\right|$ by,

$$\left|\hat{X}(k)\right| = \sqrt{\frac{\hat{\xi}(k)}{\hat{\xi}(k)+1}}\left|Y(k)\right| \tag{2.7}$$

where

$$\hat{\xi}(k) = \frac{\hat{S}_x(k)}{\hat{S}_n(k)} \tag{2.8}$$

is the estimate of ξ. And

$$\hat{S}_x(k) \cong E\left\{\left|X(k)\right|^2\right\} \tag{2.9}$$

$$\hat{S}_n(k) \cong E\left\{\left|N(k)\right|^2\right\} \tag{2.10}$$

are the estimate of the true power spectrum of the speech and noise, respectively. Both $\hat{S}_n(k)$ and $\hat{S}_x(k)$ have to be known when using the above formula. Methods for estimating $\hat{S}_n(k)$ are covered in some details in [12] and [13] and also in Section 0. $\hat{S}_x(k)$ can be evaluated by the following equation:

$$\hat{S}_x(k,i) = \alpha\hat{S}_x(k,i-1) + (1-\alpha)\max\left\{\left|Y(k,i)\right|^2 - \hat{S}_n(k,i), 0\right\} \tag{2.11}$$

where max{ } is the maximum function in order to ensure that a non-negative value is achieved in the evaluation. $\hat{S}_x(k,i-1)$ is the estimate of $S_x(k)$ in the previous frame, while α is a constant which controls the trade-off between the amount of noise reduction as well as the distortion of speech transients in a speech enhancement framework. It is popular to set the value of α to 0.98. When α is set to 1, severe distortion in the speech signal can be resulted. However, smaller values of α (e.g. 0.8) can lead to high level of musical residual noise. The effect of varying α is investigated in detail in [30], which states that the value of α ought to be greater than 0.9 in order to overcome the musical noise effect and 0.98 is thought about a practical value for α. Although the modified spectral power subtraction filter can result in fewer musical tones, the level of residual noise is still high.

2.3.2 Wiener Filter

Besides (2.7), many other gain functions have been adopted by different research groups. The most common one is the Wiener filter, which has a structure similar to (2.7) as follows:

$$\left|\hat{X}(k)\right| = G_{Wiener}(k)\left|Y(k)\right| = \frac{\hat{\xi}(k)}{\hat{\xi}(k)+1}\left|Y(k)\right| \tag{2.12}$$

where $\hat{\xi}$ is the estimated *a-priori* SNR defined as in the previous section. Amongst the simple filters mentioned in this chapter, the Wiener filter produces the highest noise attenuation, but also introduces significant distortion to the enhancement speeches.

2.3.3 Minimum Mean Square Error (MMSE) Estimator

The Minimum Mean Square Error (MMSE) estimator was proposed by Ephraim and Malah [14]. The estimator is achieved by minimizing the mean square error between the filtered coefficients and the original coefficients in the Fourier transform domain. It is defined as follows:

$$\left|\hat{X}(k)\right| = G_{MMSE}(k)\left|Y(k)\right| = \frac{M(-0.5;1;-v(k))\sqrt{\pi v(k)}}{2\gamma(k)}\left|Y(k)\right| \tag{2.13}$$

where

$$v(k) = \frac{\xi(k)\gamma(k)}{\xi(k)+1} \tag{2.14}$$

$$\gamma(k) = \frac{\left|Y(k)\right|^2}{\hat{S}(k)} \tag{2.15}$$

M(; ;) in (2.13) is the confluent hypergeometric function defined in [31] and it can be computed efficiently by the series summation as follow.

$$M(-0.5;1;-v(k)) = 1 + \frac{ac}{b1!} + \frac{a(a+1)c^2}{b(b+1)2!} + \frac{a(a+1))a+2)c^3}{b(b+1)(b+2)3!} + \ldots \tag{2.16}$$

ξ and γ are the *a-priori* and *a-posteriori* signal to noise ratios respectively. The *a-priori* SNR,

ξ, can also be estimated by the decision-directed approach as follows:

$$\xi(k) = \alpha \frac{\hat{S}_x(k,i-1)}{\hat{S}_n(k,i)} + (1-\alpha)\max\{\gamma(k,i)-1,0\} \tag{2.17}$$

Following Ephraim and Malah's work, several improvements were made to the decision-directed approach. [30] suggested limiting the smallest allowable value for ξ in (2.17). This can be easily done by the following equation:

$$\xi(k) = \alpha \frac{\hat{S}_x(k,i-1)}{\hat{S}_n(k,i)} + (1-\alpha)\max\{\gamma(k,i)-1,\xi_{\min}\} \tag{2.18}$$

where ξ_{min} is the minimum value allowed for ξ. The flooring of ξ to a small value is important for reducing low-level musical noise [30]. Other developments to the decision-directed method focus on reducing the bias and advancing the speed of adaptation [32]-[37], which is introduced by the smoothing constant in (2.17). The main advantage of this filter is that the residual noise is white and avoids the commonly encountered musical tones. Nevertheless, the residual noise can have large magnitude that affects the quality of the enhanced speeches.

2.3.4 MMSE-LSA Estimator

Ephraim [15] also proposed minimizing the log of the error instead of the error itself as it will correspond better to the human hearing mechanism. The resulting estimator is dubbed as the minimum mean square error log-spectral amplitude estimator (MMSE-LSA) estimator which performs better than the MMSE counterpart. The MMSE-LSA estimator is given by the following equation:

$$G_{LSA}(k) = \frac{\xi(k)}{\xi(k)+1}\exp\left(\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{2.19}$$

where v and ξ are defined in (2.14) and Sec. 2.3.3 respectively. The enhanced speech can be obtained by applying the gain function to the magnitude spectrum of the noisy speech as follows:

$$\left|\hat{X}(k)\right| = G_{LSA}(k)\left|Y(k)\right|. \tag{2.20}$$

The MMSE-LSA estimator trims down the residual noise without affecting the speech signal itself, that is, without introducing a large amount speech distortion.

The MMSE-LSA estimators can be extended to apply to the magnitude-squared spectrum [38] or more generally the β-order spectral magnitude spectrum [39]. It is shown in [39] that the β-order MMSE estimator provides higher attenuation for values of β>1, and less attenuation for values of β **<1**. When β is extremely small (i.e., β ≈0.001), the resulting gain curve matches closely the gain function of the MMSE-LSA estimator. Since the order β influences the amount of attenuation, it is reasonable to adjust β adaptively depending on the speech segment, rather than using a fixed value for β.

For the aforementioned MMSE algorithms, a key assumption made is that the real and imaginary parts of the clean speech spectrum can be modeled by a Gaussian distribution. Observing that the real and imaginary parts of the clean speech STFT are successfully modeled by super-Gaussian densities, the use of non-Gaussian distributions for modeling the real and imaginary parts of the speech spectrum have been further studied [40]-[44]. However, the improvement in performance of the MMSE estimators based on non-Gaussian distribution assumption is not significant [1].

2.3.5 Probability of Speech Presence

For the above speech enhancement method, the clean speech estimator is derived under the assumption that speech is always present. It is indeed not true in speech pauses or between spectral bins of the harmonics of a voiced speech. There are two forms of speech absence. The first form of speech absence is due to the speaker pausing in his speech resulting in significant portions of silence. The second form of speech absence is that although the speaker is talking, the speech energy is not present in all frequency components. For some

frequency components with insignificant energy, speech can be considered to be absent in those components. Such knowledge of speech absence can be useful to improve a speech enhancement filter. The first attempt in utilizing the uncertainty of speech absence was explored by McAulay and Malpass [13]. In their approach, they derived a filter based on a fixed probability of speech absence of 0.5. The Ephraim and Malah noise removal filter [14] adopted a more flexible approach in which different spectral frequency components can be assigned a different probability of speech absence which ranges from zero to one. The probability of speech absence is expected to be a function of time and frequency. More specifically, the *k*th spectral output, $\hat{X}(k)$, of the Ephraim and Malah noise suppression filter, taking into account the uncertainty of signal presence, is given by the following equation:

$$\left|\hat{X}(k)\right| = \left[p(k)G_{H1}(k) + (1 - p(k))G_{h0}(k)\right]\left|Y(k)\right| = G_{MMSE+SPP}\left|Y(k)\right| \tag{2.21}$$

where $p(k)$ is the speech presence probability (SPP) defined as follows:

$$p(k) = \frac{1 - q(k)}{1 - q(k) + q(k)(1 + \xi(k))\exp(-u_k)} \tag{2.22}$$

where $u(k) = \dfrac{1 - q(k)}{q(k)}$ (2.23)

where *q(k)* is the probability of speech absence. The *a-priori* SNR, ξ, can best be estimated by the decision-directed approach [14] as (2.17). For speech absence, the clean speech estimator is zero [13][45], and the gain function $G_{MMSE+SPP}$ is given by

$$G_{MMSE+SPP} = p(k)G_{H1}(k) \tag{2.24}$$

However, their work does not touch on how the probability of speech absence can be estimated, and for performance evaluation, the probability of speech absence was set to 0.2 experimentally.

As mentioned above, the MMSE-LSA estimator [15] is especially popular because of its robustness against estimation errors which results in less musical noise [46]. By incorporating the speech presence uncertainty into the MMSE-LSA filter as in Cohen's algorithm given in [46]-[47], the gain function applied to the observed noisy speech spectrum is given as follows:

$$G_{LSA+SPP}(k) = \{G_{LSA}(k)\}^{p(k)} \cdot \{G_{\min}(k)\}^{1-p(k)} \tag{2.25}$$

where $p(k)$ is the SPP as mentioned above; G_{min} is chosen to be a small constant less than 1. For a particular frequency bin k, it can be seen in (2.25) that the overall gain will approach G_{min} if $p(k)$ tends to zero. It means that the original clean speech estimator will not be used if the probability of speech presence in that frequency bin is low. Several approaches have been suggested for estimating and updating $p(k)$ [45][47]-[48]. These methods substantially reduce the residual noise when combining with the statistical estimators. However, the SPP estimator is not always accurate. Recently, in order to avoid wrongly suppressing speech components and leading to a large distortion in the enhanced speech, Gerkmann [49] proposed an improved SPP estimator $p_{fp}(k)$ which achieves by smoothing the *a-posteriori* SNR function, both temporally and spectrally, before applying to the estimation of the generalized likelihood ratio (GLR). Temporal smoothing is achieved by using a time averaging method performed across speech frames, and spectral smoothing is achieved by using a pair of local and global filters applied to the noisy *a-posteriori* SNR function in each frame. Some improvement is noted as compared to the traditional SPP based MMSE-LSA algorithms.

2.4 Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) [70][71] is a popular transform which has found applications in many image processing problems. Its basis functions are localized, and well-suited to the analysis of natural signals (e.g. images, audio, biomedical signal, etc...). They often yield a sparse representation of the signal after transformation. The structure of the DWT is similar to the subband filtering. The input signal is concurrently filtered by a high pass and low pass decomposition filter to produce two signals, detail and approximation, respectively. The approximation signal is the low-frequency component of the signal, while the detail signal is the high-frequency component. Both signals are then downsampled by 2. This composes one level of decomposition and the process can be applied again on the approximation signal to form another level of decomposition. The block diagram for wavelet decomposition is shown in Fig. 2.3.

The DWT has the advantage of using a variable length window for different frequency components. This allows the use of shorter intervals for high-frequency information as well as long-time intervals to obtain more precise low-frequency information. Thus, the characteristics of nonstationary speech signals can be more closely examined. The wavelet basis functions are well localized in time and scale domains, since the support of the wavelet basis functions can be fexibily adjusted (or selected) to achieve the locality required. This behavior of wavelet decomposition is appropriate for processing of speech signals which need high temporal resolution to analyze high-frequency components (mostly unvoiced sounds), as well as high-frequency resolution to analyze low-frequency compoments (voiced sounds, formant frequencies).

The filters can be in different orders and forms (Daubechies, Symlets, Biorthogonal, etc) [72], and the number of decompositions can also be adjusted according to the needs of the application. The main advantage of DWT is its fast implementation as its computational

complexity is in the order N, where N is the data size (in contrast to $N\log_2 N$ for fast Fourier transform (FFT)). Also, the number of taps of the digital filters used in the transform is normally small.

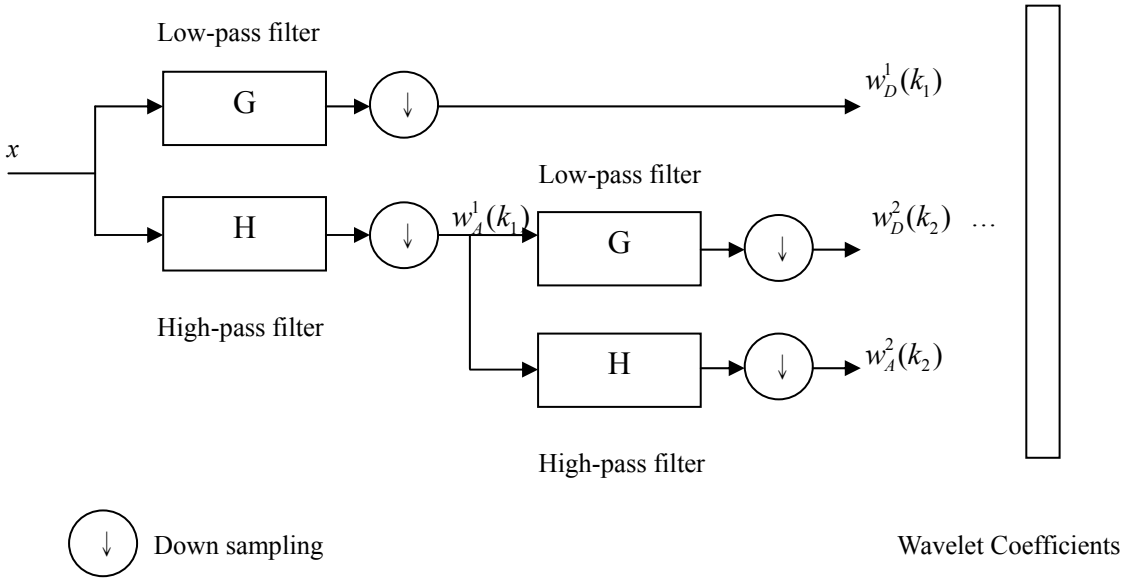


Fig. 2.3 – Block diagram of DWT

2.4.1 Noise reduction with wavelets

Signal denoising is one of the major applications of the wavelet transform. It can be achieved by identifying the singularities in the signal, which can be characterized based on the modulus maxima of the signal's wavelet coefficients [73]. Based on spatial correlation between the wavelet coefficients over adjacent scales, the singularities due to noise can be detected and remove [76]. Another class of popular wavelet denoising methods is based on thresholding the signal's wavelet coefficients [74]. Due to the vanishing moments of the wavelet functions, the wavelet transform of most natural signals will result in only small amount of significant wavelet coefficients (i.e. sparsity). It however is not the case for noises, which have their wavelet coefficients spread over the transform domain but with small magnitude. Hence by thresholding the wavelet coefficients, a large amount of noise energy

will be reduced while the energy for the signal is kept. More specifically, the wavelet thresholding algorithm can be summarized in three steps:

1. DWT of the noisy signal,
2. Thresholding the resulting wavelet coefficients, then
3. Inverse DWT (IDWT) to obtain the denoised signal.

The following soft thresholding function defined in [68] and [77] has been popularly used:

$$\widetilde{w}(k) = \begin{cases} \text{sgn}(w(k))(|w(k)| - thr) & |w(k)| > thr \\ 0 & |w(k)| \le thr \end{cases} \tag{2.26}$$

where *w(k)* represents the wavelet coefficients and *thr* is the threshold. The universal threshold defined in [68] is popularly used. It is given as follows:

$$thr = \sigma_n \sqrt{2\log(M)} \tag{2.27}$$

where σ_n is the noise variance. In practice, the noise variance may not be known. It can be estimated based on the robust statistics as σ_n = MAD/0.6745, where *M* is the number of samples; MAD is the median of the absolute value of the noisy wavelet's coefficients. A level dependent threshold can also be defined as follows [78]:

$$thr_j = \sigma_n^j \sqrt{2\log(M_j)} \tag{2.28}$$

with σ_n = MAD_j/0.6745 and MAD_j is the median of the absolute value of the coefficients, estimated with level *j* wavelet coefficients. The discriminatory threshold can also be determined by using other criterion such as the Stein's unbiased risk estimate (SURE) [79][80].

2.4.2 Application of DWT to speech enhancement

The standard wavelet thresholding has not been successfully applied to speech enhancement directly because the simple threshold cannot discriminate efficiently the speech

components from those of the noise. The wavelet transforms, however, are successfully combined with other denoising methods and can improve the performance of speech enhancement algorithms. They include using the wavelet filter bank for spectral subtraction [81], the Wiener filtering in the wavelet domain [82], or the coherence function [83][84]. Furthermore, an adaptive threshold scheme was developed in [85] on a modified hard thresholding function. A voice activity detector (VAD) was proposed to estimate noise level in colored and non-stationary noise contaminations. Besides, the bionic wavelet transform with the wavelet denoising technique was used to construct a new adaptive thresholding method for speech enhancement in [86]. However, these wavelet based methods usually need an estimation of the noise. In order to enhance the noisy speech without the requirement of an accurate estimation of the noise level, a time and scale dependent wavelet thresholding scheme for speech enhancement was proposed in [87], where the time dependency was introduced by approximating the Teager energy of the wavelet coefficients, and the scale dependency was introduced by using a level-dependent thresholding scheme based on the wavelet packet structure. Since it is known that the noise in the multi-taper spectrum is approximately white and additive, the wavelet thresholding scheme is directly applied to help in estimating the *a-priori* SNR and in turn enhancing the noisy speech [88]. However, the enhanced speech still suffers from annoying musical residual noise in the case that the noise is colored. The musical residual noise is caused by randomly spaced spectral peaks that come and go over successive frames, and occur at random frequencies [88]. To deal with this problem, a wavelet-domain optimal linear estimator which incorporates the masking properties of the human auditory system was proposed to make the residual noise inaudible [89]. In [90], a gain factor which is adapted by both the intra-frame masking properties and the inter-frame SNR variation was further proposed to enhance a noisy speech signal corrupted by non-stationary noise.

2.5 Cepstral Representation

While the Fourier transform is the core operation of a frequency domain speech enhancement algorithm, an important concept that flows directly from the Fourier transform is the cepstrum of speech. The cepstrum of a speech signal is the inverse Fourier transform of the logarithm of its magnitude spectrum [91]. The processing related to the computation of the ceptrum is a special case of the more general concept of homomorphic processing [92]. Cepstral representation is also useful in the context of stochastic signals, especially when the power spectrum of these signals is from an autoregressive model. It has been used in speech recognition for the computation of spectral features [93][94], and in speech coding for the quantization of the spectral envelop [95]. A more detailed discussion of the uses of the cepstrum in speech processing can be found in [96][97].

The cepstrum of a stationary signal x with power spectral density $S_x(k)$ is obtained from the inverse DFT of the logarithm of $S_x(k)$, i.e.,

$$C_x(q) = IDFT\{\log(S_x(k))\} \tag{2.29}$$

where q is the cepstral index, also known as the quefrency index [91]; the natural logarithm is assumed throughout this thesis. Cepstral representation of a signal with an unknown power spectral density is achieved from the inverse DFT of the logarithm of an estimate $\hat{S}_x(k)$ of $S_x(k)$ as follows:

$$\hat{C}_x(q) = IDFT\{\log(\hat{S}_x(k))\} \tag{2.30}$$

Hence $\hat{C}_x(q)$ may be considered as the estimates of the "true" cepstrum $C_x(q)$. Different power spectral density estimates generate different cepstral representation results. Parametric and nonparametric power spectral density estimates have been used in the cepstral representation of speech signals [98][94]. The autoregressive modeling of the signal is the most commonly used method to obtain the parametric power spectral density estimates. The

nonparametric power spectral density estimates for cepstral representation of speech signals consist of the periodogram [99], the smoothed periodogram obtained from the "window method" [99], as well as the closely related mel-spectrum [93][94]. Besides, improved cepstrum estimation via optimal risk smoothing for the periodogram was proposed in [100].

2.5.1 Spectral Estimation via Cepstrum Thresholding

Although periodogram is one of the most popularly used non-parametric estimates of power spectrum used in ceptrum estimation, the high variance of the periodogram, which approaches to the square of the true spectrum as data size increases, can introduce great error to the cepstrum generated. The cepstral thresholding as an approach for variance reduction in spectral estimation of a stationary signal was proposed in [101] and [102]. For many practical situations, it has been observed that a lot of these cepstral coefficients are either zeros or extremely small in magnitude [101]. In fact, many thresholding based cepstrum estimation methods were motivated by this observation. The cepstrum thresholding has been shown to be an effective and automatic way for obtaining a smoothed nonparametric estimate of the spectrum of a stationary signal.

The relationship between the true cepstrum and the cepstrum estimated from the signal's periodogram can be mathematically defined. Given that the true cepstrum is defined as,

$$\boldsymbol{C}_x(q) = \frac{1}{M}\sum_{k=0}^{M-1} \log\left(S_x(k)\right) e^{j2\pi kq/M} \quad ; \quad q=0,\ldots,M\text{-}1 \tag{2.31}$$

where M is the total number of qurefrency. Denote the periodogram of the original signal x as $\hat{S}_x(k)$ such that $\hat{S}_x(k) = \left|X(k)\right|^2$. Let $\hat{\boldsymbol{C}}_x$ be the cepstral coefficients computed from the signal's periodogram vas follows:

$$\hat{C}_x(q) = \frac{1}{M}\sum_{k=0}^{M-1} \log(\hat{S}_x(k)) e^{j2\pi fq/M} \quad ; \quad q=0,\ldots,M\text{-}1 \tag{2.32}$$

Let us further define,

$$\hat{C}_x(q) = \begin{cases} \hat{C}_x(q) + \gamma & for \quad q = 0 \\ \hat{C}_x(q) & otherwise \end{cases} \tag{2.33}$$

where γ = 0.577216 is the Euler's constant. It is shown in [101] and [105] that under some regularity conditions and for large sample size ($M >> 1$) real-valued data, the estimated cepstral coefficients $\hat{\boldsymbol{C}}_x(q)$ are even symmetric and independent random variables having normal distributions with means $\boldsymbol{C}_x(q)$ and variances $\sigma_x^2(q)$ as follows:

$$\hat{\boldsymbol{C}}_x(q) \sim N\left(\boldsymbol{C}_x(q), \sigma_x^2(q)\right) \;; \qquad q=0,\ldots,M/2 \tag{2.34}$$

where

$$\sigma_x^2(q) = \begin{cases} \pi^2 \big/ (3M) & for \quad q = 0, M/2 \\ \pi^2 \big/ (6M) & otherwise \end{cases} \tag{2.35}$$

The above result has been used in the derivation of a number of thresholding based nonparametric spectrum estimation algorithm. The thresholding method can be the simple thresholding (SThresh) [101][102][106] or the empirical Bayes thresholding (EbayesThresh) [107][108]. Further extension for complex signals was proposed in [109].

2.5.2 Speech enhancement in the Cepstral Domain

As mentioned above, spectral outliers in the adaptation of filter gains may emerge that lead to the annoying musical noise. Musical noise is particularly difficult to avoid under non-stationary noise conditions. Different methods have been used to deal with that [45][110] [111]. Many of them are based on reducing the variance using different smoothing approaches in the frequency domain [112][113][114]. A disadvantage of smoothing in the frequency domain is that the frequency and temporal resolutions are decreased. It is not desirable as the temporal smoothing spreads speech onsets and the frequency smoothing reduces the resolution of speech harmonics. Recently, it was

demonstrated that smoothing in the cepstral domain is better than the smoothing in the spectral domain [115][116]. In the cepstral domain, speeches mainly consist of the coefficients in the lower cepstrum which represent the spectral envelope as well as a peak in the upper cepstrum which represents the fundamental frequency and the harmonics [117]. Hence the speech can also be considered to have a sparse representation in the cepstrum domain. As a result, smoothing in the cepstrum domain can reduce the variance without distorting the speech signal. Generally, a cepstral variance reduction can be realized by either setting those cepstral coefficients to zero that are below a certain threshold [101][102], or by selectively smoothing the cepstral coefficients over time [115][116]. A modified cepstrum thresholding was proposed to reduce the non-stationary noise components [118]. In addition, a cepstrum based spectral estimation algorithm was proposed in [119][120]. It makes use of the knowledge about the speech spectral structure so that a better estimate of the noise power spectral density can be obtained.

2.6 Dictionary learning

When analyzing a signal, it is not uncommon to express the signal with an overcomplete representation since the redundancy in the representation often can provide extra information to better the analysis. Different criterion can be adopted when designing an overcomplete transform kernel. Recently, the sparsity of the transform coefficients is of particular interest in many signal processing applications, including signal denoising, restoration and reconstruction, etc. More specifically, if **y** is a signal and **x** is its transform coefficients, the overcomplete transform kernel D (or the so-called "dictionary") is designed based on the sparsity criterion as follows:

$$arg\min_{x}\|\mathbf{x}\|_0 \quad s.t. \quad \mathbf{y} = D\mathbf{x} \tag{2.36}$$

where $\|.\|_0$ is the sparsity measure that counts the number of nonzero coefficients. The above equation shows that the signal **y** can be expressed as the linear combination of only a few column vectors in D (which are also called "atoms"). Two main methods have appeared to determine a dictionary within a sparse decomposition: dictionary selection and dictionary learning. Dictionary selection entails choosing a pre-existing dictionary, such as the Fourier and related bases, wavelet basis, or constructing an overcomplete dictionary by forming a union of bases so that particular properties of the signal can be represented [69]. Besides, dictionary learning aims at deducing the dictionary from the training data, so the coefficients directly capture the specific features of the signal [22]. Dictionary learning methods are commonly based on an alternating optimization strategy, in which the signal representation is fixed, and the dictionary elements are learned; then the sparse signal representation is found, while the dictionary is fixed. Early dictionary learning methods were based on a probabilistic model of the observed data [121][122]. Lewicki and Sejnowski [122] clarified the relation between the independent component analysis (ICA) and the sparse coding methods, while the connection between dictionary learning in sparse coding as well as the vector quantization

problem was presented in [123]. The dictionary based on sparse representation which using variants of the focal underdetermined system solver (FOCUSS) was proposed in [124].

The Method of Optimal Directions (MOD) is one of the first methods to implement the sparsification process [125][126]. Given a set of examples $X = [x_1\ x_2\ \ldots\ x_n]$, the goal of the MOD is to find a dictionary D and a sparse matrix C which minimize the representation error,

$$\arg\ \min_{D,C}\left\|X - DC\right\|_F^2 \quad subject\ \ to\ \ \forall_i, \left\|c_i\right\|_0 \le \varepsilon \tag{2.37}$$

where c_i represent the columns of C; and ε is a very small real number. The resulting optimization problem is combinatorial and highly non-convex, and thus we can only expect for a local minimum at best. The MOD alternates sparse-coding and dictionary update steps similar to other training methods. The MOD is a very effective method as it requires only a few iterations to converge, but the drawback of this method is the relatively high complexity of the matrix inversion.

Aharon et al. [127] proposed the K-singular value decomposition (K-SVD) learning algorithm over redundant dictionaries, which involves a sparse coding stage based on a pursuit method and is followed by a dictionary matrix update step which updates the matrix one column at a time. The K-SVD algorithm defines an initial overcomplete dictionary matrix $\mathrm{D}_0 \in \mathbb{R}^{NxP}$, a set of examples arranged as the columns of the matrix $X \in \mathbb{R}^{NxR}$, and a number of iterations n. The algorithm intends to iteratively improve the dictionary by approximating the solution defined as follows:

$$\min_{D,C}\left\|X - DC\right\|_F^2 \quad subject\ \ to\ \ \begin{matrix}\forall_i, \left\|c_i\right\|_0 \le K \\ \forall_j, \left\|d_j\right\|_2 = 1\end{matrix} \tag{2.38}$$

The vectors d_j denote the rows of D. The normalization constraint on the rows of D is introduced to avoid degeneracy, but it does not have any practical significance to the result.

The K-SVD iteration involves two basic steps: (1) sparse-coding the signals in X given the current dictionary estimate, producing the sparse representations matrix C, and (2)

updating the dictionary atoms given the sparse representations. The sparse-coding step can be implemented using any sparse–approximation method. The dictionary update process of the K-SVD algorithm is performed in a simple atom-by-atom process, rather than performing matrix inversion. The approximation error could be reduced with increasing target sparsity K, but it also increases the computational time.

The atom update is achieved while preserving the sparsity constraints in (2.38). To perform this, the update step applies only those signals in X whose sparse representations use the current atom. Denoting I as the indices of the signals in X which make use of the j-th atom, the update of the atom is achieved by optimizing the following target function,

$$\left\| X_I - DC_I \right\|_F^2 \tag{2.39}$$

for both the atom and its associated coefficient row in C_I. And then, the resulting problem is a simple rank-1 approximation task given by,

$$\{d, g\} := Arg\min_{d,g} \left\| E - dg^T \right\|_F^2 \quad subject \quad to \quad \|d\|_2 = 1 \tag{2.40}$$

where $E = X_I - \sum_{i \neq j} d_i C_{i,I}$ is the error matrix without the j-th atom, d is the updated atom, and g^T is the new coefficients row in C_I. In general, the Singular-Value-Decomposition (SVD) method or further efficiently using some numerical power method can be used to solve the problem directly.

In practice, the process for obtaining the exact result of (2.40) can be rather computationally demanding, because the number of training signals is proportional to the size of E. However, the whole K-SVD algorithm does not target at converging to the global minimum, but a local minimum (hence relies on a good initial guess). Hence an exact solver is often not required. As a simple alternative, an approximate solution is proposed in [128] to reduce the complexity. It is carried out by limiting the iteration to be only one as follows:

$$d := Eg / \|Eg\|_2 \tag{2.41}$$

$$g := E^T d$$

The above process is recognized to finally converge to the optimum, and when truncated, will result in an approximation which still condenses the penalty term [128]. Also this process disposes the need to clearly compute the matrix *E*, as only its products with vectors are required.

In addition, the latest contributions to the field employ parametric models in the training process, which produce structured dictionaries [129][130]. Recently, a different improvement was achieved in online dictionary learning [131] which allows dictionary training to be carried out from very large set of data. It is found to accelerate convergence and improve the training result.

2.6.1 Speech enhancement based on sparse coding

The framework of sparse representation and dictionary learning provides new solutions for speech enhancement. Since expressing a signal by its sparse representation is similar to the traditional coding methods, the operation is also dubbed as the sparse coding. The sparse coding captures the salient features of a signal. While emphasizing the significant coefficients, it is likely that the additive noise which is often represented by the non-significant coefficients will be suppressed. This is the foundation behind the basis pursuit de-noising algorithm [132] and the later denoising algorithms specifically tailored to speech and audio signals [65][133][134][134]. For instance, a greedy adaptive dictionary learning algorithm (GAD) was proposed in [65] for sparsely approximating and denoising speech signals. In order to promote sparsity in the dictionary and in the approximation coefficients, atoms are selected iteratively from the sparsest speech frames. Besides, a non-stationary noise reduction algorithm based on the non-negative latent variable decomposition model of the speech and the noise was proposed in [134]. Then, a combined dictionary of speech and noise using the non-negative matrix factorization (NMF) method was developed to represent noisy speech

signals [135]. A similar method which recovers the clean speech using a composite dictionary consisting of the dictionary for speeches and noises was introduced in [136]. This method implements the composite dictionary learning using the K-SVD algorithm and extends the least angle regression (LARS) algorithm to include a residual coherence stopping criterion and optimized it to solve a large number of simultaneous coding problems efficiently. Finally, a speech enhancement method based on sparse coding the power spectral density (PSD) was presented in [137]. The PSD dictionary of the clean speech signal is trained by the approximate K-SVD algorithm with nonnegative constraint. By combining the estimated PSD with the signal subspace approach based on the short-time spectral amplitude (SSB-STSA), the enhanced speech signal is obtained.

2.7 Background Noise Power Estimation

Besides signal power estimation, the estimation of the background noise power spectrum is also an important task in a typical speech enhancement process. Noise, in contrast to speech, can generate from any kind of source and have any spectral and temporal characteristics. In general, there are some assumptions made about the noise when approaching the speech enhancement problem: (1) noise and speech are statistically independent; (2) noise has a longer period of stationary than speech; and (3) noise is always present in a noisy speech, although its magnitude and frequency response can be varying. When estimating the noise power, most of the proposals in the literature are based on either the bias compensated tracking of spectral minima technique ("minimum statistics") [50], voice activity detection (VAD), recursive averaging [13][17], soft-decision methods [45][51], or a combination of all above [52]-[53].

In practice, the noise power can be estimated adaptively from the silence or speech pause period. A simple method to estimate the noise spectrum is to use a VAD to classify when the speech is absent and then average the signal power spectrum during these intervals. Generally, the averaging time-constant is selected based on the assumed stationarity of the noise. The following VAD decision rule was used in this thesis:

$$VAD = \frac{1}{N}\sum \log \Lambda(k) \underset{H_1}{\overset{H_1}{>}} \delta \tag{2.42}$$

where

$$\Lambda(k) = \frac{1}{1+\xi(k)} \exp\left\{\frac{\gamma(k)\xi(k)}{1+\xi(k)}\right\} \tag{2.43}$$

where ξ and γ are the *a-priori* and *a-posterior* SNRs, respectively, as mentioned above; and ξ is computed using the decision-directed approach with $\alpha = 0.98$. N is the size of DFT, H_1

denotes the hypothesis of speech presence, H_0 denoted the hypothesis of speech absence, and δ is a fixed threshold, which was set to δ = 0.15 in [1]. If speech absence is detected, the noise power spectrum is updated according to the following formulation:

$$\hat{S}_n(k,i) = (1-\beta)|Y(k,i)|^2 + \beta\hat{S}_n(k,i-1) \tag{2.44}$$

where β = 0.98, $\hat{S}_n(k,i)$ is the estimated noise power spectrum of frame *i* (at frequency bin *k*). Such noise power estimator can give good performance for stationary noise. However, large estimation error will result when the noise is non-stationary, particularly for those noises that have their frequency spectrum changes drastically within a short period of time.

2.8 Evaluation of speech enhancement system

Controlled experiments are often conducted when evaluating the performance of a speech enhancement system. In such experiments, standard clean speech segments are added with different kinds of noise retrieved from a standard database. The noises are amplified to different levels so as to test the enhancement performance of the algorithm at different SNR. For instance in our experiments, noises are retrieved from the NOISEX-92 database [193] and are added to clean speech segments extracted from the TIMIT database [192] at different SNR. NOISEX-92 is a noise database which provides various noise signals recorded in real environments. In order to evaluate the performance of speech enhancement systems in real life applications, different noises from the NOISEX-92 database (e.g. the destroyer engine room noise, F16 cockpit noise, buccaneer noise, leopard (Military vehicle) noise, M109 (Tank) and babble noise which is recorded in a canteen with 100 people speaking.) have been used in the experiments.

Speech quality is a measure on how comfortable human listerers perceive a speech signal. Various defects in a speech can affect its quality, such as the final distortion of the speech and the level of the residual noise. The methods used to evaluate speech quality can be divided into the subjective and objective ones. Subjective methods require the participation of human listeners, and can use the preference scoring if a comparison is made between two or more speech signals, or the absolute scoring if a single stimulus is evaluated at each time. One of the most widely used absolute scoring quality measures is the Mean Opinion Score (MOS) [54]. The MOS value is calculated as the average score provided by a number of trained listeners who rate the quality of the speech using a five-point numerical scale, with one indicating “unsatisfactory” or “bad” quality and five indicating “excellent” quality.

Although subjective methods are the only way to obtain true measurements of speech quality, these are expensive in both resourses and time. And the result can be biased

according to the interest of the listeners. Objective methods, in contrast, do not require any evaluation by external personnel and estimate the quality using some analysis of the speech signal, providing an efficient approach for evaluation. Objective measures are popular for evaluating the performance of speech enhancement techniques [55]. In this thesis, we focus on the objective measures by intrusive methods which also require the original clean speech signal to evaluate speech quality. The simplest and most common objective quality measure is the segmental signal-to-noise ratio (segSNR) [56]. The segSNR is calculated by splitting the signals into frames and averaging the calculated SNR in all the frames that contain speech. Several other objective measures were proposed derived from the dissimilarity between all-pole models of the clean as well as enhanced speech signals [57]. Two of the most well-known all-pole based measures used to evaluate speed-enhancement algorithms are the Itakura-Saito (IS) and log-likelihood ratio (LLR) measures. Cepstral distance (CD) measures obtained from the linear predictive coding (LPC) coefficients were also used.

In recent years, perceptually motivated measures have also been popularly used in measuring the speech quality. The Perceptual Evaluation of Speech Quality (PESQ) is defined and becomes an ITU standard, ITU-T P862 [58]. PESQ includes a complex sequence of processing steps to produce a set of distortion scores as a function of time and frequency. The PESQ algorithm provides a quality score on a scale from 0.5 to 4.5 which has been shown to match well with subjective listening tests over a range of telephony channels [59]. The performance of PESQ on processed speech using noise-reduction algorithms was evaluated in [60], where high correlations between 0.83 and 0.96 were achieved across different processing algorithms and noise types.

On the other hand, composite methods have also been used to obtain a higher correlation to subjective measures. A composite objective measure is introduced in [61] for the objective quality rating of speech enhancement methods. It is compared with P.835 subjective

measures [62] for various SNR, noise types and enhancement algorithms [63]. It is shown that composite objective measures can give the best predictor for different subjective measures.

Chapter 3 Wavelet Based Speech Presence Probability Estimator for Speech Enhancement

3.1 Introduction

Speech enhancement is a challenging problem due to the diversity of noise sources and their effects in different applications [1]. As it has widespread applications in speech communications and recognition, continuous effort is being exerted to its investigation [14],[15],[35],[46]-[49],[88],[139]-[145] although an optimal solution is yet to be found. Many speech enhancement methods work in the frequency domain. In these approaches, a noisy speech signal is divided into overlapped frames and the short-time Fourier transform (STFT) is applied to each frame to obtain its frequency spectrum [1]. A gain function is then applied to suppress the selected frequency components in order to reduce the effect of noise to the speech [14],[15]. For these speech enhancement algorithms, a reliable estimator for speech presence probability (SPP) can significantly improve their performance. It is because clean speech estimators used in these approaches are often derived under the assumption that speech is always present. It is indeed not true in speech pauses or between spectral bins of the harmonics of a voiced speech. Consequently, SPP estimator is used [46],[47],[49],[140] to help in detecting the non-speech frequency components and further suppress them. For instance, in Cohen's algorithm given in [46],[47], the popular MMSE-LSA gain function [15] is modified such that if a frequency component is expected to have insignificant speech energy, the gain function will approach to a small constant rather than the original MMSE-LSA one. However, the SPP estimator in [46],[47] is not always accurate. Speech components can be wrongly suppressed and leads to a large distortion in the enhanced speech. In [49], it was suggested that a good SPP estimator can be achieved by

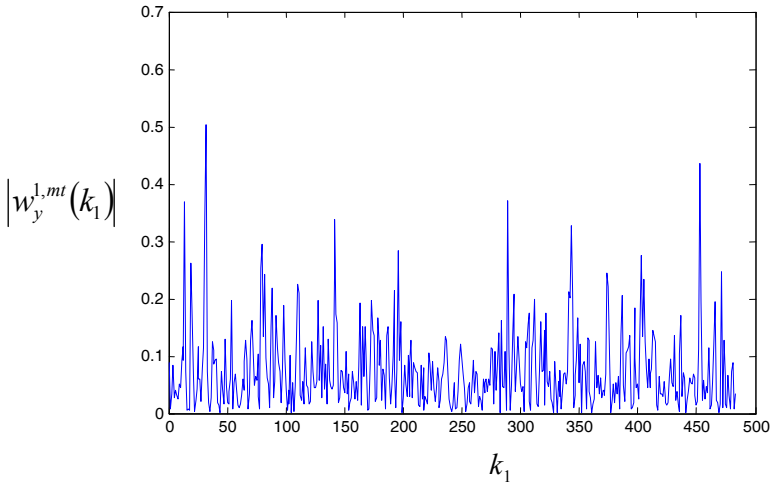
smoothing the *a-posteriori* SNR function, both temporally and spectrally, before applying to the estimation of the generalized likelihood ratio (GLR) (an important parameter for estimating SPP). Temporal smoothing is achieved by using a time averaging method performed across speech frames. Spectral smoothing is achieved by using a pair of local and global filters applied to the noisy *a-posteriori* SNR function in each frame. The resulting SPP estimator achieves probabilities close to zero for speech absence and probabilities close to one for speech presence. Although such feature is extremely useful for suppressing frequency components of noise, it has the side effect that any error in the estimation of speech absence can lead to a sudden jump in the SPP function and give rise to the musical noise [1] in the enhanced speech. To solve the problem, another approach using temporal cepstrum smoothing was proposed to improve the smoothing of the *a-posteriori* SNR [140]. Although estimation errors are noticeably reduced, the approach requires many empirically set parameters that make the generalization of the approach difficult.

As shown in [49], the performance of the smoothing process is crucial to the accuracy of SPP estimation and in turn the quality of the enhanced speech. In fact, the local and global filters used in [49] resemble a multiresolution filter bank which has been studied extensively in the wavelet community. In particular, Moulin suggested applying the wavelet denoising technique for smoothing the power spectrum of signals [146]. It was then shown in [147] that the wavelet denoising technique is particularly effective when applying to the multitaper spectrum (MTS) [149] of a signal. The MTS estimation technique was suggested to reduce the error variance when estimating the power spectrum of a signal using the STFT. It is obtained by averaging the periodograms of a signal generated using a number of orthonormal tapers. The error variance can be reduced by a factor of L where L is the number of tapers. Besides, if we let η be the error between the log MTS and the log power spectrum of the signal, it was shown that η is Gaussian distributed if the number of tapers is 5 or more. Hence,

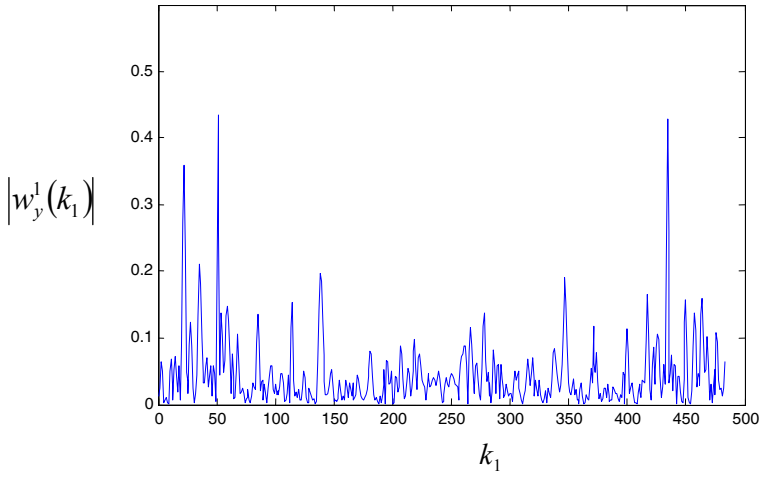
different wavelet denoising techniques, such as the wavelet shrinkage [68][79], were directly applied [147] in order to further enhance the estimation. Such technique was first applied to speech signals in [88] with some success reported. When implementing the wavelet shrinkage algorithm, the universal threshold was commonly used in the traditional approaches [88] [146]-[147] (although a SUREshrink approach was also suggested in [88] at the same time). It is known that in order to have the universal threshold effectively performed, the input noise process should be white Gaussian. However, it can be easily shown that even if the input additive noise process is white Gaussian, its log MTS can hardly be considered as white. Correlation exists among neighbored spectral components, and will increase when more tapers are used. Hence, the universal threshold is often far from optimal when using in the denoising of the log MTS of noisy speeches. In fact, the dynamic range of a speech power spectrum is highly compressed when transforming to the log domain. The spectral peaks of the speech are thus smoothened such that their wavelet coefficients in the log domain can have a magnitude similar to those of noise. It is particularly the case for some weak speech frames with low SNR. A slight error in threshold estimation can either remove a lot of speech wavelet coefficients or leave behind many wavelet coefficients of noise. The former will degrade the speech intelligibility while the later will lead to the annoying "musical" noise, which are both undesirable as far as the overall performance is concerned. As an example, we show in Fig. 3.1a and Fig. 3.1b the first level wavelet coefficients (absolute value and 4-tap symlets (sym4) wavelets) of the log MTS and periodogram, respectively, of a typical speech frame with white noise. We also show in Fig. 3.1c the periodogram of the original clean speech frame. 5 tapers are used for the generation of the MTS. From Fig. 3.1c, we know that the wavelet coefficients at both ends of Fig. 3.1a and Fig. 3.1b should correspond to the speech. It can be seen in Fig. 3.1b that the speech wavelet coefficients of the noisy periodogram are relatively easier to be identified from those of the noise floor. It is not the

case for log MTS. As can be seen in Fig. 3.1a, the wavelet coefficients of speech and the noise floor in the log MTS can have similar magnitude as mentioned above. A simple thresholding scheme will have much difficulty to separate them.

(a) Level 1 wavelet coefficients $\left|w_y^{1,mt}(k_1)\right|$ (see (3.18)) of a log MTS (noisy speech)



(b) Level 1 wavelet coefficients $\left|w_y^{1}(k_1)\right|$ (see (3.11)) of a periodogram (noisy speech)



(c) Level 1 wavelet cofficients $\left|w_x^{1}(k_1)\right|$ (see (3.11)) of a periodogram (clean speech)

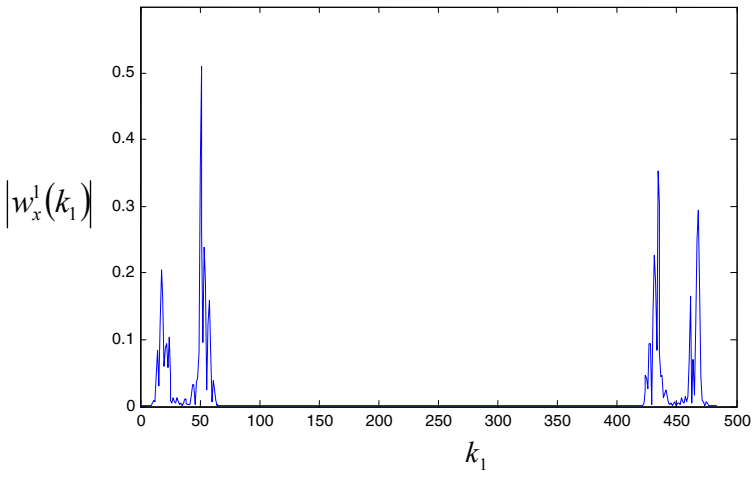


Fig. 3.1 – Level 1 wavelet coefficients (absolute value) of (a) the log MTS of a typical speech frame with white noise; (b) the periodogram of the same noisy frame; and (c) the periodogram of the same speech frame without noise.

In this chapter, we propose a new algorithm for the estimation of the SPP. Rather than using the local and global filters as in [49], we obtain a smooth *a-posteriori* SNR function by denoising the speech power spectrum using the a new wavelet based MTS estimator. The use of the wavelet transform allows a speech power spectrum to be analyzed with filters of arbitrary number of resolutions, rather than two (local and global) as in [49]. Besides, we shall benefit from the studies in the wavelet community in determining the various thresholds for denoising a speech power spectrum. The proposed estimation algorithm can be divided into two stages. First, we apply the wavelet transform to the observed noisy speech periodogram. The wavelet coefficients that are likely incurred by the spectral peaks are detected. It forms an oracle that indicates the approximate spectral locations where the wavelet coefficients of the spectral peaks can likely be found. Second, we apply another wavelet transform to the log MTS of the noisy speech. Based on the locality property of the wavelet transform, it is safe to assume that the oracle obtained in stage 1 can also indicate the wavelet coefficients of the spectral peaks in the log MTS domain. Hence, these wavelet coefficients should be kept. Besides, we also keep those wavelet coefficients which are greater than a threshold derived based on the Stein's unbiased risk estimator (SURE) and are in the vicinity of those indicated in the oracle. These coefficients are also likely to be the coefficients of spectral peaks. For the rest, they are considered as the coefficients of the noise floor and are removed. A smooth speech power spectrum can thus be reconstructed from the remaining wavelet coefficients and is then used to compute the *a-posteriori* SNR of the noisy speech signal. Due to the change in the smoothing procedure, we also propose a new estimation method of the generalized likelihood ratio (GLR), which is an important parameter for SPP estimation. The new SPP estimator can then be adopted in different speech enhancement algorithms, such as the popular MMSE-LSA [15]. When comparing with the traditional SPP estimators, better performance was achieved in most cases when using the

proposed SPP estimator evaluated using different standard measures as mentioned in [1].

This chapter is organized as follows. In Section 3.2, the background of the traditional SPP estimation method for speech enhancement, and the use of MTS and wavelet denoising for smoothing the power spectrum of a noisy speech signal are presented. The proposed 2-step wavelet denoising algorithm for smoothing the noisy speech power spectrum is described in Section 3.4. The new estimation method for the GLR is shown in Section 3.5. Simulation results are shown in Section 3.6, and we summarize the results in Section 3.7.

The results in the chapter have also been reported in [160] and [180].

3.2 SPP Estimation, MTS and Wavelet Denoising

As mention in Chatper 2, many approaches have been suggested for estimating the SPP *p(k)*. Recently, Gerkmann et al. proposed an improved SPP estimator $p_{fp}(k)$ [49] which adopts a fixed *a-priori* SNR and a fixed prior probability of speech presence. More importantly, they suggested that a good SPP estimator can be obtained by smoothing the *a-posteriori* SNR function, both temporally and spectrally, before applying to the estimation of the generalized likelihood ratio (GLR) given as follows:

$$\Lambda = \frac{q}{(1-q)} \frac{p(\gamma \mid H_1)}{p(\gamma \mid H_0)}, \tag{3.1}$$

where Λ is the GLR and q is the *a-priori* speech presence probability. $p(\gamma \mid H_1)$ is the probability density function (PDF) of γ under the hypothesis H_1, i.e. speech is present. Similarly, $p(\gamma \mid H_0)$ is the PDF of γ under the hypothesis H_0, i.e. speech is absent. The SPP can be computed based on the GLR as follows:

$$spp = P(H_1 \mid \gamma) = \frac{\Lambda}{1+\Lambda}. \tag{3.2}$$

In [49], the temporal smoothing is achieved by using a time averaging method performed across speech frames. Spectral smoothing is achieved by using a pair of local and global

filters applied to the noisy *a-posteriori* SNR function in each frame. The smoothing of the *a-posteriori* SNR function needs to be carefully designed to minimize the error in the estimation of speech absence. On the other hand, we also need to make sure the *a-posteriori* SNR function is not over-smoothed such that all spectral peaks are kept during the smoothing process. They are important to the intelligibility of the enhanced speech.

One of the approaches to smooth the *a-posteriori* SNR function is by denoising the observed noisy speech power spectrum $\hat{S}_y$. To facilitate the design of the required denoising algorithm, we need to have a better understanding of the noise that appeared in $\hat{S}_y$. Traditionally, $\hat{S}_y$ is computed using the short-time periodogram defined as follows:

$$\hat{S}_y(i,\omega) = \frac{1}{N}\left|\sum_{n=0}^{N-1} y_i(n)e^{-j\omega n}\right|^2, \tag{3.3}$$

where i is the frame index. Although y is the result of x with noise added, it is well understood that the noise found in $\hat{S}_y$ does not really follow an additive model. It is actually the result of the stochastic noise inherent in the estimation when computing the short-time periodogram. For the rest of this chapter, we dub such noise as the "structural noise" since its behavior is determined by the structure of the power spectrum estimator rather than the input signal. It should be differentiated from the "input additive noise" that is often found during speech acquisition. The variance of the structural noise however is proportional to the sum of the true power spectrum of the speech and the additive input noise process. It means that for spectral peaks in the power spectrum, the variance of the structural noise can be extremely high. For spectral valleys, the variance of the structural noise can be much lower. Hence, we cannot just use a single measure for smoothing $\hat{S}_y$. Rather, a smoothing scheme that can be adaptive to the noise variance is needed.

The MTS approach was adopted in [88] for reducing the variance of the structural noise

of speech power spectrums. The MTS reduce this variance by computing a small number (L) of direct spectrum estimators each with a different tapter (window), and then average the L spectral estimates. The MTS of frame i, where $i \in Z$, of a noisy speech y, is defined as:

$$\hat{S}_y^{mt}(i,k) = \frac{1}{L}\sum_{l=1}^{L} \hat{S}_y^{l}(i,k) \quad , \tag{3.4}$$

where $\hat{S}_y^{l}(i,k) = \left| \sum_{n=0}^{M-1} a_l(n) y_i(n) e^{-j2\pi kn/M} \right|^2$ and $k = 0,1, \ldots, M\text{-}1$.

Each noisy speech frame has N samples. If $N < M$, each speech frame will be padded with zeros to ensure the sequence length is M. The tapers a_l, for $l = 1 \ldots L$, are designed to be orthonormal. One of the popular choices is the sine tapers,

$$a_l(n) = \sqrt{\frac{2}{N+1}} \sin\left(\frac{\pi l(n+1)}{N+1}\right). \tag{3.5}$$

The sine tapers were shown to make less significant local bias with roughly the same spectral concentration [88]. When the tapers are orthonormal, the average of the output of all tapers will reduce the variance of the structural noise by a factor of L, where L is the number of tapers used. To further reduce the variance, the wavelet denoising techniques were suggested to apply to MTS. It was shown that [88][147] for nearly all k (except near $k = 0$ and $M/2$) the difference between the log MTS and the true log power spectrum is found to be Gaussian distributed if $L = 5$ or more. Besides, the variance of the structural noise in the log MTS domain will become constant, irrespective to its variance in the original MTS domain [105]. Consequently, the wavelet denoising techniques, such as the wavelet shrinkage [68][79], can be applied to the log MTS [88][147] to further reduce the noise variance.

When implementing the wavelet shrinkage, the threshold *thr* is perhaps the most important parameter that determines the denoising performance. Many important findings on its selection have been reported in recent years. Donoho first proposed the "universal

threshold" in [68]: $thr = \sigma_n \sqrt{2\log(M)}$, where σ_n^2 denotes the noise variance, and M is the number of samples. For colored noises, the SURE (based on the principle of Stein's unbiased risk estimation [150]) approach was proposed in [79] since the threshold thus derived can be level dependent. Other choices of threshold can be found in [151]-[153].

3.3 Analysis of Multitaper Spectrum of Noise

In this section, we analyze the characteristic of the structural noise in MTS to illustrate how the traditional approach can be improved. Note that in [147], the universal threshold is used to denoise the log MTS of noisy speeches. It is also used in [88] although a SUREshrink approach is suggested at the same time. In order to have the universal threshold effectively performed, the noise process should be additive and white Gaussian such that the noise variance at all levels of the wavelet transform is the same. However, as it is shown below, the noise process can hardly be considered as white particularly when many tapers are used.

In [148], it is shown that, for zero-mean white Gaussian input with variance σ^2, the covariance of the periodogram at frequencies ω_1 and ω_2 is given by:

$$\operatorname{cov}\left\{\hat{S}(k_1)\hat{S}(k_2)\right\} = \sigma^4 \left\{ \left(\frac{\sin[\pi(k_1+k_2)]}{N\sin[\pi(k_1+k_2)/N]} \right)^2 + \left(\frac{\sin[\pi(k_1-k_2)]}{N\sin[\pi(k_1-k_2)/N]} \right)^2 \right\} \tag{3.6}$$

where $k_1 = N\omega_1/2\pi$ and $k_2 = N\omega_2/2\pi$ for integer k_1 and k_2. Since both terms in (3.2) are equal to zero if $k_1 \neq k_2$, a conclusion is then drawn in [148] that the periodogram is also white since the covariance between adjacent frequency components is zero. However in actual implementation, the frame size N is often different from the FFT length M as indicated in (3.4). A speech frame is zero-padded to form a longer sequence for computing its periodogram [1]. In this case, (3.6) should be rewritten as follows:

$$\operatorname{cov}\left\{\hat{S}(k_1)\hat{S}(k_2)\right\} = \sigma^4\left\{\left(\frac{\sin[\pi(k_1+k_2)N/M]}{N\sin[\pi(k_1+k_2)/M]}\right)^2 + \left(\frac{\sin[\pi(k_1-k_2)N/M]}{N\sin[\pi(k_1-k_2)/M]}\right)^2\right\} \tag{3.7}$$

where $k_1 = M\omega_1/2\pi$ and $k_2 = M\omega_2/2\pi$ for integer k_1 and k_2 and $M > N$. It can be seen that both terms in (3.7) are not necessarily equal to zero if $k_1 \neq k_2$. It depends on the shape of the taper, in this case the sinc function since the rectangular taper is assumed. In general, it can be easily shown that for any taper al for $l \in \mathrm{Z}$, the covariance of the periodogram at frequencies ω_1 and ω_2 is given by:

$$\operatorname{cov}\left\{\hat{S}^{a_l}(\omega_1)\hat{S}^{a_l}(\omega_2)\right\} = \frac{\sigma^4}{C_l^2 N^2}\left\{A_l(\omega_1+\omega_2)^2 + A_l(\omega_1-\omega_2)^2\right\} \tag{3.8}$$

where

$$A_l(\omega) = \sum_{n=0}^{N-1} a_l(n)^2 e^{-j\omega n} \quad \text{and} \quad C_l = \frac{1}{N}\sum_{n=0}^{N-1} a_l(n)^2 . \tag{3.9}$$

(3.8) shows that the covariance of the periodogram at frequencies ω_1 and ω_2 depends on the power spectrum of the taper applied. The second term of (3.8) is particularly important. It indicates that there will be covariance between adjacent frequencies ω_1 and ω_2, $\omega_1 \neq \omega_2$, if A_l is not an impulse. For most commonly used tapers such as Hanning or Hamming window, their energy is highly concentrated at DC but not an impulse. Hence the covariance between adjacent frequency components of a periodogram is not exactly zero, although the value will die away quickly as ω_1 and ω_2 become farther apart.

Let us further consider the case when the MTS of a white Gaussian input additive noise is computed. Since the tapers used for computing the MTS are chosen to be orthonormal, the cross covariance between the periodograms generated by different tapers should be zero. Hence it can be shown that:

$$\operatorname{cov}\left\{\hat{S}^{mt}(\omega_1)\hat{S}^{mt}(\omega_2)\right\} = \sum_{l=1}^{L}\frac{\sigma^4}{C_l^2 N^2 L^2}\left\{A_l(\omega_1+\omega_2)^2 + A_l(\omega_1-\omega_2)^2\right\} \tag{3.10}$$

where C_l is defined as in (3.9). (3.10) shows that the covariance of the MTS at frequencies ω_1 and ω_2 is the sum of the covariance of the periodogram at frequencies ω_1 and ω_2 generated by all tapers. Similarly, the second term in the bracket of (3.10) indicates that there will be covariance between adjacent frequencies ω_1 and ω_2, $\omega_1 \neq \omega_2$, if A_l is not an impulse for all l. Fig. 3.2 shows the power spectrum of the sine tapers generated by using (3.5), where L = 5. It can be seen that except the first taper, all other tapers do not have energy centered at dc. Hence when summing up their power spectrum, there will be a wide spread of energy around dc that the resulting power spectrum is certainly far from an impulse.

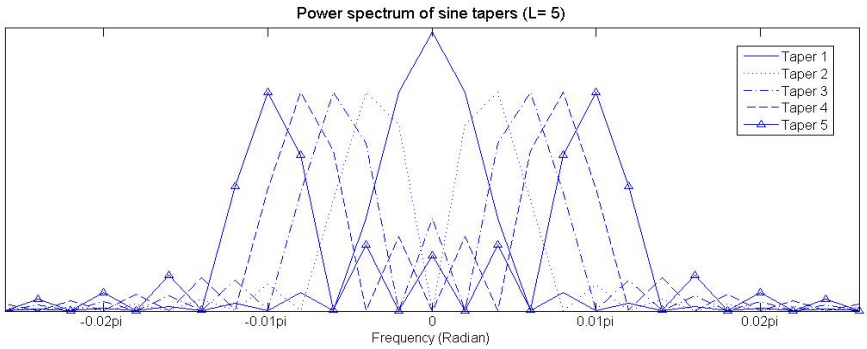


Fig. 3.2 – The power spectrum of the tapers generated by (3.5), where L = 5, taper size N = 480 computed using FFT with size M = 960

Eqn. (3.10) shows that the covariance between adjacent frequency components of MTS can hardly be considered as zero. The covariance will further increase when more tapers are used in the MTS evaluation. It will be the same after taking log of the MTS since the logarithm operator will not change the covariance between MTS frequency components. This local covariance of the log MTS introduces much difficulty when denoising it using wavelet shrinkage and universal threshold. It is because in this case the noise variance at each level of the wavelet transform will be different. Using the same universal threshold for all levels will introduce great error to the denoising process. To illustrate this, an experiment was conducted to apply the wavelet transform to the log MTS of non-speech frames with only white noise.

In the experiment, the discrete wavelet transform (DWT) with sym4 wavelet was applied to the log MTS ($L = 5$) of a number of non-speech frames (i.e. these frames contain only the additive white noise). The variance of the resulting wavelet coefficients at each level is recorded as follows: level 1 – 0.016; level 2 – 0.112; level 3 – 0.474; level 4 – 1.078. Since DWT is linear and orthogonal, the variance of the wavelet coefficients at all levels should be similar if the structural noise is white. The above result however shows that it is not the case.

3.4 Proposed 2-stage wavelet denoising algorithm

In this section, the proposed algorithm for smoothing the noisy speech power spectrum is described. Similar to the traditional approaches [88][147], the new algorithm works in the log MTS domain and uses the wavelet denoising method for reducing the variance of the structural noise. Denoising in the log MTS domain rather than the MTS domain is preferred because it is generally believed that the enhancement process can follow better the perceptual characteristics of the human auditory system when it is carried out in the log domain [15]. Besides, denoising in the log MTS domain avoids the ambiguity arisen from the negative valued power spectral coefficients generated due to the non-linear wavelet denoising process. The proposed algorithm can be divided into two stages as described below.

A. First Stage

For a noisy speech power spectrum, we can often find spectral peaks contributed by the speech and/or the input additive noise (for certain kinds of colored noise). On the other hand, we can also find regions where no spectral peaks can be found. Let us call these regions as the noise floor. It is important to have a smooth noise floor since any large variance structural noise that exists on the noise floor will likely contribute to the annoying musical noise in the enhanced speech. Although the noise floor contains no spectral peak, we have shown in Fig. 3.1a that its log MTS can have large wavelet coefficients with magnitude similar to those of

speech. It means that directly thresholding the wavelet coefficients in the log MTS domain cannot smooth the noise floor. To solve the problem, it is desirable if we have an oracle that indicates the locations of the wavelet coefficients of the noise floor in the log MTS. Such oracle indeed can be obtained from the periodogram of the noisy speech. Let $\hat{S}_y^{a_l}$ be the periodogram of a noisy speech frame y generated using a taper a_l and

$$w_y^j = W\left\{\hat{S}_y^{a_l}\right\} \tag{3.11}$$

be its level j wavelet coefficients, where $W\{.\}$ is the wavelet transform. While $\hat{S}_y^{a_l}$ behaves also like a noisy signal, it is obvious that w_y^j will be spread out for all j with magnitude depending on the local variance of $\hat{S}_y^{a_l}$. That is, if $\hat{S}_y^{a_l}$ has a large variance for some frequencies k, the corresponding w_y^j for all j will also have large magnitude due to the locality property of the wavelet transform. Let us denote $\left(\sigma_{nfloor}^j\right)^2$ to be the variance of the wavelet coefficients of the noise floor at level j. Then if $\hat{S}_y^{a_l}$ contains a spectral peak at frequency bin k, the magnitude of the respective wavelet coefficients $w_y^j(k_j)$ are likely to be much bigger than σ_{nfloor}^j:

$$\left|w_y^j(k_j)\right| >> \sigma_{nfloor}^j \quad \forall\, j \tag{3.12}$$

This allows us to use a simple thresholding scheme to identify the wavelet coefficients of the noise floor. Firstly we need to have a good estimation of σ_{nfloor}^j. An intuitive approach is to use a Voice Activity Detector (VAD) (such as [155]-[156]) to find out the noise frames (frames that contain only noise) and then estimate the standard deviation of the noise wavelet coefficients by averaging across these frames. However, due to the frequency response of different colored noise (such as pink noise, high frequency noise, etc.), the noise power spectrum can contain a limited number of spectral peaks and induce coefficients with large

magnitude in the wavelet domain. Directly evaluating the standard deviation by averaging the noise frames will have significant error. To solve the problem, robust statistics [157] is adopted to avoid the estimation from being affected by the outliers. Among various robust estimators, the median absolute deviation (MAD) is a robust measure of the variability of a univariate sample of quantitative data. For a univariate data set $X_1, X_2, ..., X_n$, the MAD is defined as the median of the absolute deviations from the data's median:

$$MAD = median_i\left(\left|X_i - median_j\left(X_j\right)\right|\right) \quad . \tag{3.13}$$

In order to use the MAD as a consistent estimator of the standard deviation σ of the data set, one takes

$$\hat{\sigma} = K * MAD \quad . \tag{3.14}$$

where K is a constant scale factor, which depends on the distribution. For Gaussian distributed data, K is taken to be $1/\Phi^{-1}(3/4) \approx 1.4826 = 1/0.6745$, where Φ^{-1} is the inverse of the cumulative distribution function for the standard normal distribution, i.e., the quantile function. Let us further define $w_{yn}^j(k_j)$ be the wavelet coefficients of the noise frames. Follow the same argument as in [154], it is reasonable to regard $w_{yn}^j(k_j)$ as approximately Gaussian distributed since $w_{yn}^j(k_j)$ are just linear combinations of $\hat{S}_y^{a_l}$, which are independent random variables. Hence we can apply the MAD for estimating σ_{nfloor}^j. Knowing that $\text{median}\left\{ w_{yn}^j(k_j)\right\} = \text{mean}\left\{ w_{yn}^j(k_j)\right\} = 0$, we have

$$\sigma_{nfloor}^j = K * MAD = \text{median}\left\{ \left|w_{yn}^j(k_j)\right|\right\} / 0.6745 \quad \forall j \quad . \tag{3.15}$$

Based on σ_{nfloor}^j, we can develop a threshold thr_j such that if $\left|w_y^j(k_j)\right| < thr_j$, we consider $w_y^j(k_j)$ belongs to the noise floor. Since $w_y^j(k_j)$ is approximately Gaussian distributed, we propose to use the following level dependent universal threshold to carry out the above

classification:

$$thr_j = \sigma_{nfloor}^{j}\sqrt{2\log(M_j)}, \tag{3.16}$$

where M_j is the number of wavelet coefficients at level j. The universal threshold is chosen because, as shown by Picklands [158] that for a stationary $\varepsilon(i) \sim N(0,1)$ with $\lim_{k\to\infty} E\big(\varepsilon(i+k)\varepsilon(i)\big)=0$, $\max\{\varepsilon(i)\}/\sqrt{2\log M} \to 1$ almost surely as $n\to\infty$. It means that given a set of Gaussian random variables, the universal threshold is their maximum limit asymptotically [68]. Note that the threshold thr_j is level dependent since $\hat{S}_y^{a_l}$ is in general not white. In practice, there can be wavelet coefficients of the noise floor having magnitude greater than the threshold. These outliers, although in a small amount, can still introduce large errors in the estimated SPP and become the source of musical noise. To take care of these outliers, we propose to combine the results of using two orthonormal tapers. It is because the outliers found in the periodogram generated by one taper may not exist in the periodogram generated by the other. To be specific, if a wavelet coefficient is smaller than the threshold in any of the two periodograms, it will be classified as the wavelet coefficient of the noise floor. σ_{nfloor}^{j} needs to be updated from time-to-time using the noise frames detected by a VAD. By using thr_j, an oracle of the spectral locations of the noise floor in the wavelet domain is obtained and will be used in the second stage of the algorithm. As an example, we show in Fig. 3.3 the classification result using the proposed approach. In Fig. 3.3a, the level 2 wavelet coefficients (absolute value) of the periodogram of a typical speech frame with pink noise are shown. The classification result is shown in Fig. 3.3b: a '0' represents the corresponding wavelet coefficient classified as belonging to the noise floor; a '1' represents the corresponding wavelet coefficient classified as belonging to a spectral peak. It can be seen that the proposed approach accurately classifies the wavelet coefficients of the spectral peaks.

To summarize, the procedure of the first stage of the proposed algorithm is as follows:

1. Evaluate two periodograms $\hat{S}_y^{a_1}$ and $\hat{S}_y^{a_2}$ of the observed noisy speech frame *y* using two orthonormal sine tapers a_1 and a_2, respectively.
2. Generate 4 levels of wavelet coefficients of the two periodograms, i.e. $w_y^{j,1} = W\left\{\hat{S}_y^{a_1}\right\}$ and $w_y^{j,2} = W\left\{\hat{S}_y^{a_2}\right\}$, where $W\{\cdot\}$ is the wavelet transform and $j = 1,2,\ldots,4$.
3. From the noise frames, compute σ_{nfloor}^{j} for all *j*.
4. Generate an oracle V_j for classifying the speech and noise wavelet coefficients as follows:

$$V_j(k_j) = \begin{cases} 1 & \left|w_y^{j,1}(k_j)\right| > thr_j \text{ and } \left|w_y^{j,2}(k_j)\right| > thr_j \\ 0 & \text{otherwise} \end{cases} \tag{3.17}$$

where thr_j is defined in (3.15).

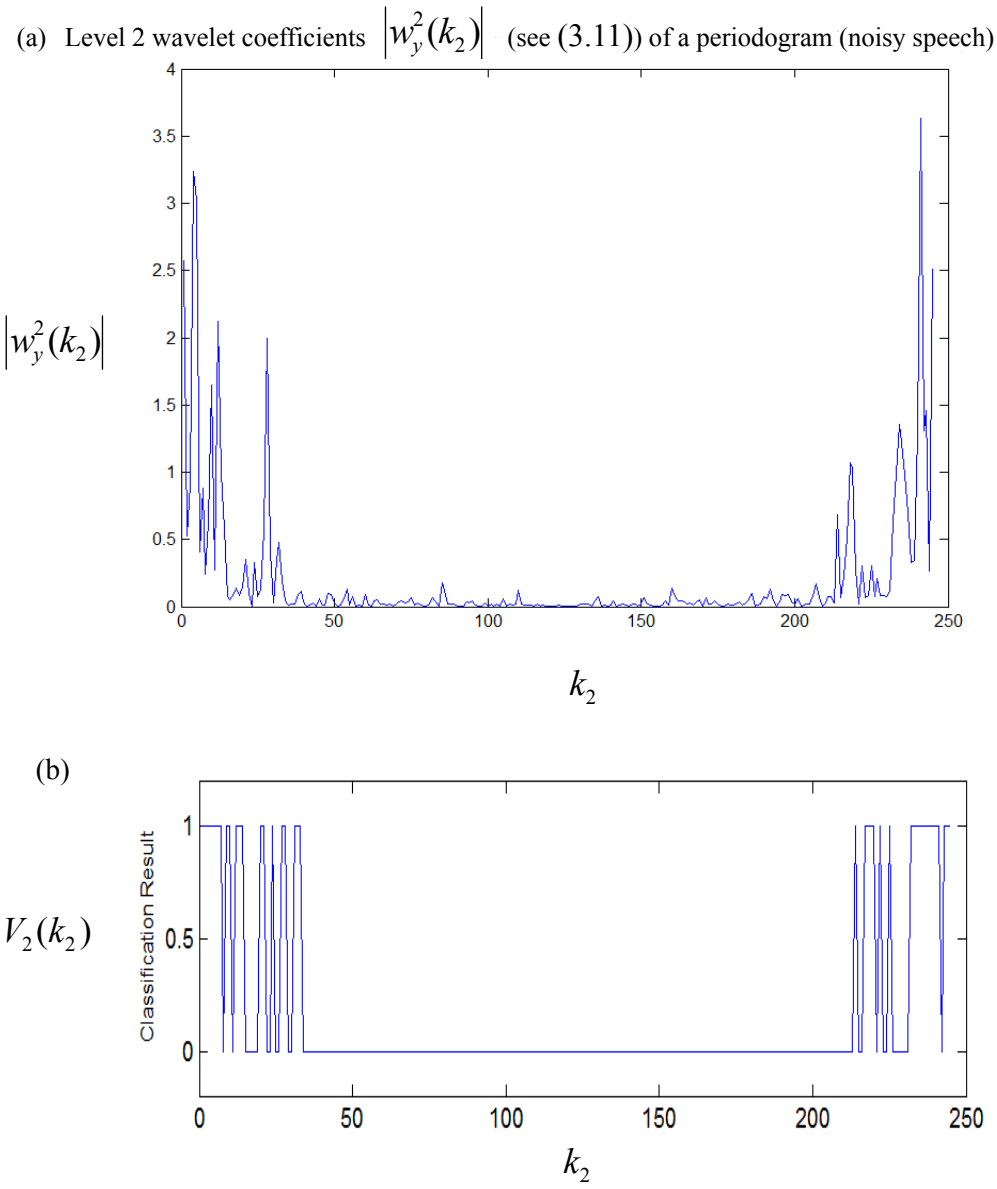


Fig. 3.3 – (a) Level 2 wavelet coefficients (absolute value) of the periodogram of a speech frame with pink noise. (b) The classification result $V_2(k_2)$ (see (3.17)).

B. Second Stage

Note that the actual denoising operation is performed in the log MTS domain. With the locality property of the wavelet transform, the oracle obtained in stage 1 can also be used for the classification of the wavelet coefficients in the log MTS domain. Let

$$w_y^{j,mt}(k_j) = W\left\{\log\left(\hat{S}_y^{mt}\right)\right\} \qquad \text{for all } j. \tag{3.18}$$

A hard thresholding procedure is applied to $w_y^{j,mt}(k_j)$ based on the oracle V_j as follows:

$$\tilde{w}_y^{j,mt}(k_j) = \begin{cases} w_y^{j,mt}(k_j) & \text{if } V_j(k_j) = 1 \\ w_y^{j,mt}(k_j) & \text{if } w_y^{j,mt}(k_j) > thr2_j \text{ and } V_j(k_j \pm \varepsilon) = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } j \tag{3.19}$$

In general if the oracle $V_j(k_j) = 1$, it indicates that the wavelet coefficient $w_y^{j,mt}(k_j)$ at k_j likely belongs to a true spectral peak. Hence, it should be kept. Besides, the wavelet coefficients $w_y^{j,mt}(k_j \pm \varepsilon)$ in the vicinity of k_j also has a good chance to belong to a true spectral peak, particularly if it has a large magnitude. For the proposed hard thresholding procedure as shown in (3.19), $w_y^{j,mt}(k_j)$ is kept if the oracle $V_j(k_j)$ equals to 1. Besides, if $w_y^{j,mt}(k_j)$ has a large magnitude that is greater than a threshold $thr2_j$, it should also be kept if in the vicinity of k_j, i.e. $k_j \pm \varepsilon$, $V_j(k_j \pm \varepsilon)$ equals to 1. The threshold $thr2_j$ is obtained using the standard SUREshrink approach [79] due to the fact that the log MTS of a speech signal in general is not white. The SUREshrink approach usually can give a more accurate threshold than the universal threshold particularly for colored noises. The limit of ε is selected to be

$$\varepsilon \le \lfloor l_w / 2 \rfloor \tag{3.20}$$

where l_w is length of the wavelet filter and $\lfloor x \rfloor$ stands for the nearest integer smaller than x. ε is selected as in (3.20) because, it can be shown that, if there is a signal change in $\hat{S}_y^{a_1}$ and $\log\left(\hat{S}_y^{mt}\right)$ at frequency index k that will lead to a strong wavelet coefficient $w_y^j(k_a)$ and

$w_y^{j,mt}(k_b)$, respectively, at level *j*, then $|k_a - k_b|$ is bounded approximately to $\lfloor l_w / 2 \rfloor$. The term 1/2 comes from the decimation operator of the wavelet transform. Consequently, the proposed algorithm in the second stage can be summarized as follows:

1. Evaluate the log MTS, i.e. $\log\left(\hat{S}_y^{mt}\right)$, of the observed noisy speech frame *y* using two orthonormal sine tapers.
2. Generate 4 levels of wavelet coefficients of the log MTS, i.e. $w_y^{j,mt}(k_j) = W\left\{\log\left(\hat{S}_y^{mt}\right)\right\}$, where $W\{\cdot\}$ is the wavelet transform and j = 1,2,…,4.
3. Compute $thr2_j$ using the standard SUREshrink approach.
4. Remove the wavelet coefficients of the noise floor based on (3.18).
5. Inverse transform the denoised wavelet coefficients to obtain a smoothed $\hat{S}_y$.

When generating the MTS, we observe in the experiment that using a large number of tapers often over-smoothes the estimated speech power spectrum and leads to low intelligibility for the resulting enhanced speeches. Hence, in step 1, we suggest using only 2 sine tapers and it usually gives better performance than using more tapers.

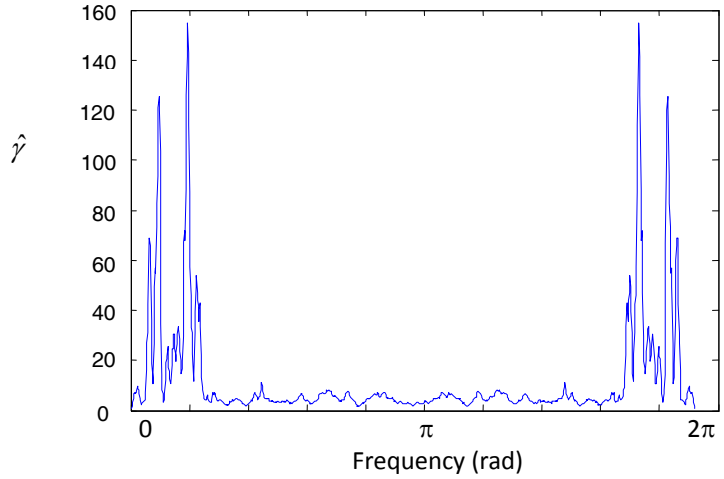
Based on the proposed two-stage wavelet denoising algorithm, the smoothed $\hat{S}_y$ is used for the evaluation of the *a-posteriori* SNR function as follows:

$$\hat{\gamma}(k) = \frac{\hat{S}_y(k)}{\hat{S}_n(k)}. \tag{3.21}$$

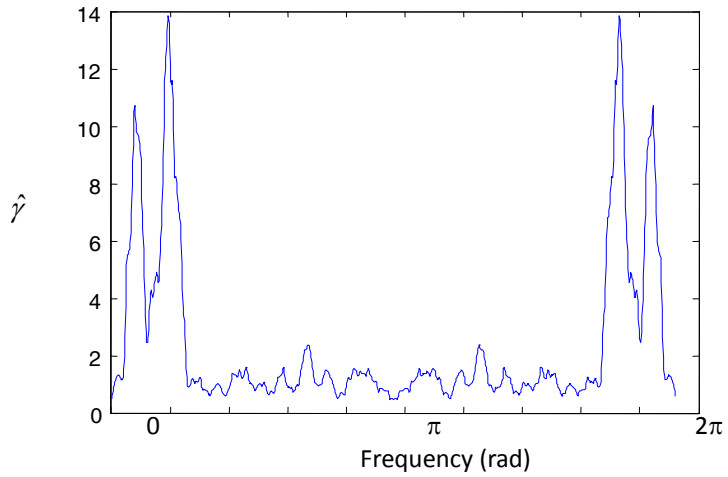
Following the same approach as in [49], $\hat{\gamma}(k)$ is further smoothed temporally by averaging $\hat{\gamma}(k)$ with those obtained in the last 4 frames. Fig. 3.4 shows a comparison of the *a-posteriori* SNR function generated from a typical noisy speech frame (pink noise, 0dB segSNR) using the proposed two-stage wavelet denoising algorithm, the local and global filtering approach used in [49], and the wavelet based MTS denoising method with universal

thresholding [147]. It can be seen that the proposed algorithm can better preserve the spectral peaks of the speech while reducing the noise variance of the noise floor. As shown in the results of using universal thresholding and [49], the spectral peaks are obviously over-smoothed. Besides, a large variance noise floor is noticed in the result of using [49].

(a) *A-posteriori* SNR $\hat{\gamma}$ generated by the proposed approach



(b) *A-posteriori* SNR $\hat{\gamma}$ generated using the smoothing approach in [49]



(c) *A-posteriori* SNR $\hat{\gamma}$ generated using the universal thresholding approach [147]

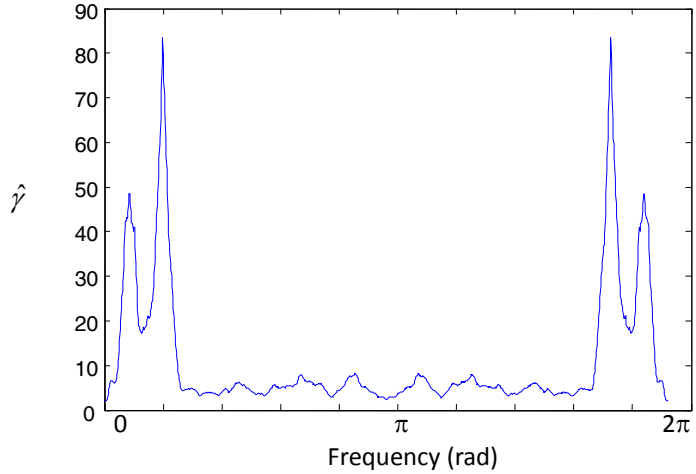


Fig. 3.4 – Performance of *a-posteriori* SNR estimation using different approaches: (a) the proposed 2-stage wavelet denoising algorithm; (b) the local and global filtering method in **[49]**; and (c) the wavelet based MTS denoising method with universal thresholding [147].

3.5 New approach for estimating GLR

The smoothed *a-posteriori* SNR is then used for the estimation of the SPP. Recall that the SPP can be estimated based on the GLR, which is a weighted ratio between $p(\gamma \mid H_1)$ – the PDF of γ under the hypothesis H_1, i.e. speech is present; and $p(\gamma \mid H_0)$ – the PDF of γ under the hypothesis H_0, i.e. speech is absent, as given in (3.1). In [49], $p(\gamma \mid H_0)$ is approximated to be chi-squared distributed with the degree of freedom depending on the order of the temporal and spectral smoothing operators. With the introduction of the proposed wavelet based smoothing operator, $\hat{\gamma}$ will have a value close to 1 if the current frame is known to be a noise frame. And although the PDF of $\hat{\gamma}$ can still be approximated as chi-squared distributed as in [49], the degree of freedom will be very difficult to estimate due to the non-linear thresholding operation in the wavelet domain. Here we propose to generate $p(\hat{\gamma} \mid H_0)$ by directly computing the histogram of $\hat{\gamma}$ in noise frames. To be specific, we first make use of a VAD given by [154] to identify the noise frames. For each noise frame, we compute $\hat{\gamma}$ by using (3.21) and bin the resulting $\hat{\gamma}$ at different frequencies into 200 equally spaced containers with centres ranged from 0.2 to 40. The number of data in each container is recorded and finally the histogram of $\hat{\gamma}$ is obtained. It is then normalized with the total number of data and the bin size, i.e. 0.2, to serve as an approximation of the PDF. Fig. 3.5 shows a typical $p(\hat{\gamma} \mid H_0)$ under pink noise contamination. It can be seen that it is seldom to have large $\hat{\gamma}$ with value greater than 5 since the proposed two-stage wavelet denoising algorithm has effectively removed the outliers.

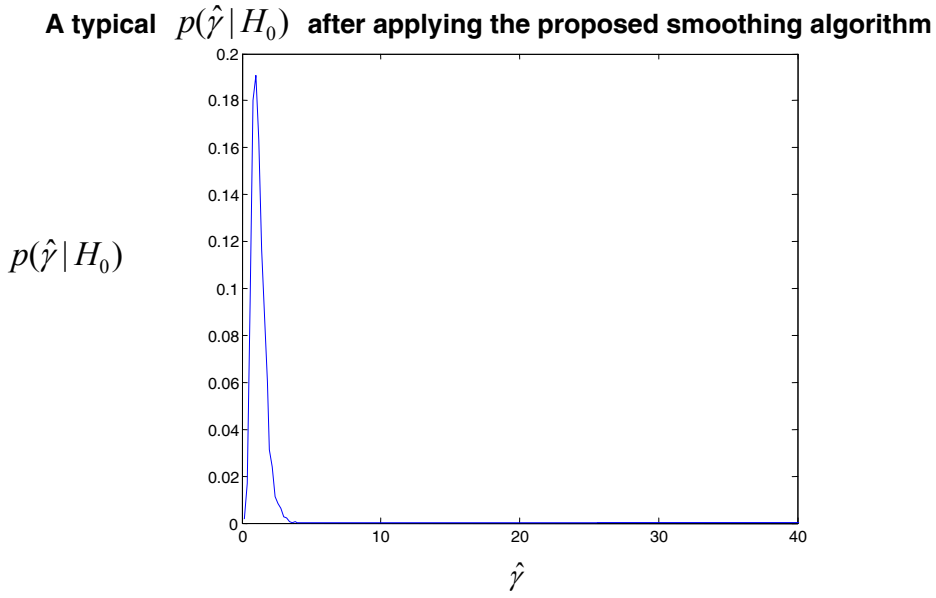


Fig. 3.5 – A typical $p(\hat{\gamma}\,|\,H_0)$ after using the proposed 2-stage wavelet denoising algorithm.

It is known that when the observation *y(n)* is stationary with a relatively small span of correlation and the frame size is large, the real and imaginary part of a DFT coefficient ***Y(k)*** can be considered to be independent and can be modeled as zero mean Gaussian random variables. Consequently, the PDF of $|Y(k)|^2$, which is the sum of the square of the real and imaginary parts of ***Y(k)*** can be described by a gamma distribution as follows:

$$|Y(k)|^2 = \hat{S}_y \sim \Gamma(1, 2\sigma_y^2) \tag{3.22}$$

As to the estimation of $p(\hat{\gamma}\,|\,H_1)$, it should be noted that the proposed two-stage wavelet denoising algorithm mainly smoothens the noise floor using a hard thresholding approach. It basically does not make any modification to the wavelet coefficients which do not belong to the noise floor. Thus, the spectral peaks in the resulting power spectrum (after antilog) are just the original MTS, which is an average of two periodograms generated using two orthonormal tapers. Hence

$$\hat{S}_y^{mt} = (\hat{S}_y^a + \hat{S}_y^b)/2 \sim \Gamma(2, \sigma_y^2), \tag{3.23}$$

And then since the proposed approach adopts the averaging scheme as in [49], $\hat{S}_y^{mt}$ is further

averaged with that of the previous 5 frames. That is,

$$\hat{S}_y^{amt}(i) = \frac{1}{5}\sum_{n=-4}^{0} \hat{S}_y^{mt}(i) \sim \Gamma\left(2*5*c_{dof}, \sigma_y^2/\left(5*c_{dof}\right)\right) \tag{3.24}$$

for $i > 4$. The term c_{dof} is an adjustment factor since correlation exists between adjacent frames. It is mainly due to the overlapping of speech samples when windowing the speech. In our simulation, an overlapping factor of 0.75 is adopted. Hence we choose c_{dof} to be 0.25. Consequently,

$$\hat{\gamma} = \frac{\hat{S}_y^{amt}}{\sigma_n^2} \sim \Gamma\left(\frac{\bar{r}}{2}, \frac{4}{\bar{r}}\frac{\sigma_y^2}{\sigma_n^2}\right) = \Gamma\left(\frac{\bar{r}}{2}, \frac{4}{\bar{r}}\left(1+\bar{\xi}\right)\right) \tag{3.25}$$

where $\bar{r}$ = 2*2*5*c_{dof}, and $\bar{\xi}$ is the *a-priori* SNR. Hence

$$p\left(\hat{\gamma} \mid H_1\right) = \left(\frac{\bar{r}/2}{2\left(1+\bar{\xi}\right)}\right)^{\frac{\bar{r}}{2}} \frac{\hat{\gamma}^{\frac{\bar{r}}{2}-1}}{\Gamma\left(\frac{\bar{r}}{2}\right)} \exp\left(-\frac{\hat{\gamma}\bar{r}/2}{2\left(1+\bar{\xi}\right)}\right), \tag{3.26}$$

Similar to [49], $\bar{\xi}$ is selected to be a fixed constant 8dB. Consequently, the GLR can be estimated as follows:

$$\tilde{\Lambda} = \frac{q}{(1-q)} \frac{p(\hat{\gamma} \mid H_1)}{p(\hat{\gamma} \mid H_0)}, \tag{3.27}$$

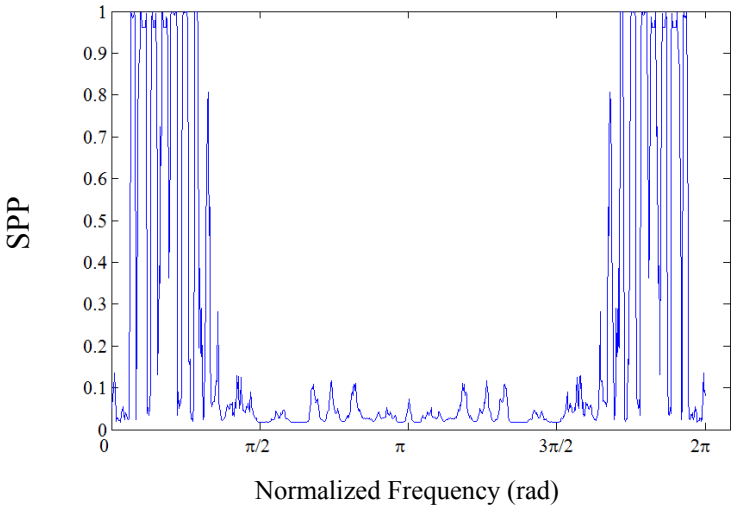
where q is set as 0.5 as in [49] in our experiments. And the SPP can be obtained based on the estimated GLR $\tilde{\Lambda}$ as follows:

$$\overline{spp} = \frac{\tilde{\Lambda}}{1+\tilde{\Lambda}}. \tag{3.28}$$

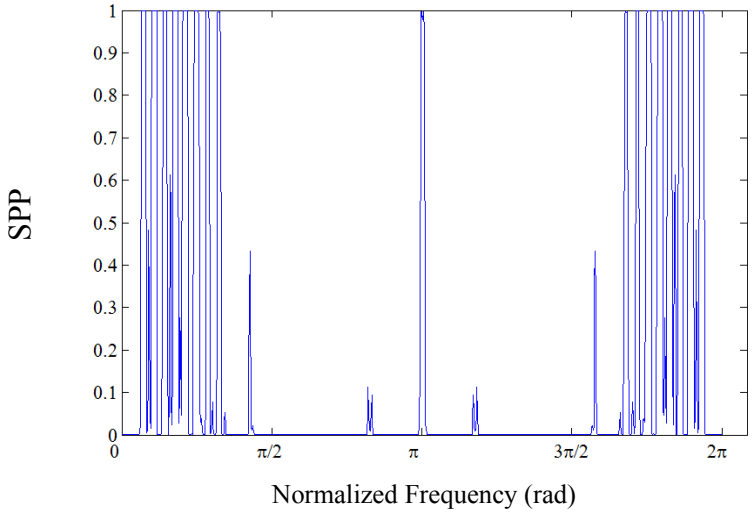
The resulting SPP is then applied to (2.25) to obtain a new gain function for enhancing the noisy speeches. As an illustration, Fig. 3.6a, Fig. 3.6b and Fig. 3.6c show the SPP estimated using the proposed 2-stage wavelet denoising algorithm, the local and global filtering method in [49], and the wavelet based MTS denoising method with universal thresholding [147], respectively, for a typical noisy speech frame. Note that the GLR used for these approaches is

different. For the proposed approach, the GLR as given in (3.27) is used. For the local and global filtering method, the GLR suggested in [49] is used. As to the one using the universal thresholding, essentially the same GLR as the proposed algorithm is used. However, the parameter $\bar{r}$ for computing $p(\hat{\gamma} \mid H_1)$ is increased to 2*5*5*c_{dof} since 5 tapers are used in [147] for generating the MTS. Besides, c_{dof} needs to be empirically adjusted due to the shrinkage operation applied to the wavelet coefficients. As can be seen in Fig. 3.6b, the SPP given by [49] can contain large spurious impulses (such as that near $\pi/2$). They often contribute to the musical noise in the final enhanced speech. On the other hand, the SPP estimator using the universal thresholding method, as shown in Fig. 3.6c, can merge all spectral peaks of the speech. It will certainly affect the final speech quality. Since the proposed approach can accurately estimate the spectral locations of the noise floor, the resulting SPP can largely preserve the spectral peaks while achieving a good control of spurious impulses on the noise floor.

(a) SPP generated using the proposed smoothing algorithm



(b) SPP generated using the smoothing algorithm **[49]**



(c) SPP generated using wavelet based MTS smoothing approach using universal thresholding **[147]**

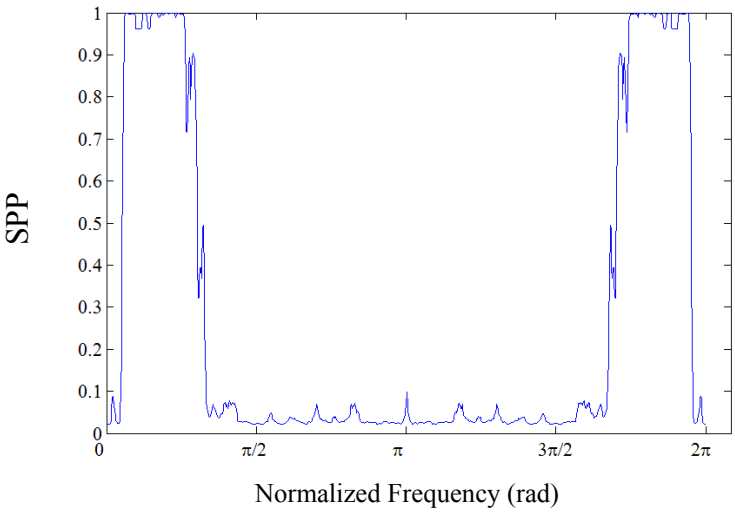


Fig. 3.6 – SPP estimated using: (a) the proposed 2-stage wavelet denoising algorithm; (b) the local and global filtering method in **[49]**; and (c) the wavelet based MTS denoising method with universal thresholding **[147]**.

3.6 Simulations and results

The best approach for evaluating the performance of the new SPP estimator is by directly assessing the quality of the resulting speeches when it is applied to some traditional speech enhancement algorithms, such as the MMSE-LSA. In this study, a series of simulations have been performed for comparing the performance between different SPP estimators. Table **3.1** gives a summary of the algorithms that have been compared. They are different only in the way the SPP is estimated. For instance, LSA+SPP uses the traditional SPP estimator given by Cohen [47]. LSA+FPSPP uses the SPP estimator given in [49]. The LSA+2sSPP use the proposed 2-stage wavelet denoising algorithm. Other simulation details are listed as follows:

- Speech sampling rate: 16 kHz
- Frame size: 480 samples (~30ms)
- FFT size: 960 samples (zeros padded each frame with 480 samples)
- Window shift step size: 120 samples (75% overlap rate)
- Wavelet function used in the computation of the wavelet transform: "sym4" – order-4 least asymmetric orthogonal wavelet [159].

For all algorithms, the noise power spectrum is estimated following the same approach described in [1]. The noise power spectrum is estimated by first using the initial frames that are assumed to have no speech energy. It is then updated whenever a frame is detected to have no speech energy by using a VAD, such as [155]-[156]. Note that the accuracy of the VAD can affect the performance of the enhanced speech signals. However, it will barely affect the comparison results since the same VAD is used in all compared algorithms.

Table 3.1 - Summary of the algorithms compared in the simulations.

Methods	Description
LSA+SPP	MMSE-LSA [15] with Cohen's SPP [47]
LSA+FPSPP	MMSE-LSA [15] with SPP estimated using the local and global filtering method for smoothing the *a-posteriori* SNR [49]
LSA+2sSPP	MMSE-LSA [15] with SPP estimated using the proposed 2-stage wavelet denoising algorithm for smoothing the *a-posteriori* SNR

In the simulation, we arbitrarily selected 40 male and 40 female test speeches from the TIMIT database [192]. White noise and colored (pink) noises adopted from the NOISEX-92 database [193] were added to the speeches at different input segSNR. We use the three composite objective measures [1] as shown in Table 3.2 as the performance/design criteria in the simulation. The three composite objective measures are used traditionally to predict the quality of noisy speech enhanced by noise suppression algorithms. They are obtained by linearly combining existing objective measures as follows: (a) C_{sig} for measuring signal distortion (SIG) – it is formed by linearly combining the log-likelihood ratio (LLR), PESQ, and weighted-slope spectral distance (WSS) measures; (b) C_{bak} for measuring noise distortion (BAK) – it is formed by linearly combining the segSNR, PESQ, and WSS measures; and (c) C_{ovl} for measuring overall quality (OVL) – it is formed by linearly combining the PESQ, LLR, and WSS measures. The definition of all these measures can be found in [1]. Table 3.2 shows that the proposed LSA+2sSPP often outperforms the other two algorithms particularly when the noise level is high. Specifically, while LSA+FPSPP gives an average increase of 0.58, 0.61 and 0.57 in C_{sig}, C_{bak}, and C_{ovl} respectively compared with the noisy signal (pink noise at different input SNRs), the proposed LSA+2sSPP gives an average increase of 0.67, 0.66 and 0.64, which account to 15.5%, 8.2% and 12.2% improvement respectively over the LSA+FPSPP algorithm.

To verify the improvement of the 2-step wavelet denoising method over the traditional universal thresholding method, we applied both approaches to the proposed SPP estimation procedure. We compared the PESQ scores (the Perceptual Evaluation of Speech Quality) of

the enhanced speeches generated by both methods. PESQ is an ITU standard for evaluating speech quality [58]. The same set of speeches from the TIMIT database as mentioned above were used in the simulation. As shown in the comparison result in Fig. 3.7, the enhanced speech using the proposed SPP estimator with the 2-step wavelet denoising method has a much higher PESQ score. While both approaches are able to suppress musical and other residue noises, the SPP generated by using the universal thresholding method is over-smoothed that cannot resolve the spectral harmonics and leads to degraded speeches. The result in Fig. 3.7 has verified this observation.

Table 3.2 - Composite measurement comparison of LSA+SPP, LSA+FPSPP and the proposed LSA+2sSPP.

	Noise	Method	Input SNR										
			-5	-4	-3	-2	-1	0	1	2	3	4	5
Csig	White	Noisy	1.089	1.114	1.154	1.206	1.265	1.34	1.432	1.537	1.654	1.782	1.917
		LSA+SPP [47]	1.412	1.547	1.687	1.841	1.995	2.153	2.303	2.449	2.587	2.721	2.849
		LSA+FPSPP [49]	1.565	1.703	1.847	1.999	2.145	2.280	2.409	2.535	2.651	2.766	2.868
		Proposed LSA+2sSPP	**1.636**	**1.777**	**1.919**	**2.065**	**2.203**	**2.337**	**2.468**	**2.593**	**2.710**	**2.823**	**2.923**
	Pink	Noisy	1.322	1.396	1.481	1.578	1.688	1.809	1.935	2.063	2.193	2.323	2.454
		LSA+SPP [47]	1.811	1.946	2.080	2.221	2.357	2.489	2.622	**2.746**	**2.869**	**2.988**	**3.097**
		LSA+FPSPP [49]	1.843	1.962	2.088	2.212	2.332	2.443	2.556	2.658	2.751	2.844	2.928
		Proposed LSA+2sSPP	**1.955**	**2.072**	**2.190**	**2.306**	**2.420**	**2.531**	**2.639**	2.739	2.831	2.919	2.996
Cbak	White	Noisy	1.332	1.401	1.474	1.552	1.633	1.716	1.802	1.890	1.980	2.072	2.165
		LSA+SPP [47]	1.978	2.064	2.148	2.236	2.321	2.407	2.493	2.578	2.663	2.746	2.825
		LSA+FPSPP [49]	2.045	2.124	2.205	2.288	2.371	2.451	2.531	2.611	2.688	2.766	2.838
		Proposed LSA+2sSPP	**2.089**	**2.172**	**2.253**	**2.336**	**2.417**	**2.498**	**2.578**	**2.659**	**2.737**	**2.815**	**2.887**
	Pink	Noisy	1.283	1.332	1.389	1.453	1.528	1.609	1.696	1.788	1.882	1.978	2.075
		LSA+SPP [47]	1.844	1.919	1.995	2.077	2.160	2.244	2.332	2.419	2.507	2.597	2.683
		LSA+FPSPP [49]	1.859	1.927	2.005	2.083	2.163	2.242	2.326	2.408	2.489	2.572	2.655
		Proposed LSA+2sSPP	**1.894**	**1.967**	**2.045**	**2.125**	**2.208**	**2.290**	**2.376**	**2.458**	**2.539**	**2.623**	**2.703**
Covl	White	Noisy	1.079	1.109	1.151	1.206	1.271	1.345	1.432	1.535	1.642	1.754	1.866
		LSA+SPP [47]	1.423	1.552	1.683	1.816	1.943	2.071	2.195	2.316	2.432	2.544	2.651
		LSA+FPSPP [49]	1.559	1.682	1.805	1.931	2.051	2.164	2.272	2.378	2.476	2.573	2.658
		Proposed LSA+2sSPP	**1.616**	**1.742**	**1.864**	**1.986**	**2.101**	**2.214**	**2.323**	**2.430**	**2.529**	**2.626**	**2.712**
	Pink	Noisy	1.171	1.226	1.289	1.367	1.458	1.558	1.665	1.775	1.886	1.997	2.108
		LSA+SPP [47]	1.627	1.740	1.851	1.969	2.085	2.197	2.312	2.421	**2.530**	**2.637**	**2.737**
		LSA+FPSPP [49]	1.649	1.751	1.860	1.967	2.072	2.171	2.273	2.368	2.457	2.545	2.626
		Proposed LSA+2sSPP	**1.737**	**1.837**	**1.941**	**2.044**	**2.147**	**2.248**	**2.349**	**2.442**	2.529	2.615	2.692

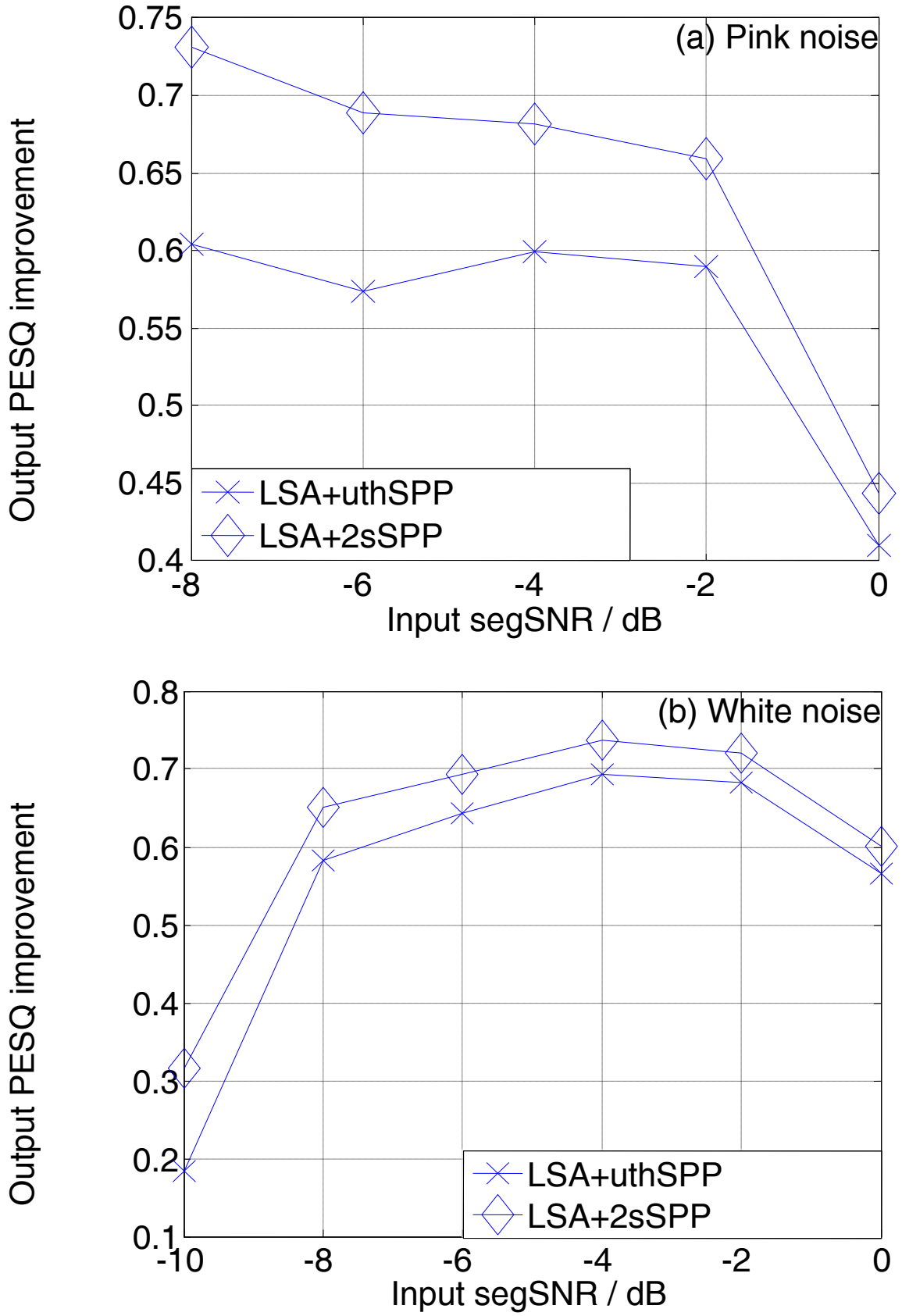


Fig. 3.7 – Comparison of using the 2-step wavelet denoising (LSA+2sSPP) and universal thresholding (LSA+uthSPP) in terms of PESQ improvement for the cases of (a) pink and (b) white noise contamination.

As in [49], the missed-hit rate and false-alarm rate of SPP could be evaluated by the Speech Distortion (SD) and Noise Leakage (NL) measurement, respectively. In Table 3.3, the SD and NL are measured to indicate the percentage of the speech energy that the corresponding SPP estimator neglects as well as how much energy from the noise-only bins is not attenuated.It can be clearly seen that the SD for the proposed LSA+2sSPP method reduce the speech distortion as compared to LSA+FPSPP. Moreover, both the proposed LSA+2sSPP and LSA+FPSPP methods can provide low noise leakage, but the proposed method yields a better tradeoff than LSA+FPSPP.

Table 3.3 - Speech Distortion (SD) and Noise Leakage (NL) measurement comparison of LSA+FPSPP and the proposed LSA+2sSPP for the cases of white noise contamination.

	Method	Input SNR (dB)				
		-10	**-5**	**0**	**5**	**10**
SD (%)	**LSA+FPSPP [49]**	29.6	13.8	5.6	2.1	0.7
	Proposed LSA+2sSPP	22.2	10.1	4.4	1.8	0.7
NL (%)	**LSA+FPSPP [49]**	1.2	1.4	2.0	3.0	5.1
	Proposed LSA+2sSPP	1.3	1.5	1.8	2.6	4.2

Fig. 3.8 shows a comparison of the spectrogram of the enhanced speeches generated using different algorithms. The speech is added with color (pink) noise at input segSNR 0dB. It can be seen that the proposed algorithm in general preserves much better the speech contents while effectively removing the background noise. We particularly circle the parts in the spectrograms where improvement can easily be seen. For LSA+SPP, it can be seen in region A to C that the speech spectrum is over-smoothed such that many of the speech contents are removed. Particularly, the distortion in region B has introduced some difficulty in understanding the enhanced speech since the spectrum in region B refers to a word in the sentence. For LSA+FPSPP, musical noises are noticed in region D mainly due to the errors in the SPP estimation. Besides, at low frequencies where SNR is low (pink noise has higher

noise level at low frequencies), speech distortion is also noticed (such as in region C). Comparing with the other two algorithms, the proposed LSA-2sSPP removes musical noise without sacrificing speech quality. The enhanced spectrogram has much improvement over those generated by the other algorithms as can be seen in region A to D. The result in Fig. 3.8 conforms to the objective comparison results as shown in **Table 3.2**.

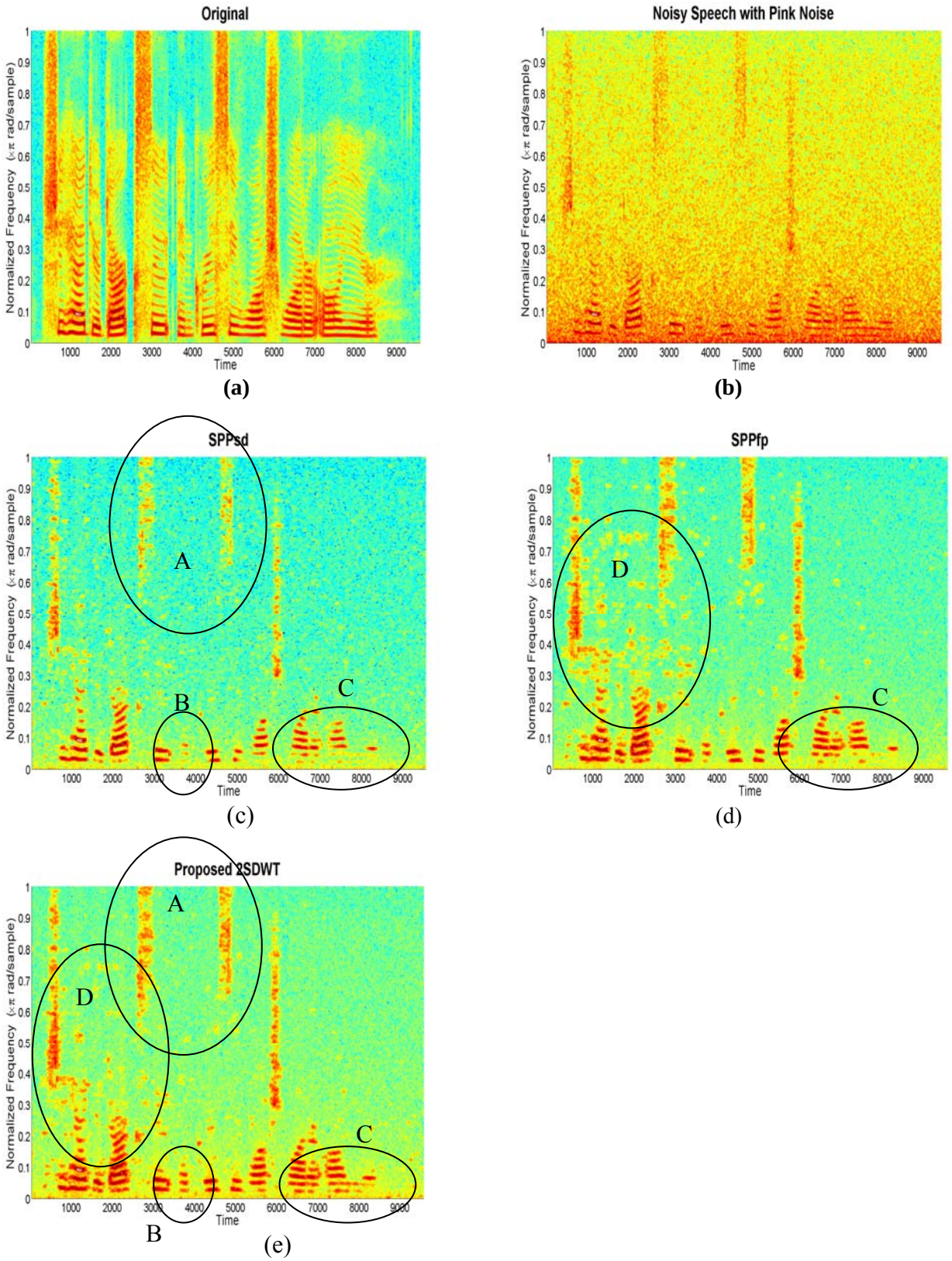


Fig. 3.8 – Spectrogram of (a) a speech selected from TIMIT database; (b) speech contaminated by color (pink) noise at input segSNR 0dB; and enhanced speech using (c) LSA+SPP, (d) LSA+FPSPP, and (e) the proposed LSA+2sSPP algorithm.

3.7 Chapter Summary

In this chapter, we proposed a new algorithm for the estimation of the speech presence probability (SPP) of a noisy speech signal. Although it is known that a good estimator of SPP can be obtained by smoothing the observed noisy speech power spectrum before using it in the estimation process, care must be taken to ensure the smoothing operation will not wash away the spectral peaks which are important to the intelligibility of the enhanced speech. The major contribution of this work is two-folded. First, we successfully developed a two-stage wavelet denoising algorithm that effectively removes the noise while preserving the spectral peaks in a noisy speech power spectrum. It outperforms the traditional approaches by combining the information of noise and spectral peaks in both the periodogram and the log MTS of a noisy speech. The denoised speech power spectrum in turn lets us generate a smooth *a-posteriori* SNR function. Second, we proposed a new method for estimating the generalized likelihood ratio (GLR). It is by directly estimating the PDF of the *a-posteriori* SNR under the hypothesis H_0, i.e. speech is absent, using the data in different noise frames. It simplifies the estimation process and avoids the use of many empirically selected parameters in the traditional approaches. The new SPP estimator was then applied to the MMSE-LSA speech enhancement algorithm. Compared with the traditional SPP estimators, up to 15% improvement was noted for different noises at different noise levels when measuring using the standard composite objective measures. When inspecting the spectrogram of the enhanced speeches using different approaches, the proposed algorithm in general preserves much better the speech contents while effectively removing the background noise.

The proposed algorithm enhances a noisy speech by making use of the discrete wavelet transform to successfully detect the speech's spectral peaks. However, problems may arise for certain kinds of noise which also have spectral peaks similar to those of speeches. This can make the SPP thus estimated erroneous. More effort is needed to differentiate speeches from

noises perhaps in a different domain.

Chapter 4 A Novel Expectation-Maximization Framework for Speech Enhancement in Non-stationary Noise Environments

4.1 Introduction

In Chapter 3, we have discussed a new method for estimating the SPP and its application to speech enhancement. We have shown that by using the wavelet techniques (which is a kind of sparse representation techniques), we can improve the estimation of the SPP so that it can be applied to assist the speech enhancement gain function, such as the MMSE-LSA or the Wiener filter, etc., to improve the suppression of the spectral components with no speech information. While the SPP is an important parameter in speech enhancement, the importance of the *a-priori* SNR is by no means inferior since it is the major parameter of almost all gain functions. The accuracy of its estimation can significantly affect the performance of a speech enhancement algorithm. The *a-priori* SNR is defined as the ratio between the true power spectra of speech and noise. While the estimation of the true speech power spectrum is known to be difficult, the estimation of the true noise power spectrum is not easy either. This is particularly the case when the contaminating noise signal is non-stationary. It ends up with the musical noises [1] introduced to the resulting enhanced speech, which is extremely annoying to human listeners.

To improve the estimation of the *a-priori* SNR, the temporal cepstrum smoothing (TCS) technique was recently proposed [115][116]. Since voiced speeches are quasi-periodic in nature, their magnitude spectrum exhibits peaks and valleys separated by harmonics of the fundamental frequency which can be compactly represented in the cepstral domain. As most noises do not have such harmonic structure, it allows us to selectively reduce the variance of the cepstral coefficients which are likely contributed by noise. In general, the TCS method

can improve the accuracy in estimating the *a-priori* SNR of a noisy speech comparing with the traditional spectral subtraction methods. It was also reported that the method works well in some non-stationary noise environments [116]. Nevertheless, the TCS method requires a set of empirically selected parameters to control the cepstrum smoothing process. As there is not a mechanism to automatically adjust the parameters, the TCS method cannot adapt itself to the changes in SNR of the noisy speech signal across time frames or frequency components. The problem is particularly obvious if there are some parts of a noisy speech spectrum having significantly low SNR (e.g. the noise is composed by a few strong tones of varying frequencies). The TCS method cannot fully remove the related cepstral coefficients with the speech content intact. To deal with the problem, it was suggested to further apply a SPP estimator with the TCS technique to remove the outliers in the enhanced speech [140]. The result however is still not very satisfactory since the accuracy of the SPP estimators will also deteriorate when the SNR is low or when the noise is non-stationary as mentioned above.

In this chapter, we present an improved speech enhancement algorithm based on a novel expectation-maximization (EM) framework working in the cepstral domain. The EM algorithm was discovered and employed independently by several different researchers until Dempster [161] brought their ideas together and coined the term EM algorithm. It is particularly suitable to the parameter estimation problems in which the data for evaluating the parameters are missing or incomplete. It is known to produce the maximum-likelihood (ML) parameter estimates when there is a many-to-one mapping from an underlying distribution to the distribution governing the observation. The algorithm contains two main steps. The E-step (expectation) gives an expectation of the unknown underlying distribution based on the observed data and the M-step (maximization) estimates the parameters by maximizing the expectation. The E-step and M-step then iterate alternately until converged. The EM algorithm has widespread applications in digital image and speech processing [162]-[171].

One of the widely cited applications of the EM algorithm is the estimation of the hidden Markov models (HMMs), which is particularly relevant to speech processing [166]-[169]. In [168], a HMM-based gain modeling algorithm was proposed for the enhancement of speech in noise. It applies the EM algorithm for offline training and the recursive EM algorithm for online estimation of the HMM parameters. The EM algorithm is also used in an approximate Bayesian based speech enhancement algorithm [169] for learning the speech and noise spectra under the Gaussian approximation. Similar to the conventional model based speech enhancement methods, these approaches require prior knowledge about the noise model, or it has to be detected online. Degraded performance will be resulted if there is error in detection or the detected noise model is not in the training database. Besides HMM models estimation, the EM algorithm is also used in the estimation of autoregressive (AR) model for speech enhancement [170], where the E-step is in fact the Kalman filter and the M-step is similar to the standard Yule-Walker solution for estimating the coefficients of AR processes. It is noted that the performance of the method is rather unstable (particularly at input SNR from 4dB to 10dB). Effort was made [171] to improve the problem by using the Rao-Blackwellized particle filter in the E-step to replace the Kalman filter. However, the overall performance, particularly at low SNR, still has much room to improve. In general, the performance of model based speech enhancement methods depends heavily on the accuracy in model estimation, which is often a challenge when they are working in open environments. There are many other applications of the EM algorithm. More details can be found in [172][173].

Similar to the abovementioned approaches, the proposed algorithm makes use of the EM algorithm to define a theoretical framework for the design of an iterative speech enhancement process. However, it is non-parametric, hence it does not require specific prior knowledge about the speech or noise model. In the proposed algorithm, the parameters to be estimated

are the cepstral coefficients of the true speech power spectrum, of which their accurate estimation is important in speech enhancement. It enables the computation of the *a-priori* SNR of the noisy speech, which is one of the most essential parameters required in different speech enhancement gain functions as mentioned above. The proposed algorithm first makes use of the TCS technique to generate an initial guess of the clean speech periodogram, which is the complete data set of our problem. It is applied to an L_1 norm regularizer [175] in the M-step of our EM framework to give the first estimate of the required cepstral coefficients of the true speech power spectrum. They are then used to compute the *a-priori* SNR that is needed for the MMSE-LSA gain function to refine the estimation of the clean speech periodogram, which is the E-step of our EM framework. Subsequently, the estimate is fed back to the M-step to refine the estimation of the cepstral coefficients of the true speech power spectrum. The E-step and M-step iterate alternately until convergence is reached. The operation is illustrated in Fig. 4.1. A notable improvement of the proposed algorithm over the traditional non-parametric speech enhancement methods is that, due to the iterative process, the proposed algorithm can adapt to the changes (even abrupt changes) in SNR of the noisy speech. In addition, the proposed algorithm fully utilizes the sparsity of speeches in the cepstral domain by adopting an L_1 norm regularizer in the M-step. It enables the regularization process to be carried out on coefficients with improved SNR hence reduces the effect due to the error in estimating the non-stationary noise statistical characteristics. As a result, the proposed algorithm has outstanding performance when working in non-stationary noise environments. Extensive performance evaluations have been conducted using the speech samples from the TIMIT database [192] contaminated by many different noise signals. Significant improvement is noted in almost all cases over the competing speech enhancement methods measured using standard performance metrics.

This chapter is organized as follows. In Section 4.2, a brief review of the traditional

temporal cepstrum smoothing algorithm is given. It is followed by a brief introduction of the EM algorithm in Section 4.3. The new EM framework for speech enhancement in non-stationary environments is described in Section 4.4. The simulation results are shown in Section 4.5, and conclusions are drawn in Section 4.6.

The results in the chapter have also been reported in [181] and [182].

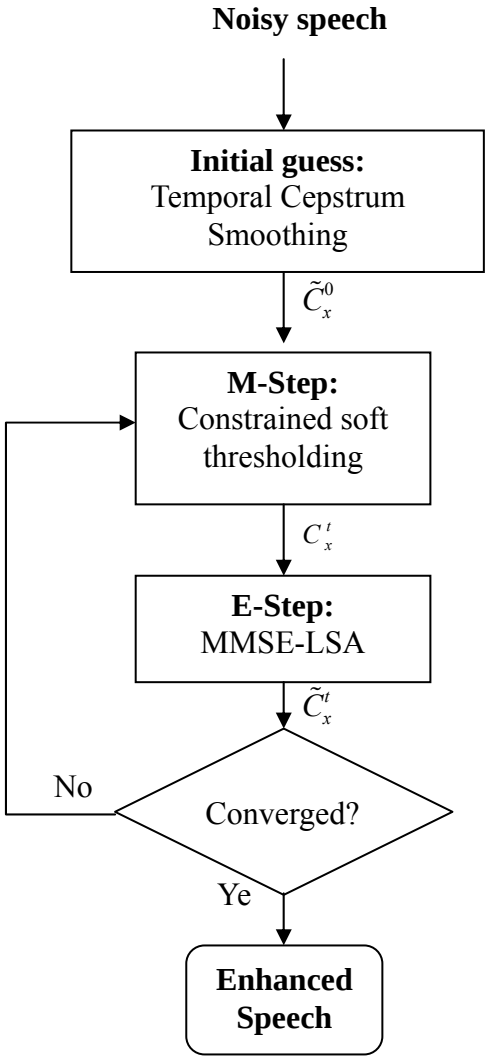


Fig. 4.1 – The operation of the proposed speech enhancement algorithm based on the new EM framework

4.2 Spectral Subtraction and Cepstrum Smoothing

Let us begin with a brief introduction of the traditional cepstrum smoothing method that is applied to the spectral subtraction speech enhancement algorithms. In fact, we have introduced the essence of the spectral subtraction techniques in Chapter 2. The spectral subtraction methods are still popularly used in speech enhancement due to their simplicity and efficiency. For the spectral subtraction methods, a spectral gain function $G(k,i)$ is applied to each noisy short-time frame $Y(k,i)$ to enhance the speech signal. Two most popular gain functions are the Wiener filter and the MMSE-LSA gain functions as follows [1]:

$$G_{wiener}(k) = \frac{\xi(k)}{\xi(k)\ +1} \tag{4.1}$$

$$G_{\log mmse}(k) = \frac{\xi(k)}{\xi(k)+1}\exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\} \quad \text{where} \quad v(k) = \frac{\xi(k)\gamma(k)}{\xi(k)+1} \tag{4.2}$$

It is seen in (4.1) and (4.2) that the determination of both gain functions requires the evaluation of two parameters: (i) *a-posteriori* SNR $\gamma(k)$ which is defined as $\gamma(k) = \frac{|Y(k)|^2}{S_n(k)}$, where $|Y(k)|^2$ is also referred as the periodogram of *y*; (ii) *a-priori* SNR $\xi(k)$ which is defined as $\xi(k) = \frac{S_x(k)}{S_n(k)}$. In practice, the power spectrum of noise $S_n(k) = E\left(|N(k)|^2\right)$ is estimated by averaging the periodograms of all the noise frames detected using a voice activity detector (VAD) [53]. We denote the estimation of S_n as $\hat{S}_n$. Obviously, the estimation error can be high if *n* is not stationary.

The estimation of $S_x(k) = E\left(|X(k)|^2\right)$ is even more difficult since *x* is not known. Hence $\xi(k)$ cannot be exactly evaluated. Different approaches were suggested to estimate

$\xi(k)$. The maximum likelihood (ML) estimate of the *a-priori* SNR $\hat{\xi}^{ML}(k)$ can be obtained as follows [1]:

$$\hat{\xi}^{ML}(k) = \hat{\gamma}(k) - 1, \text{ where } \hat{\gamma}(k) = \frac{|Y(k)|^2}{\hat{S}_n(k)} \tag{4.3}$$

The variance of $\hat{\xi}^{ML}(k)$ however is often too large for using in the traditional gain functions. The inaccuracy of $\hat{\xi}^{ML}(k)$ is further amplified due to the estimation error of $\hat{S}_n$. To reduce the variance in estimation, the decision-direct approach is often used in practice where the *a-priori* SNR is estimated based on a previous clean-speech estimate as follows [14]:

$$\hat{\xi}^{DD}(k,i) = \alpha \frac{\hat{S}_x(k,i-1)}{\hat{S}_n(k,i)} + (1-\alpha)\max\{\hat{\gamma}(k,i) - 1, \xi_{\min}\} \tag{4.4}$$

where the parameters α and $\xi_{\min}$ control the trade-off between the amount of noise reduction and the distortion of speech transients in a speech enhancement framework. Although much effort has been devoted to resolve the difficulties, in general the performance of current spectral subtraction algorithms will still degrade significantly when the SNR is low or if the noise is non-stationary.

It is shown in [116] that TCS method can give a good estimation of the *a-priori* SNR for some non-stationary noise environments. The algorithm can be implemented by first computing the ML estimation of the *a-priori* clean speech power spectrum as follows:

$$\hat{S}_x^{ML}(k,i) = \hat{S}_n(k,i)\max\left\{\hat{\xi}^{ML}(k), \hat{\xi}_{\min}^{ML}\right\} \tag{4.5}$$

where $\hat{\xi}_{\min}^{ML}$ is the minimum value allowed for $\hat{\xi}^{\text{ML}}$ and $\hat{\xi}^{\text{ML}}$ is the maximum-likelihood estimate of the *a-priori* SNR given by (4.3). Next, the cepstral representation of $\hat{S}_x^{ML}(k,i)$ is computed as,

$$\widehat{C_x}(q) = IDFT\{log(\hat{S}_x^{ML}(k))\} \tag{4.6}$$

where q is the cepstral index, also known as the quefrency index. Next, the selected cepstral

coefficients are recursively smoothed over time with a quefrency dependent parameter $\alpha(q)$ as follows:

$$\tilde{C}_x^{TCS}(q,i) = \alpha_{TCS}(q)\tilde{C}_x^{TCS}(q,i-1) + (1-\alpha_{TCS}(q))\hat{C}_x(q,i) \tag{4.7}$$

The parameter $\alpha(q)$ is also smoothed recursively using:

$$\alpha_{TCS}(q,i) = \begin{cases} \alpha_{pitch} & if\ q \in C_{pitch}(i) \\ \beta\alpha_{TCS}(q,i-1) + (1-\beta)\alpha_q^{const} & otherwise \end{cases} \tag{4.8}$$

where C_{pitch} refers to the set of cepstral bin indices associated with the fundamental frequency. All parameters including α_{pitch}, β and α_q^{const} for different q need to be determined empirically. It is believed that these parameters should be adaptively adjusted in order to achieve optimal performance for noisy speeches of different SNR values. It is however difficult to derive an efficient algorithm for such purpose due to the empirical nature of these parameters.

4.3 The Expectation Maximization Algorithm

In this section, the basic idea behind the EM algorithm is described. Assume that θ is the parameter set we would like to estimate and the probability density function (PDF) $f(x|\theta)$ of some data set x given θ is known, where x is referred as the complete data set in the context of the EM algorithm. Let us also assume the PDF f is a continuous and appropriately differentiable function of θ. If x is known, θ can be readily evaluated by maximizing $f(x|\theta)$:

$$\theta = \arg\max_{\theta}\left(\log f(x|\theta)\right). \tag{4.9}$$

Unfortunately in many practical applications, some or all elements of x cannot be obtained directly from the experiments but only by means of another observed data set y. Besides, there can be a many-to-one mapping between x and y. So for the E-step of the EM algorithm,

we first compute the expectation of $\log f\left(x|\theta\right)$ given the data set *y* and our current estimate of θ. That is,

$$Q\left(\theta|\theta^t\right) = E\left(\log f\left(x|\theta\right)\middle|y,\theta^t\right) \tag{4.10}$$

where θ^t is the *t*-th estimation of θ and *t* is the iteration index. Then for the M-step of the EM algorithm, we find the value of θ which maximizes $Q\left(\theta|\theta^t\right)$ as:

$$\theta^{t+1} = \arg\max_{\theta} Q\left(\theta|\theta^t\right) \tag{4.11}$$

where θ^{t+1} is our refined estimation of θ. The E-step and M-step then iterate alternately and will converge to give the ML estimation of θ as proven in [161].

4.4 The New EM Framework for Speech Enhancement in Non-stationary Noise Environments

When applying the EM algorithm to speech enhancement, we consider the cepstral coefficients of the true clean speech power spectrum $S_x(k)$ to be the parameters for estimation. It is defined as,

$$C_x(q) = \frac{1}{M}\sum_{k=0}^{M-1}\log\left(S_x(k)\right)e^{j2\pi kq/M} \quad ; q=0,\ldots M\text{-}1 \tag{4.12}$$

where *M* is the total number of frequency components and the frame index *i* is dropped for notation simplicity. As it is explained in Section 4.2, the *a-priori* SNR of the noisy speech is one of the most important parameters to be estimated in speech enhancement applications. The objective of the proposed algorithm is to obtain a good estimation of C_x so as to compute $S_x(k)$ based on (4.12). Then the *a-priori* SNR of the noisy speech can be obtained and used in the traditional speech enhancement gain functions such as (4.1) or (4.2) for the estimation of the unknown clean speech periodogram.

We propose to use the EM algorithm to help in the estimation of C_x. To do so, let us

first select the cepstral coefficients of the original clean speech periodogram, denoted as $\hat{C}_x$, to be the complete data set in our EM framework. $\hat{C}_x$ can be computed from the periodogram of the original speech *x*, i.e. $\hat{S}_x(k)$ where $\hat{S}_x(k) = |X(k)|^2$, as follows:

$$\hat{C}_x(q) = \frac{1}{M}\sum_{k=0}^{M-1} \log\left(\hat{S}_x(k)\right) e^{j2\pi kq/M} + \gamma\delta_q \quad ; \qquad q=0,\ldots,M\text{-}1 \tag{4.13}$$

where

$$\delta_q = \begin{cases} 1 & for \quad q=0 \\ 0 & otherwise \end{cases} \tag{4.14}$$

and $\gamma \approx 0.577216$ is the Euler's constant. It is shown in [174][101] that under some regularity conditions and for large sample size ($M >> 1$) real-valued data, the estimated cepstral coefficients $\hat{C}_x(q)$ are even symmetric and independent random variables having normal distributions with means $C_x(q)$ and variances $\sigma_e^2(q)$ as follows:

$$\hat{C}_x(q) \sim N\left(C_x(q), \sigma_e^2(q)\right) \quad ; \qquad q=0,\ldots,M/2 \tag{4.15}$$

where

$$\sigma_e^2(q) = \begin{cases} \pi^2/(3M) & for \quad q=0, M/2 \\ \pi^2/(6M) & otherwise \end{cases} . \tag{4.16}$$

In the remaining of this section we shall drop the index *q*, where appropriate, for simplifying the equations. The dependency of *q* on the relevant quantities should be apparent.

It can be seen in (4.15) that the required parameter C_x in fact is the mean of the complete data set $\hat{C}_x$, which has a normal distribution. However, $\hat{C}_x$ is unknown since we do not have the original speech data. Hence we cannot directly compute C_x from $\hat{C}_x$. We have to rely on the observed noisy speech periodogram $\hat{S}_y = |Y|^2$ to help us in estimating C_x. The expectation of $\log f\left(\hat{C}_x | C_x\right)$ given the data set $\hat{S}_y$ and the current estimate of C_x can

be expressed as follows:

$$Q\left(C_x|C_x^t\right)=E\left(\log f\left(\hat{C}_x|C_x\right)\middle|\hat{S}_y,C_x^t\right). \tag{4.17}$$

From (4.15) and (4.16), we know that,

$$f\left(\hat{C}_x|C_x\right)=\frac{1}{\sigma_e\sqrt{2\pi}}\exp\left(\frac{-\left(\hat{C}_x-C_x\right)^2}{2\sigma_e^2}\right), \tag{4.18}$$

hence

$$\begin{aligned} Q\left(C_x|C_x^t\right) &= E\left(\left(\log\left(\frac{1}{\sigma_e\sqrt{2\pi}}\right)+\frac{-\left(\hat{C}_x-C_x\right)^2}{2\sigma_e^2}\right)\middle|\hat{S}_y,C_x^t\right) \\ &= \log\left(\frac{1}{\sigma_e\sqrt{2\pi}}\right)+\frac{-E\left(\left(\hat{C}_x-C_x^t\right)^2\middle|\hat{S}_y,C_x^t\right)}{2\sigma_e^2} \end{aligned} \tag{4.19}$$

We shall apply $Q\left(C_x|C_x^t\right)$ to the M-step of the proposed algorithm. The purpose of the M-step is to optimize C_x^t in order to maximize $Q\left(C_x|C_x^t\right)$. From the recent research in iterative regularization [175], it is known that if the signal is sparse or if the signal can be transformed into a domain where its coefficients are sparse, the inclusion of a penalty term made up by the L_1 norm of the signal or its coefficients can significantly improve the chance for the iterative process to reach its global optimum point. Such idea has been popularly adopted in some image restoration and image reconstruction applications [176][177]. For the proposed method, the EM algorithm is operating in the cepstral domain. Since the coefficients of speech in the cepstral domain are sparse and are very much different from noise, we include a penalty term made up by the L_1 norm of the desired cepstral coefficients in the optimization process. More specifically, the M-step of the proposed algorithm is given as follows:

$$
\begin{aligned}
C_x^{t+1} &= \arg\min_{C_x^t}\left\{-Q\left(C_x \middle| C_x^t\right) + pen(C_x^t)\right\} \\
&= \arg\min_{C_x^t}\left\{E\left(\left(\left(\hat{C}_x - C_x^t\right)^2\right)\middle|\hat{S}_y, C_x^t\right) + 2\sigma_e^2 pen(C_x^t)\right\}
\end{aligned}
\tag{4.20}
$$

where $pen\left(C_x^t\right)$ is the penalty term, which is selected as follows:

$$
pen(C_x^t) = \tau\left\|AC_x^t\right\|_1 = \tau\sum_q\left|A(q)C_x^t(q)\right|. \tag{4.21}
$$

In (4.21), τ is a free parameter that adjusts the amount of regularization applied to the maximization process. Its selection method will be discussed at the end of this section. A is dependent on q and it is defined as a weakly differentiable binary function such that,

$$
A(q) = \begin{cases} 0 & q \le q_l \ and \ q \in \mathrm{C}_{pitch} \\ 1 & otherwise \end{cases} \quad . \tag{4.22}
$$

The introduction of the penalty term imposes a constraint to the optimization process such that the energy of the estimated cepstral coefficients will concentrate at very low quefrencies as well as the quefrencies associated with the fundamental frequency. They are exactly the features of voiced speeches in the cepstral domain. By doing so, the optimization process can be carried out on coefficients with improved SNR and hence reduces the effect due to the estimation error of the non-stationary noise characteristics. It turns out to be the major factor that leads to the good performance of the proposed algorithm in non-stationary noise environments. The use of the L_1 norm in the penalty term in (4.21) is based on the assumption that C_x has a Laplacian prior. From (4.20), C_x^{t+1} can be obtained by taking the derivative of the right hand side of (4.20) and setting the result to 0.That is, let

$$
\Theta = E\left(\left(\left(\hat{C}_x - C_x^t\right)^2\right)\middle|\hat{S}_y, C_x^t\right) + 2\sigma_e^2 pen(C_x^t) \tag{4.23}
$$

Then

$$\begin{aligned}\frac{\partial\Theta}{\partial C_x^t} &= -2E\left(\hat{C}_x\middle|\hat{S}_y, C_x^t\right) + 2C_x^t + 2\sigma_e^2\tau\frac{\partial\sum_q\left|A(q)C_x^t(q)\right|}{\partial C_x^t(q)} \\ &= -2E\left(\hat{C}_x\middle|\hat{S}_y, C_x^t\right) + 2C_x^t + 2\sigma_e^2\tau.A.\text{sign}\left(C_x^t\right)\end{aligned} \tag{4.24}$$

where *sign(x)* returns {1, 0, -1} if x is positive, 0 or negative, respectively. Hence the optimal C_x^t is the one such that $\frac{\partial\Theta}{\partial C_x^t} = 0$. It is then used as the new estimate of C_x^t. That is,

$$\begin{aligned}&-2E\left(\hat{C}_x\middle|\hat{S}_y, C_x^t\right) + 2C_x^{t+1} + 2\sigma_e^2\tau.A.\text{sign}\left(C_x^{t+1}\right) = 0 \\ &C_x^{t+1} + \sigma_e^2\tau.A.\text{sign}\left(C_x^{t+1}\right) = E\left(\hat{C}_x\middle|\hat{S}_y, C_x^t\right) \\ &C_x^{t+1} = \begin{cases}\tilde{C}_x^t + T & \tilde{C}_x^t < -T \\ 0 & -T \le \tilde{C}_x^t \le T \\ \tilde{C}_x^t - T & T < \tilde{C}_x^t\end{cases}\end{aligned} \tag{4.25}$$

where $\tilde{C}_x^t = E\left(\hat{C}_x\middle|\hat{S}_y, C_x^t\right)$ and $T = \sigma_e^2\tau A$. (4.25) is indeed the well-known soft thresholding non-linearity [77][175] with an additional constraint A.

To implement (4.25), we need to have a good estimate of $\tilde{C}_x^t$. For the proposed algorithm, we estimate $\tilde{C}_x^t$ by the following,

1. For initial guess: $\tilde{C}_x^0 = \tilde{C}_x^{TCS}$

2. For subsequent iterations:

$$\tilde{C}_x^t = IDFT\left\{\log\left\{\left(G_{\log mmse}\left(\hat{\xi}^t\right)\cdot\sqrt{\hat{S}_y}\right)^2\right\}\right\} \tag{4.26}$$

where $\hat{\xi}^t = \frac{\exp\left(DFT\left(C_x^t\right)\right)}{\hat{S}_n}$. In both cases, bias is removed using the approach in [101] whenever transforming data between the cepstral domain and the spectral domain. As shown in (4.26), we adopt the TCS method [116] to obtain the initial guess of $\tilde{C}_x^t$ at t=0. Afterwards, we update $\tilde{C}_x^t$ by using an MMSE-LSA gain function [15] in which the *a-priori* SNR is computed based on the current estimate C_x^t. The MMSE-LSA gain function

theoretically gives the minimum mean square error estimation of the log-magnitude spectra. From [15], we know that,

$$E\left(\log \hat{S}_x \middle| \hat{S}_y, C_x^t\right) = \log\left(G_{\log mmse}\left(\hat{\xi}^t\right)\cdot\sqrt{\hat{S}_y}\right)^2 \quad . \tag{4.27}$$

Hence,

$$\begin{aligned} IDFT\left\{\log\left\{\left(G_{\log mmse}\left(\hat{\xi}^t\right)\cdot\sqrt{\hat{S}_y}\right)^2\right\}\right\} &= IDFT\left\{E\left(\log \hat{S}_x \middle| \hat{S}_y, C_x^t\right)\right\} \\ &= E\left(IDFT\left(\log \hat{S}_x\right)\middle| \hat{S}_y, C_x^t\right) \\ &= E\left(\hat{C}_x^t \middle| \hat{S}_y, C_x^t\right) = \hat{C}_x^t \end{aligned} \tag{4.28}$$

It can be seen in (4.28) that the MMSE-LSA gain function can give a good estimation of $\tilde{C}_x^t$. However, we do not use it for the initial guess since without a good *a-priori* SNR estimator, the MMSE-LSA gain function can accidentally remove speech spectral components of low SNR. This is particularly the case if the noise is non-stationary. The TCS method gives a reasonably good estimation of $\tilde{C}_x^t$ without the need of a very good *a-priori* SNR estimator. It also works reasonably well for non-stationary noises. It is thus used as the initial guess of $\tilde{C}_x^t$ and is afterwards refined by the MMSE-LSA gain function using the *a-priori* SNR estimate obtained by the M-step of the proposed algorithm.

To summarize, the proposed speech enhancement algorithm based on the new EM framework can be described as follows:

A. Initial guess:

   Compute the initial guess of $\tilde{C}_x^t$ using the TCS method, i.e. $\tilde{C}_x^0 = \tilde{C}_x^{TCS}$ (see (4.26)).

B. M-step:

   Estimate the parameter C_x^t using the constrained soft thresholding method (see (4.25)).

C. E-step:

Refine the estimation of $\tilde{C}_x^t$ using the MMSE-LSA gain function with the estimated C_x^t obtained from the M-step (see (4.26)).

D. Iterate the M-step and E-step alternately until convergence is reached. The enhanced speech $\tilde{X}$ can thus be obtained by

$$\begin{aligned} \left|\tilde{X}\right| &= G_{\log mmse}\left(\hat{\xi}^{converge}\right).|Y| \\ \tilde{X} &= \left|\tilde{X}\right|\exp(j\angle Y) \end{aligned} \tag{4.29}$$

where $\angle Y$ is the phase angle of Y. The operation of the algorithm is also described in Fig. 4.1.

Besides those required in the original TCS method [116], the proposed algorithm has very few free parameters. Once these parameters are set, they can be used for speeches of different genders and at different noise levels, as it is the case in our simulations. More specifically, the setting of the free parameters in the TCS method can be found in [116]. The parameters for determining the soft threshold T in (4.25) can be obtained as follows: (i) the value of σ_e^2 can be found in (4.16); (ii) the setting of parameter A requires the parameter q_l in (4.22), which is set as *0.025M* in our simulations. C_{pitch} in (4.22) can be obtained from the TCS method given by [116] when generating the initial guess. Finally, the parameter τ is set as

$$\tau = \frac{\sqrt{2}}{\sigma} \tag{4.30}$$

where σ^2 is the variance of C_x and is approximated by,

$$\sigma^2 = Var[\hat{C}_x] - \sigma_e^2 . \tag{4.31}$$

We show in the Appendix A that by setting τ as in (4.30), the soft thresholding operation in (4.25) indeed achieves a good approximation of the maximum *a-posterior* (MAP) estimation of C_x. Moreover, as shown in Appendix B, a bias compensation method is applied in log-spectral domain to improve the spectral estimator. We would like to emphasize that

due to the initial guess and the additional constraints applied to the algorithm, the proposed algorithm need not be iterated many times to achieve satisfactory performance. In our experiments, iterating 3 times already gives very good results. Iterating further in most cases will only increase the computation time but not further improve the performance. Hence it is not recommended.

4.5 Simulations and Results

In this section, the performance of the proposed algorithm is shown and compared with the state-of-the-art speech enhancement methods. To start with, we use an example to illustrate the deficiency of the traditional TCS method and how it is solved by the proposed algorithm. For the ease in presentation, let us first denote the proposed algorithm as Logmmse-L1-EM, which represents the two major operations (MMSE-LSA gain function and L_1 norm regularization) used in the new EM framework. Fig. 4.2 shows a segment of a typical noisy speech periodogram (red line), its original clean speech periodogram (black line), the enhanced speech periodogram using the traditional TCS (green line) and the proposed Logmmse-L1-EM algorithm (yellow line). It can be seen that the noisy periodogram consists of a sharp noise spectral peak at frequency index about 170. The TCS method tries to remove the noise spectral peak by reducing the variance of the cepstral coefficients. It however is only partially successful and leaves behind a rather strong noise spectral peak. In addition, the TCS method over-smoothes the speech spectral peak at indices near 100 and 120. It is clear that a fixed set of smoothing parameters is difficult to handle noisy speech spectral components equally well if they have great difference in SNR. On the contrary, the proposed Logmmse-L1-EM algorithm gives extremely good performance in removing the noise spectral peak while keeping the speech spectral peaks. It is due to the iterative process through EM by which the noise spectral peak is reduced successively in each

iteration while the speech spectral peaks are also gradually adjusted. Fig. 4.3 illustrates how the proposed algorithm refines the estimation in each iteration.

The results in Fig. 4.2 and Fig. 4.3 clearly explain why the proposed Logmmse-L1-EM algorithm can provide a better performance than the original TCS method and the traditional MMSE-LSA method. As mentioned above, the TCS method cannot take care of noisy components with great difference in SNR. However, it serves as a good initial guess of the clean speech power spectrum since it is less sensitive to the accuracy of the *a-priori* SNR estimation. The initial guess is then applied to the L_1 norm regularizer in the M-step of the proposed EM framework, which is indeed a constraint soft thresholding process. Besides a good MAP estimation of the true power spectrum of the clean speech, the applied constraint fully utilizes the sparsity feature of speech signals in the cepstral domain. Noises, no matter stationary or non-stationary, will be rejected since most of them cannot fulfill this constraint. For every iteration the proposed algorithm performs, the same constraint is imposed to the observed noisy data and gradually improves the estimation, which is shown in Fig. 4.3. The M-step of the proposed EM algorithm enables a gradually improved *a-priori* SNR estimate for use in the E-step of the proposed algorithm, which is just the traditional MMSE-LSA method. With a good *a-priori* SNR, the MMSE-LSA method can often give a good estimate to the original clean speech periodogram for the M-step again.

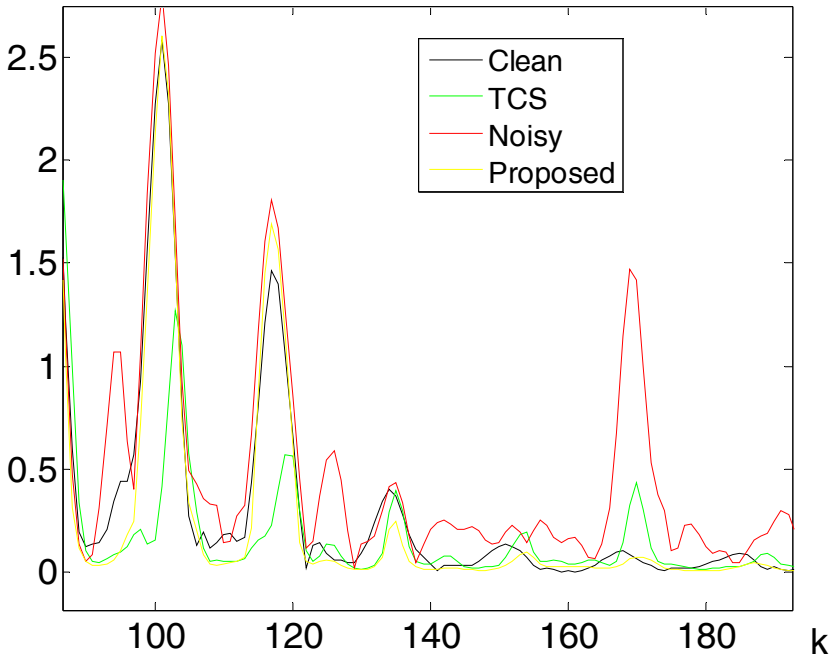


Fig. 4.2 – A comparison of the traditional TCS method and the proposed Logmmse-L1-EM algorithm

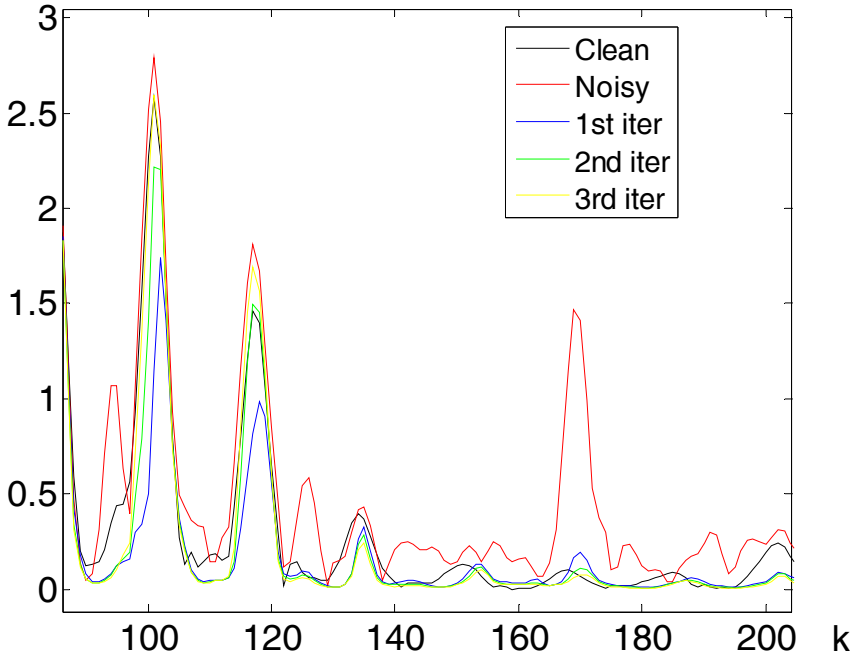


Fig. 4.3 – The result of the proposed Logmmse-L1-EM algorithm after each iteration

A comparison of the spectrogram of the enhanced speeches generated using different algorithms with different colored noises at input is shown in Fig. 4.4 to Fig. 4.5. Table **4.1** gives a summary of the algorithms that have been compared. We start with the case that the speech is contaminated by pink noise, which is a relatively stationary noise. Fig. 4.4a shows the clean speech spectrogram of a female speech selected from the TIMIT database [192] saying the following sentence: "She had your dark suit in greasy wash water all year". Fig. 4.4b shows the result when pink noise is added to the speech with input segSNR about 5dB. Fig. 4.4c depicts the spectrogram using the traditional TCS method. It can be seen that although the TCS method can recover much speech contents, its noise control is not sufficient and strong background residue noise remains in the enhanced speech. Fig. 4.4d shows the spectrogram using the MMSE-LSA method plus SPP with fixed prior [49], which is a relatively recent MMSE-LSA estimator enhanced by using a special speech presence probability function. It has better control of the background noise however it also removes speech content and the intelligibility of the enhanced speech is reduced. Furthermore, musical noise appears particularly at the beginning of the speech. Fig. 4.4e shows the spectrogram given by a combination of the TCS method and the MMSE-LSA plus SPP [140]. In that approach, the TCS is used for the estimation of the *a-priori* SNR and is used in the MMSE-LSA and the SPP estimation. It has better noise control compared with that in Fig. 4.4c and Fig. 4.4d. However, some speech contents are removed as it is indicated in the circled areas. It seems that the speech contents removed by the SPP estimator cannot be recovered although the TCS method is used. Fig. 4.4f shows the spectrogram using the proposed Logmmse-L1-EM algorithm. It has very well background noise control and the speech content is also better preserved as indicated in the circled areas.

Fig. 4.5 shows the case where the speech is contaminated by the buccaneer noise, which is a non-stationary noise such that two tones of varying frequencies together with other

background noises are added to the speech. The noisy speech is shown in Fig. 4.5b. When speeches are contaminated by non-stationary noises, traditional decision-direct approach will generate large error when estimating the a-priori SNR. Hence the performance of the traditional the MMSE-LSA approach, although using SPP, will not be good. As can be seen in Fig. 4.5d, the strong tones cannot be removed and will be quite annoying in human auditory. The TCS method improves slightly as shown in Fig. 4.5c. The combined TCS and MMSE-LSA approach gives better result as can be seen in Fig. 4.5e. However in both cases, the tones still cannot be sufficiently removed while the background noises remain. The result of the proposed Logmmse-L1-EM algorithm is illustrated in Fig. 4.5f. It is clear that the time varying tones are largely suppressed while the speech contents are well preserved comparable with that using the TCS method. The background noise is also significantly reduced. The overall performance is very promising.

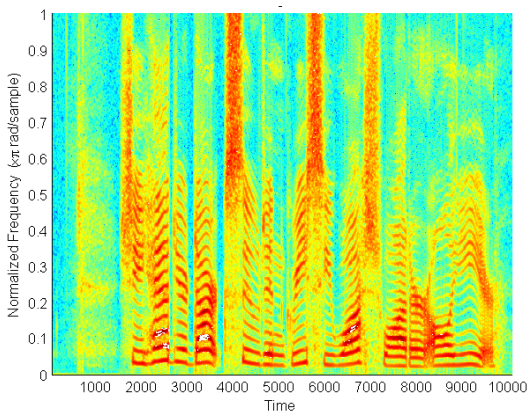
The performance of the proposed Logmmse-L1-EM algorithm is further evaluated using standard evaluation measures. A series of simulations have been performed for comparing the performance between the following approaches: MMSE-LSA [15], MMSE-Gamma [178], MMSE-LSA plus SPP with fixed prior [49], MMSE-LSA plus TCS method [140], and the proposed Logmmse-L1-EM algorithm. The speech sampling rate is 16kHz. Simulation details are listed as follows: frame size – 512 samples (~32ms), FFT size – 1024 samples (zeros padded each frame with 512 samples), window shift step size – 128 samples (75% overlap). For all algorithms, the noise power spectrum is estimated by first using the initial frames that are assumed to have no speech energy; then updated whenever a frame is detected to have no speech energy by using a VAD [53]. For the algorithms using the MMSE-LSA gain function, G_{min} is set at *-25dB* which helps masking musical noise and limits speech distortion.

In the simulation, 40 male and 40 female test speeches were arbitrarily selected from the TIMIT database [192]. The noise signals were adopted from the NOISEX-92 database [193]

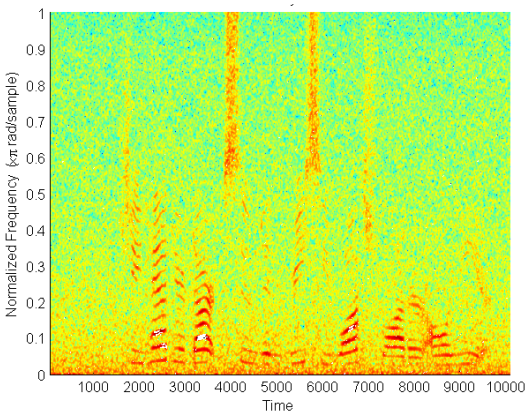
and added to the speeches with input segmental signal-to-noise ratio (segSNR) [1] ranging from about -10dB to +10dB. The resulting enhanced speeches generated by all algorithms were evaluated using standard measures including (i) the segSNR; and (ii) the perceptual evaluation of speech quality (PESQ), which is an ITU standard for evaluating speech quality [58]. The results are shown in Fig. 4.6. It can be seen that the performance of the proposed algorithm is always the best in all cases. For instance, when comparing with the MMSE-LSA plus TCS approach [140], the proposed algorithm can always give an improvement in segSNR and PESQ score for all noise signals. More specifically, for the pink noise case, the average improvement in segSNR and PESQ is about 0.9dB and 0.1, respectively. For the buccaneer noise case, the average improvement in segSNR and PESQ is about 0.85dB and 0.12, respectively. Similar results can be found in Fig. 4.6 for other kinds of noise contamination.

Table 4.1 - Summary of the algorithms compared in the simulations.

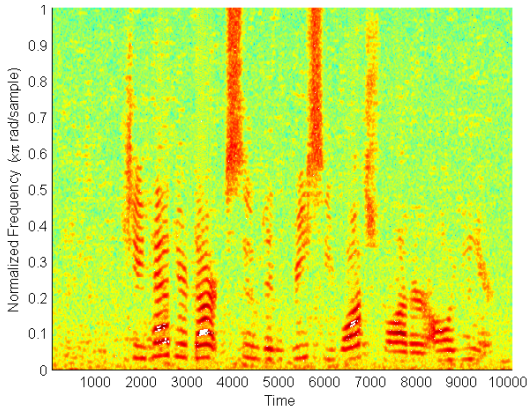
Method	**Description**
MMSE-LSA	Minimum mean-square error log-spectral amplitude estimator [15]
MMSE-Gamma	Minimum mean-square error spectral amplitude estimator with generalized gamma speech priors [178]
TCS	Temporal cepstrum smoothing method [116]
MMSE-LSA FP SPP	Using the MMSE-LSA gain function plus SPP estimated with fixed prior [49]
MMSE-LSA TCS SPP	Using the MMSE-LSA gain function plus SPP estimated using the TCS method [140]
Logmmse-L1-EM	The proposed algorithm based on the new EM framework



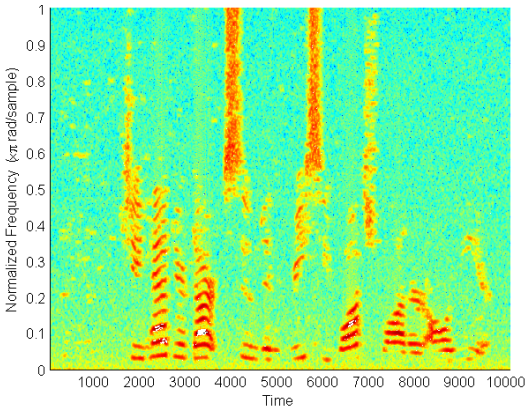
(a) Original speech



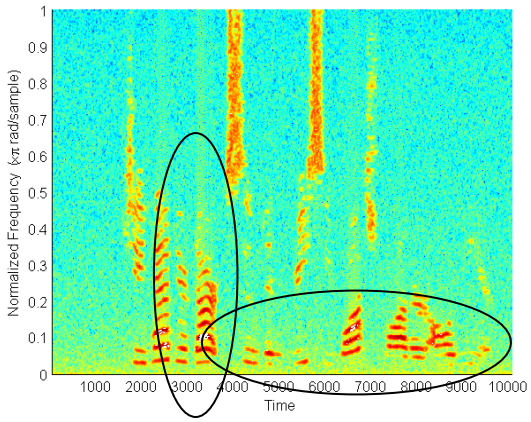
(b) Noisy speech (Noise: Pink, ~5dB segSNR)



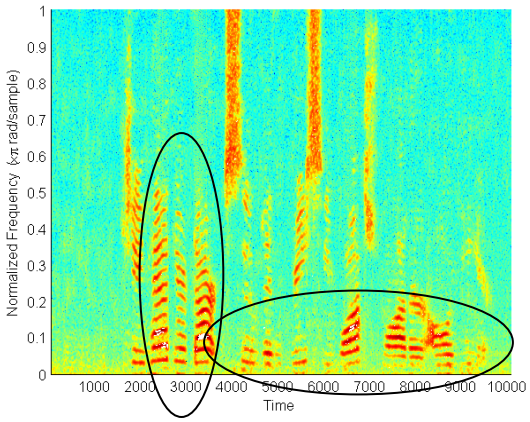
(c) Traditional TCS method [116]



(d) MMSE-LSA plus SPP with fixed prior [49]

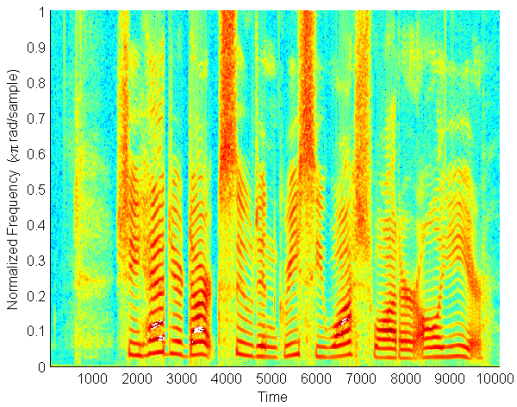


(e) MMSE-LSA plus SPP and TCS [140]

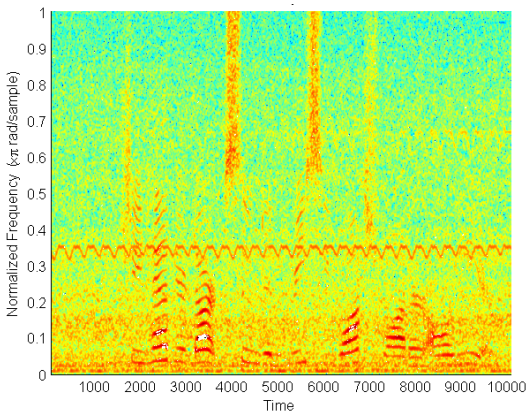


(f) The proposed Logmmse-L1-EM

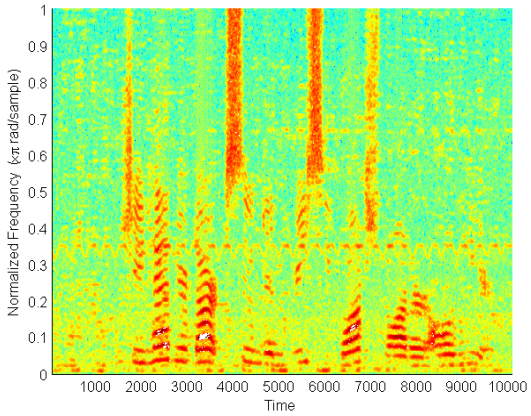
Fig. 4.4 – Spectrogram of the original, noisy and enhanced speeches (pink noise)



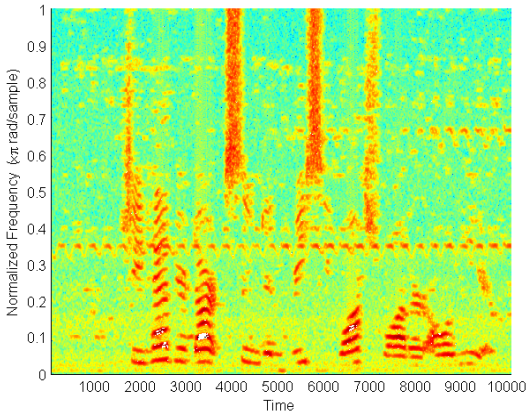
(a) Original speech



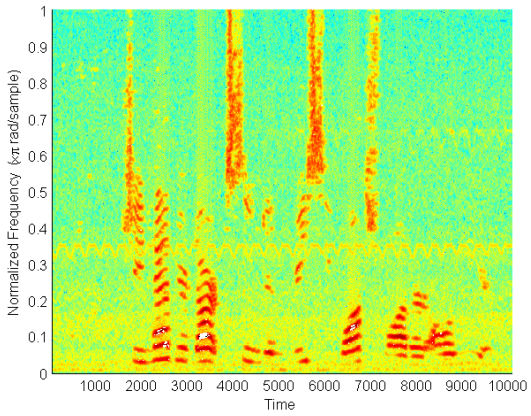
(b) Noisy speech (Noise: buccaneer1, ~5dB segSNR)



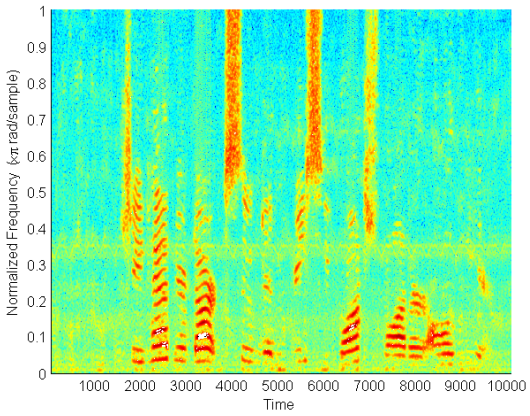
(c) Traditional TCS method



(d) MMSE-LSA plus SPP with fixed prior

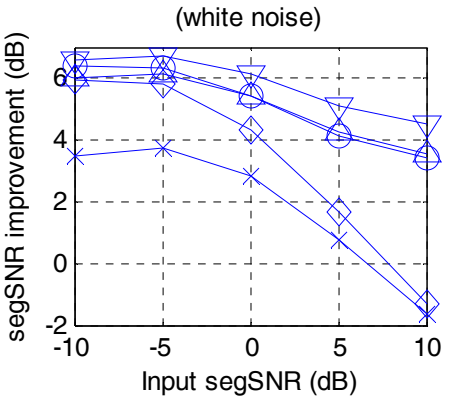


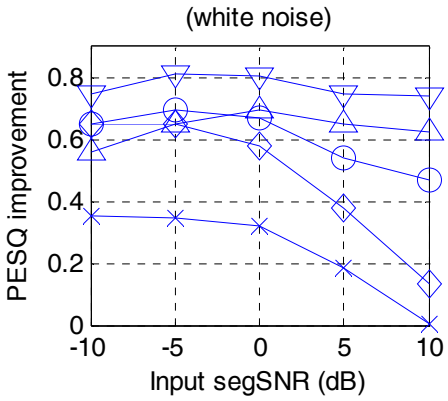
(e) MMSE-LSA plus SPP and TCS

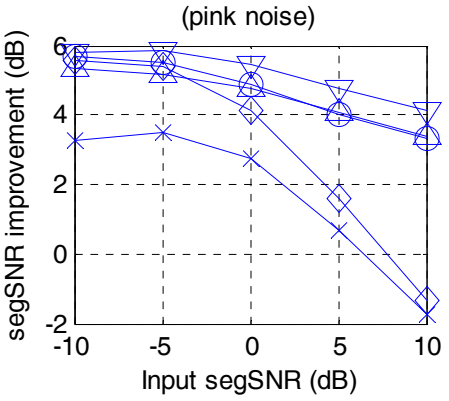


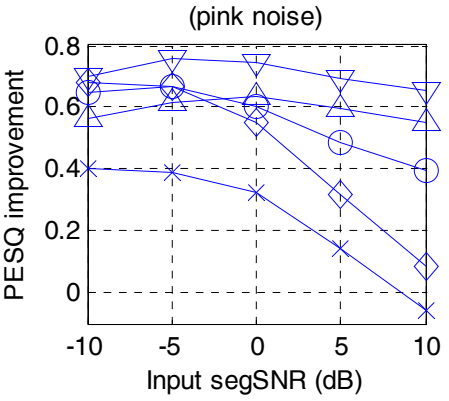
(f) The proposed logmmse-L1-EM

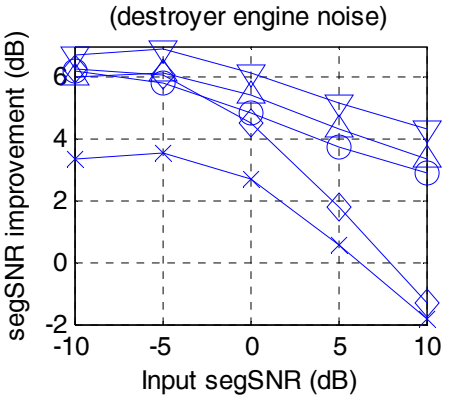
Fig. 4.5 – Spectrogram of the original, noisy and enhanced speeches (buccaneer noise)

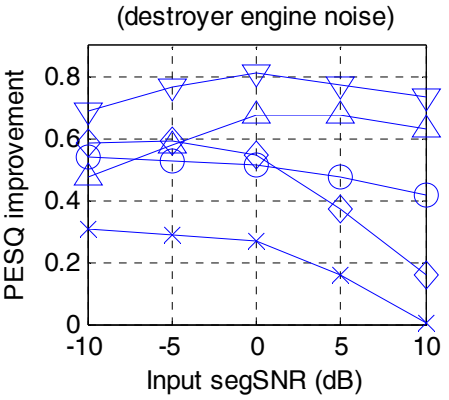

(white noise)
segSNR improvement (dB)
Input segSNR (dB)

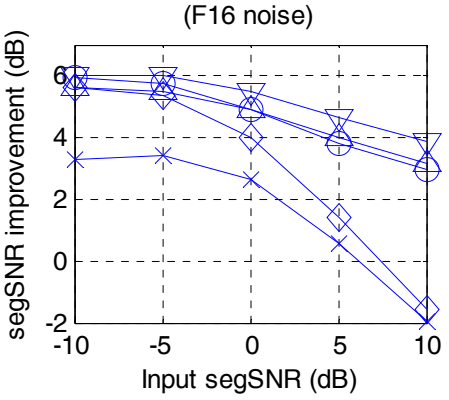

(white noise)
PESQ improvement
Input segSNR (dB)

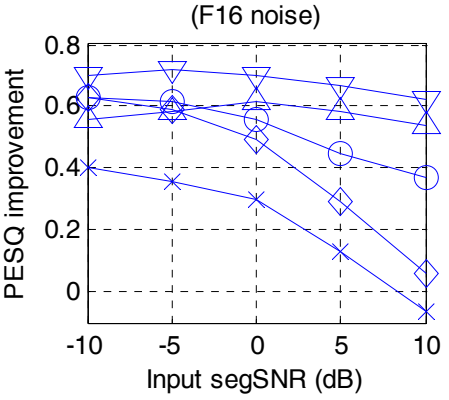

(pink noise)
segSNR improvement (dB)
Input segSNR (dB)


(pink noise)
PESQ improvement
Input segSNR (dB)


(destroyer engine noise)
segSNR improvement (dB)
Input segSNR (dB)


(destroyer engine noise)
PESQ improvement
Input segSNR (dB)


(F16 noise)
segSNR improvement (dB)
Input segSNR (dB)


(F16 noise)
PESQ improvement
Input segSNR (dB)

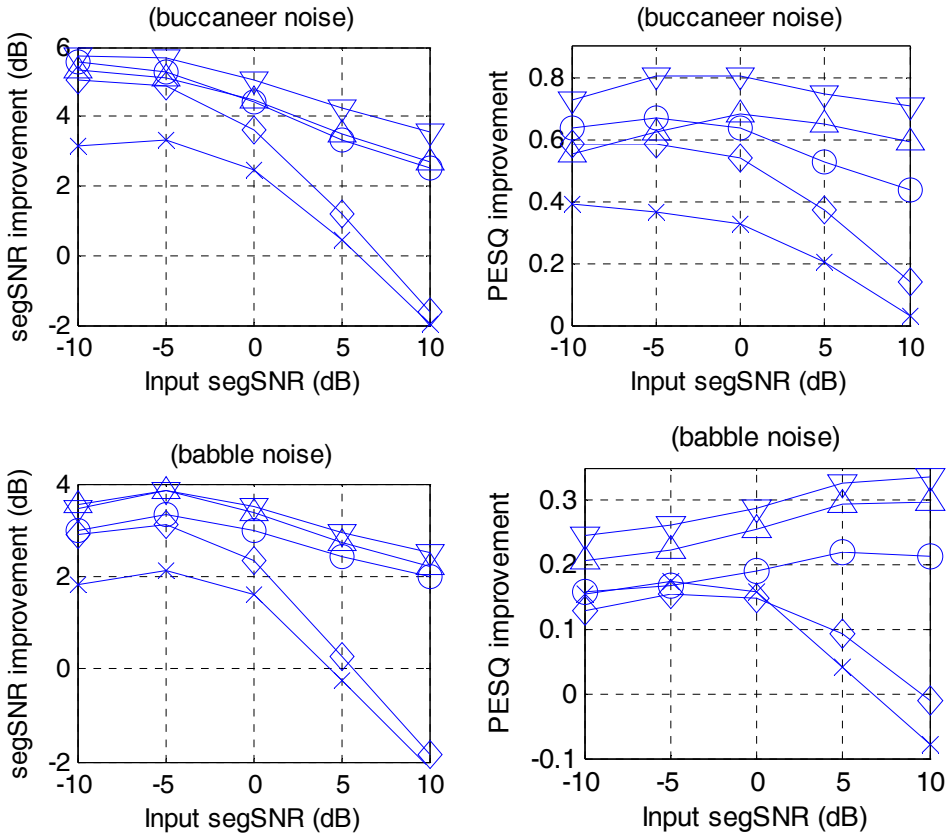


Fig. 4.6 – (Left) Segmental SNR improvement over the noisy speech; (Right) PESQ improvement over the noisy speech achieved by using MMSE-LSA [15] (‘x’), MMSE-Gamma [178] (‘◇’), MMSE-LSA FP SPP [49] ('O') MMSE-LSA TCS SPP [140] (‘Δ’) and the proposed Logmmse-L1-EM (‘∇’) for the case of white noise, pink noise, destroyer engine noise, F16 noise, buccaneer noise and babble noise contamination.

To predict the quality of noisy speech enhanced by noise suppression algorithms, three composite objective metrics [1] are often used in the literature, which include (a) C_{sig}: Signal distortion (SIG) formed by linearly combining the LLR, PESQ, and WSS measures; (b) C_{bak}: Noise distortion (BAK) formed by linearly combining the segSNR, PESQ, and WSS measures. (c) C_{ovl}: Overall quality (OVL) formed by linearly combining the PESQ, LLR, and WSS measures. Table 4.2 lists the performance of 5 different algorithms for noisy speech signals contaminated by 6 different kinds of noise, both stationary and non-stationary, at input SNR ranging from 0dB to 20dB. Concurring to the results in segSNR and PESQ, the proposed Logmmse-L1-EM algorithm always outperforms the other 4; and in many cases, the improvement is significant. These results have demonstrated the robustness of the proposed algorithm. Its performance is consistent when enhancing speeches made by people of different genders and are contaminated by different kinds of noise at different noise levels.

Table 4.2 - Composite measurement comparison of different algorithms.

Noise	Method \ Input SNR	C_{sig}					C_{bak}					C_{ovl}				
		0	5	10	15	20	0	5	10	15	20	0	5	10	15	20
White	Noisy	1.34	1.92	2.61	3.26	3.87	1.72	2.17	2.64	3.12	3.59	1.35	1.87	2.42	2.95	3.45
	MMSE-LSA	1.84	2.54	3.16	3.67	4.05	2.18	2.59	2.97	3.28	3.51	1.80	2.37	2.86	3.26	3.55
	MMSE-Gamma	2.09	2.78	3.31	3.71	3.98	2.36	2.76	3.10	3.35	3.55	2.04	2.60	3.03	3.35	3.57
	MMSE-LSA FP SPP	2.30	2.99	3.52	3.93	4.29	2.47	2.88	3.25	3.62	3.99	2.19	2.75	3.19	3.56	3.88
	MMSE-LSA TCS SPP	2.11	2.91	3.50	3.96	4.33	2.41	2.87	3.29	3.69	4.08	2.05	2.70	3.21	3.62	3.98
	Proposed Logmmse-L1-EM	**2.47**	**3.17**	**3.71**	**4.17**	**4.56**	**2.62**	**3.05**	**3.43**	**3.82**	**4.23**	**2.35**	**2.94**	**3.38**	**3.79**	**4.17**
babble	Noisy	2.52	3.09	3.63	4.10	4.50	1.68	2.15	2.64	3.13	3.61	1.96	2.47	2.95	3.39	3.78
	MMSE-LSA	**2.53**	**3.11**	3.60	3.97	4.22	1.87	2.35	2.79	3.16	3.44	**2.04**	2.56	3.01	3.36	3.59
	MMSE-Gamma	1.95	2.61	3.17	3.60	3.89	1.66	2.20	2.69	3.10	3.40	1.64	2.23	2.74	3.15	3.44
	MMSE-LSA FP SPP	2.15	2.89	3.53	4.06	4.49	1.73	2.27	2.80	3.30	3.79	1.76	2.40	2.96	3.45	3.86
	MMSE-LSA TCS SPP	2.21	2.94	3.56	4.07	4.49	1.86	2.39	2.90	3.39	3.85	1.84	2.47	3.03	3.50	3.91
	Proposed Logmmse-L1-EM	2.41	3.09	**3.69**	**4.19**	**4.59**	**1.95**	**2.47**	**2.96**	**3.46**	**3.92**	2.00	**2.59**	**3.12**	**3.59**	**3.99**
destroyer engine	Noisy	2.00	2.59	3.19	3.75	4.25	1.52	1.98	2.46	2.96	3.47	1.65	2.14	2.64	3.12	3.57
	MMSE-LSA	2.45	3.02	3.51	3.89	4.15	1.96	2.39	2.79	3.14	3.41	2.04	2.53	2.95	3.29	3.55
	MMSE-Gamma	2.71	3.19	3.55	3.80	3.98	2.26	2.66	3.00	3.28	3.49	2.32	2.75	3.09	3.35	3.53
	MMSE-LSA FP SPP	2.58	3.03	3.48	3.94	4.39	2.18	2.56	2.97	3.39	3.83	2.21	2.62	3.04	3.44	3.85
	MMSE-LSA TCS SPP	2.57	3.20	3.67	4.09	4.49	2.19	2.71	3.16	3.59	4.00	2.19	2.78	3.24	3.64	4.01
	Proposed Logmmse-L1-EM	**3.00**	**3.55**	**3.97**	**4.36**	**4.72**	**2.49**	**2.94**	**3.35**	**3.74**	**4.16**	**2.57**	**3.08**	**3.48**	**3.85**	**4.20**
F16	Noisy	1.95	2.58	3.20	3.77	4.28	1.57	2.05	2.54	3.05	3.55	1.63	2.17	2.70	3.20	3.66
	MMSE-LSA	2.45	3.05	3.56	3.94	4.20	2.03	2.47	2.87	3.21	3.46	2.09	2.60	3.04	3.37	3.61
	MMSE-Gamma	2.57	3.11	3.51	3.80	4.00	2.19	2.61	2.98	3.28	3.50	2.23	2.71	3.09	3.36	3.55
	MMSE-LSA FP SPP	2.69	3.23	3.64	4.05	4.46	2.28	2.72	3.12	3.52	3.93	2.32	2.81	3.20	3.57	3.94
	MMSE-LSA TCS SPP	2.56	3.18	3.67	4.07	4.48	2.21	2.72	3.18	3.61	4.04	2.21	2.79	3.26	3.65	4.03
	Proposed Logmmse-L1-EM	**2.89**	**3.45**	**3.88**	**4.30**	**4.68**	**2.43**	**2.89**	**3.32**	**3.74**	**4.16**	**2.50**	**3.01**	**3.43**	**3.82**	**4.19**
pink	Noisy	1.81	2.45	3.10	3.70	4.24	1.61	2.08	2.57	3.08	3.57	1.56	2.11	2.65	3.17	3.64
	MMSE-LSA	2.36	3.01	3.55	3.96	4.24	2.09	2.53	2.93	3.25	3.49	2.05	2.60	3.06	3.40	3.64
	MMSE-Gamma	2.50	3.12	3.57	3.88	4.07	2.25	2.68	3.04	3.33	3.54	2.23	2.75	3.15	3.43	3.60
	MMSE-LSA FP SPP	2.61	3.20	3.65	4.02	4.37	2.30	2.73	3.16	3.58	3.99	2.30	2.82	3.23	3.59	3.92
	MMSE-LSA TCS SPP	2.45	3.15	3.67	4.06	4.41	2.23	2.72	3.20	3.64	4.08	2.17	2.78	3.27	3.66	4.01
	Proposed Logmmse-L1-EM	**2.81**	**3.44**	**3.91**	**4.28**	**4.63**	**2.45**	**2.92**	**3.35**	**3.78**	**4.20**	**2.47**	**3.03**	**3.46**	**3.84**	**4.19**
buccaneer	Noisy	1.79	2.40	3.04	3.63	4.16	1.54	2.00	2.49	2.99	3.49	1.50	2.01	2.55	3.06	3.53
	MMSE-LSA	2.27	2.90	3.45	3.88	4.17	1.99	2.42	2.82	3.17	3.43	1.93	2.46	2.93	3.30	3.56
	MMSE-Gamma	2.25	2.88	3.37	3.72	3.95	2.04	2.48	2.88	3.21	3.45	1.97	2.52	2.95	3.27	3.49
	MMSE-LSA FP SPP	2.44	3.04	3.53	3.98	4.39	2.18	2.61	3.02	3.43	3.85	2.13	2.66	3.10	3.49	3.86
	MMSE-LSA TCS SPP	2.38	3.06	3.59	4.01	4.42	2.16	2.64	3.10	3.53	3.95	2.07	2.68	3.17	3.58	3.96
	Proposed Logmmse-L1-EM	**2.71**	**3.35**	**3.84**	**4.27**	**4.67**	**2.36**	**2.82**	**3.25**	**3.67**	**4.10**	**2.37**	**2.93**	**3.37**	**3.78**	**4.15**

4.6 Chapter Summary

In this chapter, an improved speech enhancement algorithm based on a novel expectation-maximization framework is proposed. The new algorithm makes use of the EM algorithm to define a theoretical framework for the estimation of the true power spectrum of the original speech and its periodogram from a noisy observation. The proposed algorithm starts with the traditional cepstrum smoothing method which gives the initial guess of the periodogram of the clean speech. It is applied to an L_1 norm regularizer in the M-step of the EM framework to estimate the cepstral coefficients of the true speech power spectrum. It enables the estimation of the *a-priori* SNR and is used in the E-step, which is indeed an MMSE-LSA gain function, to refine the estimate of the clean speech periodogram. The M-step and E-step then iterate for 2 more times, with which we have shown to be sufficient in most cases to achieve good result. The proposed algorithm fully utilizes the sparsity of speeches in the cepstral domain by adopting the L_1 norm regularizer. It enables the optimization process to be carried out on coefficients with improved SNR and hence reduces the effect due to the estimation error of the non-stationary noise characteristics. As a result, the proposed algorithm works particularly well when the input speech is contaminated by non-stationary noises. Besides, due to the iterative process, the proposed algorithm has very good control of the residue background noises which makes it outperform the traditional methods. Simulation results have verified that the proposed algorithm improves over the competing speech enhancement methods in almost all testing conditions, such as different kinds of noise at different noise levels using different evaluation measures. They have clearly demonstrated the robustness of the proposed algorithms in general speech enhancement applications.

4.7 Appendices

4.7.1 Appendix A – MAP Estimation of C_x

Given $\hat{C}_x$, the MAP estimator of C_x is,

$$\tilde{C}_x = \arg\max_{C_x} f\left(C_x \mid \hat{C}_x\right). \tag{4.32}$$

We obtain by using the Bayes' rule,

$$\tilde{C}_x = \arg\max_{C_x} \frac{f(\hat{C}_x \mid C_x) f(C_x)}{f(\hat{C}_x)} \quad . \tag{4.33}$$

Since the value of C_x that maximizes the right-hand side is not influenced by the denominator, the MAP estimate of C_x can be rewritten as,

$$\tilde{C}_x = \arg\max_{C_x} \left[f(\hat{C}_x \mid C_x) f(C_x) \right] \tag{4.34}$$

The logarithm function can be applied to (4.34) because it is monotonic. Hence,

$$\tilde{C}_x = \arg\max_{C_x} \left[\log\left(f(\hat{C}_x \mid C_x) \right) + \log\left(f(C_x) \right) \right]. \tag{4.35}$$

As $\hat{C}_x$ is normal distributed with mean C_x [101],

$$f(\hat{C}_x \mid C_x) = \frac{1}{\sigma_e \sqrt{2\pi}} \exp\left(-\frac{(\hat{C}_x - C_x)^2}{2\ \sigma_e^2} \right). \tag{4.36}$$

By using (4.36), (4.35) becomes,

$$\tilde{C}_x = \arg\max_{C_x} \left[\frac{-(\hat{C}_x - C_x)^2}{2\sigma_e^2} + \log\left(f(C_x) \right) \right] \tag{4.37}$$

Let us define $g\left(C_x\right) = \log\left(f(C_x) \right)$. Then we have,

$$\tilde{C}_x = \arg\max_{C_x} \left[\frac{-(\hat{C}_x - C_x)^2}{2\sigma_e^2} + g(C_x) \right]. \tag{4.38}$$

We can obtain the MAP estimate of C_x by taking the derivative of the terms in the square

bracket in (4.38) with respect to C_x. Then,

$$\frac{(\hat{C}_x - \tilde{C}_x)}{\sigma_e^2} + g'(\tilde{C}_x) = 0 . \tag{4.39}$$

We now need the prior $f(C_x)$, i.e. the distribution of cepstral coefficients, C_x, of the clean speech. In Fig. 4.7, we show the PDF $f(\hat{C}_x)$ obtained from the cepstral coefficients of 40 male and 40 female test speeches from the TIMIT database [192], and fit it with different distributions. It is seen that it can be modeled by a Laplacian, or Gaussian or Generalized Gaussian Distribution (GGD) without large error. It is highly likely that this will also be the case for $f(C_x)$. Assume that $f(C_x)$ is modeled using a Laplacian PDF:

$$f(C_x) = \frac{1}{\sigma\sqrt{2}} \exp\left(-\frac{\sqrt{2}}{\sigma} |C_x| \right) \tag{4.40}$$

In this case,

$$g(C_x) = -\log(\sigma\sqrt{2}) - \frac{\sqrt{2}}{\sigma} |C_x| . \tag{4.41}$$

As a result,

$$g'(C_x) = -\frac{\sqrt{2}}{\sigma} sign(C_x) . \tag{4.42}$$

Put (4.42) into (4.39), we have,

$$\hat{C}_x = \tilde{C}_x + \frac{\sigma_e^2 \sqrt{2}}{\sigma} sign(\tilde{C}_x) . \tag{4.43}$$

Therefore, $\tilde{C}_x$ as a function of $\hat{C}_x$ is given by

$$\tilde{C}_x = \begin{cases} \hat{C}_x + T, & \hat{C}_x < -T \\ 0, & -T \le \hat{C}_x \le T \\ \hat{C}_x - T, & T < \hat{C}_x \end{cases} . \tag{4.44}$$

This is the soft threshold nonlinearity. T in this case is given by,

$$T = \frac{\sigma_e^2 \sqrt{2}}{\sigma} \tag{4.45}$$

which is similar to that in (4.25) except the omission of the constraint A. To summarize, the soft thresholding operation as defined in (4.25) is a good approximation of the MAP estimate of C_x with the assumption that C_x has a Laplacian prior.

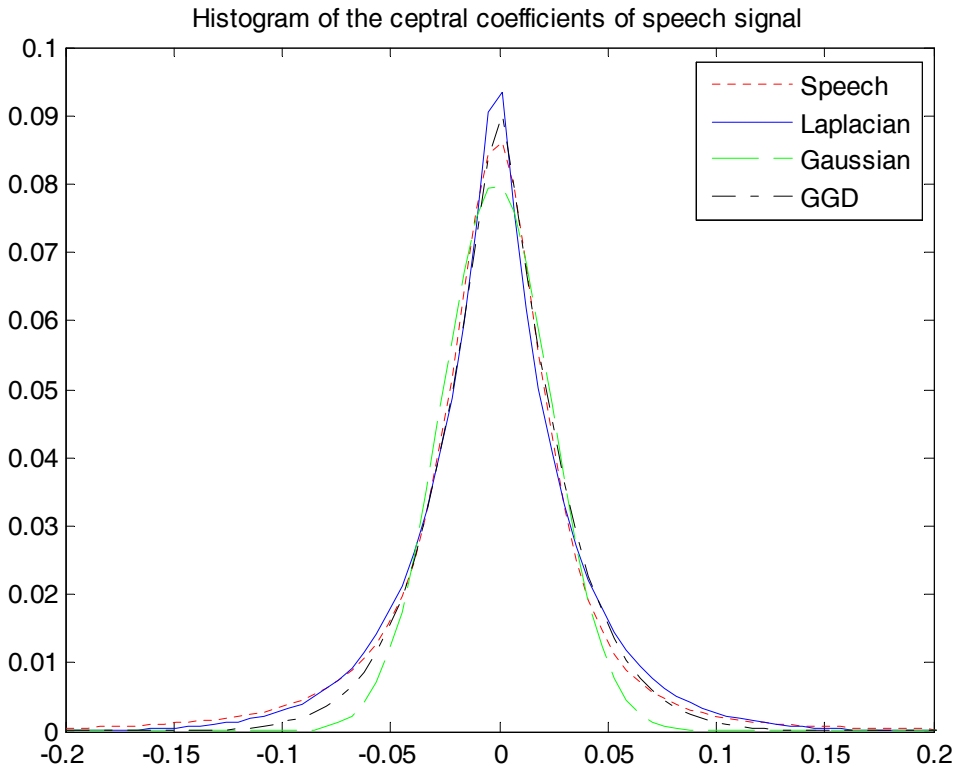


Fig. 4.7 – The PDF *of* cepstral coefficients of 40 male and 40 female test speeches from the TIMIT database [192]. The PDF approximated directly from the speeches (red dotted line). The PDF fit by using a Laplacian model (σ=0.02792, $\beta=1$) (blue solid line), Gaussian model (σ =0.02510, $\beta=2$) (green dashed line), and GGD model (σ =0.03223, $\beta=1.2783$) (black dash-dot line)

4.7.2 Appendix B – Bias compensation in log-spectral domain

In general, the applied smoothing and thresholding (ST) process in the cepstral domain is biased, as it does not only reduce the variance of power spectrum but also affects its means. Since it is believed that the enhancement process can follow better the perceptual characteristics of human auditory system when it is carried out in log domain, the scaling factor α_{ls} of bias correction is proposed to minimize the sum of square in log-spectral domain.

$$\min_{\alpha_{ls}} \sum_{k=0}^{N-1} \left(\log\left(\hat{S}_x(k)\right) - \alpha_{ls} \log\left(\hat{S}_x^{ST}(k)\right) \right)^2 \tag{4.46}$$

where

$$\log\left(\hat{S}_x^{ST}(k)\right) = FFT\left(\widetilde{\boldsymbol{C}}_{\boldsymbol{x}}^{ST}(q)\right) \tag{4.47}$$

The least square estimate of the force constant, α_{ls} is given by

$$\hat{\alpha}_{ls} = \frac{\sum_{K=0}^{N-1} \log\left(\hat{S}_x(k)\right) \log\left(\hat{S}_x^{ST}(k)\right)}{\sum_{K=0}^{N-1} \left(\log\left(\hat{S}_x(k)\right) \right)^2} \tag{4.48}$$

Then, the proposed nonparametric spectral estimate is obtained from the $\widetilde{\boldsymbol{C}}_{\boldsymbol{x}}^{ST}(q)$ by the simple scaling α_{ls} as

$$\widetilde{S}_x(k) = \exp\left[\hat{\alpha}_{ls} FFT\left(\widetilde{C}_x^{ST}(q)\right)\right] \tag{4.49}$$

The effect of bias compensation in log-spectral is shown in Fig. 4.8. It is seen that the ST process introduces a signal power bias, and is successfully compensated with the bias compensation in log-spectral. And then, a refined *a-priori* SNR could be generated to be used in the later enhancement process.

$$\hat{\xi}^{ST}(k) = \frac{\widetilde{\boldsymbol{S}}_{\boldsymbol{x}}(k)}{\hat{S}_n(k)} \tag{4.50}$$

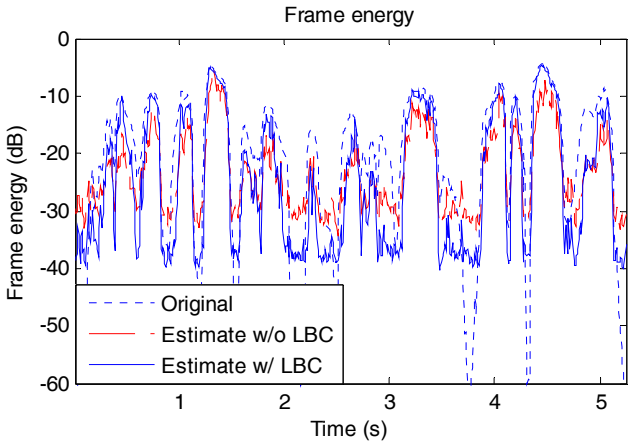


Fig. 4.8 – Effect of the bias compensation in log-spectral for estimate speech. Original speech segment energies (dotted line), the speech segment energies after the proposed ST estimation (dashed line) and the speech segment energies after the proposed ST estimation and bias compensation in log-spectral (solid line).

Chapter 5 A Speech Enhancement Method Based on Sparse Reconstruction on Log-Spectra

5.1 Introduction

In Chapter 4, we proposed a new EM framework for speech enhancement. In the core of the framework, a new L_1-norm regularization process is developed for the estimation of the cepstral coefficients of the speech true power spectrum. The regularizer is very effective since the cepstral coefficients are in fact a kind of sparse representation of the speech power spectrum, as explained in Chapter 2. A natural question thus arises if the cepstral coefficients are the best representation of the speech power spectrum as far as the sparsity is concerned. If there is a sparser representation available, it may be possible to further improve the estimation and hence the speech enhancement performance.

As it is introduced in Chapter 2, the techniques of sparse representation and dictionary learning are widely investigated and have provided possible solutions for many signal processing problems. The goal of these techniques is to look for the sparsest representation of a signal in terms of linear combination of atoms in an overcomplete dictionary. They have been adopted in many applications in speech processing, such as speech recognition [184], voice activity detection (VAD) [186] and speaker identification [187], etc. They have also been introduced to speech enhancement methods [134]-[137] as well. For instance, in [136], the approximation of K-SVD [128] algorithm and the Least Angle Regression (LARS) with a coherence criterion (LARC) are used to learn the composite dictionary and reconstruct the speech spectral amplitude. The LARC method extends the LARS algorithm to include a residual coherence stopping criterion and optimize it to solve a large number of simultaneous coding problems efficiently. The residual coherence stopping criterion of LARC is invariant

to changes in signal energy. In addition, the LARC method allows the code and the dictionary entries to assume values of the entire real domain. However, the composite dictionary depends on the noise dictionary which is difficult to train and inconvenient to use because the background noise type is diversified and the property is not known in advance for general applications. On the other hand, a speech enhancement method using the sparse reconstruction of the approximated PSD through the magnitude-squared spectrum is presented in [137]. The approximate K-SVD algorithm with nonnegative constraint is applied to train the PSD dictionary of the clean speech signal. The enhanced speech is obtained by combing the estimated PSD with the signal subspace approach based on the short-time spectral amplitude (SSB-STSA). The above studies have demonstrated that the sparse coding method can improve the quality of noisy speeches. In this work, we further extend the abovementioned methods to apply the sparse coding techniques to the reconstruction of the log power spectrum of speech. It is based on the well-known fact that distortion measures based on the mean-square error of the log-spectra are more appropriate for speech processing [15]. In addition, we propose a new adaptive residual coherence threshold as the stopping criterion. It enables the speech dictionary to adapt to various noise environments in order to improve the enhancement performance. Finally, a modified two-step noise reduction (TSNR) technique with the MMSE-LSA estimator is applied to estimate the clean speech signal. As shown in the simulation results, the proposed algorithm has outstanding performance particularly when the contaminating noises are not totally random but contain certain structure in the frequency domain. Better performance is obtained in all cases evaluated by using different standard measures as compared with the state-of-the-art speech enhancement techniques.

This chapter is organized as follows. In Section 5.2, some formulations used in the traditional two-step noise reduction speech enhancement algorithms are shown. It is followed

by a brief introduction of the traditional framework of sparse coding and dictionary learning for speech enhancement in Section 5.3. The new algorithm is described in Section 5.4. The simulation results are shown in Section 5.5, and conclusions are drawn in Section 5.6.

5.2 Two-step noise reduction

In Chapter 1, we have briefly introduced the essence of some noise reduction techniques. Here we would also like to introduce a two-step noise reduction (TSNR) technique [37] which is used in our proposed algorithm. In the first step of TSNR, the spectral gain $G_{Wiener}(k)$ is computed as in (2.12) and is used to generate the initial guess of the enhanced speech. It is used in the second step to estimate the *a-priori* SNR as follows:

$$\hat{\xi}_{\mathrm{TSNR}}(k,i) = \alpha' \frac{\left|G_{Wiener}(k,i)Y(k,i)\right|^2}{\hat{S}_n(k,i)} + (1-\alpha')\max\left\{\hat{\gamma}\ (k,i+1) - 1, 0\right\} \tag{5.1}$$

In order to avoid the additional processing delay due to the usage of the future frame $(i+1)$, the parameter α' is set to 1. (5.1) becomes

$$\hat{\xi}_{\mathrm{TSNR}}\ (k,i) = \frac{\left|G_{Wiener}(k,i)Y(k,i)\right|^2}{\hat{S}_n(k,i)} \tag{5.2}$$

Finally, the new *a-priori* SNR is used in the Wiener filter gain function as shown below:

$$G_{TSNR}(k) = \frac{\hat{\xi}_{\mathrm{TSNR}}(k)}{1+\hat{\xi}_{\mathrm{TSNR}}(k)} \tag{5.3}$$

The TSNR improves the noise reduction performance because the gain matches to the current frame at all SNR. Experimental result shows that it preserves speech onsets and offsets, and successfully removes the annoying reverberation effect by the decision-directed approach. It will be modified and applied to the proposed algorithm.

5.3 Sparse coding techniques for speech enhancement

The basic principle of sparse coding is that natural signals can be efficiently represented as the linear combinations of pre-specified atom signals in overcomplete dictionaries, where the coefficients are sparse (most of them are zeros or insignificant). Formally, if *y* is a column signal and *D* is the dictionary (whose columns are the atom signals), the sparse coding can be explained using a cardinality constraint:

$$c^* = \arg\ \min_{\mathrm{c}} \|y - \mathrm{Dc}\|_2 \text{ Subject To } \|\mathrm{c}\|_0 \leq K \ll P, \tag{5.4}$$

or using an error constraint

$$c^* = \arg\ \min_{\mathrm{c}} \|\mathrm{c}\|_0 \text{ Subject To } \|y - \mathrm{Dc}\|_2 \leq \varepsilon \tag{5.5}$$

where ε is the error tolerance; $\|\cdot\|_0$ is the L_0 pseudo-norm, which is counting the non-zero entries of a vector; and *K* is the target sparsity. The matrix $D \in \mathbb{R}^{NxP}$ with $N<P$ is called the atoms, which is an overcomplete dictionary usually normalized by the L_2 norm. The vector $c \in \mathbb{R}^P$ is the sparse coefficients of the signal $y \in \mathbb{R}^N$. Since $N < P$, an infinite number of solutions are available for the problem, which is called NP-hard problem. The desired solution can be estimated by using (i) the greedy searching methods such as the matching pursuit (MP), orthogonal MP (OMP) and gradient pursuits (GP); (ii) the nonconvex local optimization methods such as the focal underdetermined system solver (FOCUSS); or (iii) the convex relaxation methods such as the least absolute shrinkage and selection operator (LASSO), LARS and LARC, and others [128], [188]. LARS [189] is a very efficient model selection algorithm that gives a solution closely resembling LASSO (and with a simple modification, it can be made to exactly give the LASSO solution). As with OMP, each iteration of LARS consists of an atom selection and a coding coefficient update step. Atom selection is based on the maximal correlation to the current residual. For the coefficient update step, LARS selects atoms in the equiangular direction, until a new atom has equal

correlation with the residual as all atoms in the active set. The terminate criterion of LARS is based on a coding cardinality or a residual norm value. Based on LARS, the coding algorithm LARC [136] uses a residual coherence threshold as the sparsity parameter. As the LARC method is also a greedy algorithm, the coherent components will be coded before the incoherent components, and the maximum residual coherence will decrease in each iteration. A residual coherence threshold τ_l is used as the stopping criterion, which does not depend on the magnitude of the observation. It is not necessary to adapt the residual coherence threshold to the data on a frame by frame basis in contrast to specifying cardinality or a residual norm. It enables a trade-off between source distortion and source confusion by controlling the coding sparsity.

The above methods are based on the assumption that the speech dictionary is known already. When applying to the speech enhancement problems, a proper speech dictionary needs to be designed. The methods of dictionary learning have been presented in [190]. The K-SVD algorithm [127] is one of the most popular methods for dictionary learning. It has been briefly discussed in Chapter 2. The K-SVD algorithm includes an initial overcomplete dictionary D_0, a set of training signals arranged as the columns of a matrix Y, and the iteration number k. It target iteratively improving the dictionary to achieve the sparse coding of the signal in Y. It is achieved by solving the following optimization problem:

$$\min_{\mathrm{D,C_s}} \left\| \underline{\mathrm{Y}} - \mathrm{DC_s} \right\|_{\mathrm{F}}^2 \quad \text{Subject To } \forall_{\mathrm{i}} \quad \|\mathrm{c_i}\|_0 \leq K \tag{5.6}$$

where $\|\cdot\|_{\mathrm{F}}^2$ denotes the Frobenius norm; c_i is the i-th row of C_s, and K is the desired sparsity. As D and C_s are unknown, the objective function (5.6) is not convex. The K-SVD algorithm solves it by alternating between the sparse coding of the examples based on the current data and a updating of the dictionary atoms. The sparse-coding step is commonly implemented by employing the OMP. When updating the dictionary, one atom will be processed at a time to optimize the target function while keeping the rest fixed. To further reduce the complexity,

the approximate K-SVD algorithm is proposed in [128]. It limits the iteration to be only one but with a different updating step for the dictionary. Such simplified procedure is shown to be able to give very close results to the full computation. More details about the K-SVD algorithm can be found in Chapter 2.

When applying the sparse coding technique to speech enhancement, it is desirable to have the dictionary $D^{(s)}$ trained to be coherent to the speech signal and incoherent to the background noise signal. It can be relatively easily achieved when the background noise is white Gaussian that does not contain any structure. Such background noise is incoherent to any fixed dictionary and in particular to the speech dictionary [67]. However, many relevant kinds of background noise contain structure. If the background noise is partially coherent to the speech dictionary, it will also incur strong coding coefficients which will be very difficult to remove. To solve the problem, it is suggested to train a coherent noise dictionary $D^{(i)}$ for structured background noises. It is shown in [136] that by using a composite dictionary $D = [D^{(s)}\ D^{(i)}]$, significant improved enhancement performance can be achieved comparing with using a single speech dictionary.

5.4 The new framework of sparse coding and dictionary learning of log power spectrum for speech enhancement

In this section, a new speech enhancement algorithm based on sparse coding is proposed. The new algorithm extends the traditional sparse coding based speech enhancement methods in a few aspects: (i) working on log-spectra; (ii) estimating the clean speech with a modified two-step noise reduction (TSNR) procedure; and (iii) using a new noise adaptive stopping criterion for atoms updating.

It is well-known that distortion measures based on the mean-square error of the log-spectra are more appropriate for speech processing [15]. In fact, the log power spectrum

of a signal is related to its log periodogram as in the following formulation [146]:

$$log\left(\tilde{S}_x\ (k)\right) = log\ (S_x(k)) + R(k)\ \text{ where }\ R(k) = \varepsilon(k) - \gamma \tag{5.7}$$

where $\varepsilon(k)$ are *i.i.d.* with zero mean and a fixed variance $\pi^2/6$, and $\gamma \approx 0.577216$ is the Euler's constant. We can also express the log power spectrum of the estimated speech in vector form as follows:

$$\tilde{S}_{log} = \mathrm{D_{log}c_{log}} + R \tag{5.8}$$

where $\tilde{S}_{log} = \log\left(\tilde{\mathrm{S}}_x\right) \in \mathbb{R}^N$ is a column vector, the matrix $\mathrm{D_{log}} = \{\mathrm{d_{log_j}}\}_{\mathrm{j=1}}^{\mathrm{P}} \in \mathbb{R}^{NxP}\ (N < P)$ containing P atoms is an overcomplete dictionary usually normalized by the L_2 norm, the vector $c_{log} \in \mathbb{R}^N$ is the sparse coding coefficients such that

$$\mathrm{S}_{log} = \mathrm{D_{log}c_{log}} \tag{5.9}$$

where $\mathrm{S}_{log} = \log\left(S_x\right)$. Using an error constraint, the sparse coding can be described as the following minimization process:

$$\hat{c}_{log} = \arg\ \min_{\dot{\mathrm{c}}_{\log}} \left\|\dot{\mathrm{c}}_{\log}\right\|_0 \text{ Subject To } \left\|\tilde{S}_{log} - \mathrm{D_{log}}\dot{\mathrm{c}}_{\log}\right\|_2 \leq \varepsilon \tag{5.10}$$

For solving the above minimization problem, we directly adopt the LARC algorithm which is a greedy method consisting of an atom selection and a coding coefficient update step in each iteration. Different from that in [136], the LARC algorithm is applied to reconstruct the log power spectrum of the clean speech. The detailed procedure is shown in Algorithm 1.

Algorithm 1: Batch LARC on log power spectrum

1. Input: $\tilde{S}_x \in \mathbb{R}^N$; $\mathrm{D_{log}} \in \mathbb{R}^{NxP}$; $\mathrm{G}=\mathrm{D_{log}^T D_{log}}$; τ_l
2. $\tilde{S}_{log} \leftarrow \log(\tilde{S}_x)$
3. Normalize $x_N \leftarrow \{\tilde{S}_{log} - mean(\tilde{S}_{log})\}/std(\tilde{S}_{log})$
4. Initialize coefficient vector $\hat{c}_{\log}^{(0)} \leftarrow 0$ and fitted vector $y_N{}^{(0)} \leftarrow 0$
5. Initialize active set $\mathrm{A} \leftarrow \{\ \}$; number of active set $K_A \leftarrow 0$
6. $\mu^{(x)} \leftarrow \mathrm{D_{log}^T} x_N$; $\mu^{(y)} \leftarrow 0$
7. While $|\mathrm{A}| < D_{\log}$ do
8. $\mu \leftarrow \mu^{(x)} - \mu^{(y)}$
9. $j^* \leftarrow \arg\max_j \left|\mu_j\right|$, $j \in \mathrm{A^c}$
10. $\mathrm{A} \leftarrow \mathrm{A} \cup \{j^*\}$; $K_A \leftarrow K_A + 1$
11. if $\left|\mu_{j^*}\right| / \|x_N - y_N\|_2 < \tau_l$ then break
12. $\mathrm{s} \leftarrow \mathrm{sign}(\mu_A)$
13. $\mathrm{g} \leftarrow \mathrm{G}_{(A,A)}^{-1}\mathrm{s}$
14. $\mathrm{b} \leftarrow (\mathrm{g^T s})^{-1/2}$
15. $\mathrm{w} \leftarrow \mathrm{bg}$
16. $\mathrm{u} \leftarrow \mathrm{D}_{\log_{(:,A)}}\mathrm{w}$
17. $\mathrm{a} \leftarrow \mathrm{G}_{(:,A)}\mathrm{w}$
18. Calculate step length $\varphi \leftarrow \min_{e\in A^c}^{+}\left(\frac{|\mu_{j^*}|-\mu_e}{b-a_e}, \frac{|\mu_{j^*}|+\mu_e}{b+a_e}\right)$
19. Update fitted vector $y_N \leftarrow y_N + \varphi u$
20. Update regression coefficients $\hat{c}_{log_A} \leftarrow \hat{c}_{log_A} + \varphi w$
21. $\mu^{(y)} \leftarrow \mu^{(y)} + \varphi a$
22. End while
23. Output coefficients: $\hat{c}_{\log} \in \mathbb{R}^P$, K_A

The above procedure shows that if we are given the periodogram of the clean speech, we can use Algorithm 1 to obtain a good estimate of the sparse coding coefficients of its true power spectrum based on the speech dictionary *D*. The problem is how to obtain the periodogram of the clean speech in the first place. To do so, we propose a modified TSNR procedure which is similar to the EM algorithm we proposed in Chapter 4. First, we make a rough estimation of the periodogram of the clean speech from the noisy observation using a traditional speech enhancement algorithm. Then based on Algorithm 1, we can have the initial estimate of the true power spectrum of the clean speech. We then use it to compute the *a-priori* SNR and in turn the MMSE-LSA gain function to refine our estimation of the clean

speech periodogram. Finally the enhanced speech is obtained.

More specifically, we adopt the temporal cepstrum smoothing (TCS) method to obtain the initial estimate of the clean speech periodogram. As explained in Chapter 4, the TCS method [116] works reasonably well for non-stationary noises and can give an acceptably good estimation of $\tilde{\mathrm{S}}_x$ from the noisy $\tilde{\mathrm{S}}_y$ without the need of a very accurate *a-priori* SNR estimator. Thus we adopt the TCS method to obtain the initial guess of $\tilde{\mathrm{S}}_x$ for the proposed algorithm. Then, the sparse coefficient vector $\hat{c}_{\log}$ is estimated by Algorithm 1. By substituting $\hat{c}_{\log}$ into (5.9), the log true power spectrum of the clean speech is estimated as follows:

$$\hat{S}_{log} = \mathrm{D}_{\log}\hat{c}_{\log} \tag{5.11}$$

Since the observed log periodogram is normalized in step 3 of Algorithm 1, a denormalization procedure should be performed to obtain the estimated log power spectrum of the clean speech as follows:

$$\bar{\mathrm{S}}_{log} = \hat{S}_{log} * std(\hat{S}_{log}) + mean(\hat{S}_{log}) \tag{5.12}$$

The enhanced speech $\tilde{X}$ can thus be obtained by using the modified TSNR gain function in which the *a-priori* SNR is computed based on the current estimate $\bar{\mathrm{S}}_{\log}$. The MMSE-LSA gain function theoretically gives the minimum mean square error estimation of the log-magnitude spectra of speeches. More specifically, we refine our estimate of the *a-priori* SNR in the first step as follows:

$$\hat{\xi}_k^{\mathrm{TSL}}(k) = \frac{\left|G_{fs}(k)Y(k)\right|^2}{\hat{S}_n(k)} \tag{5.13}$$

where $G_{fs}(k) = G_{logmmse}(\hat{\gamma}(k), \tilde{\xi}^{fs}(k))$ and $\tilde{\xi}^{fs} = \frac{\exp(\bar{S}_{log})}{\hat{S}_n}$. Then, the MMSE-LSA gain function is computed and applied to obtain the clean speech estimate as follows:

$$G_{SRTSL}(k)=\frac{\hat{\xi}^{\mathrm{TSL}}(k)}{1+\hat{\xi}^{\mathrm{TSL}}(k)}\exp\left\{\frac{1}{2}\int_{\nu_{TSL}(k)}^{\infty}\frac{e^{-t}}{t}dt\right\}\gamma(k) \quad \text{where} \quad \nu_{TSL}(k)=\frac{\hat{\xi}^{\mathrm{TSL}}(k)}{1+\hat{\xi}^{\mathrm{TSL}}(k)}\gamma(k) \tag{5.14}$$

The enhanced speech $\tilde{X}$ can thus be obtained by,

$$\begin{aligned}\left|\tilde{X}\right| &= G_{SRTSL}(k)\cdot|Y| \\ \tilde{X} &= |X|\exp\left(j\angle Y\right)\end{aligned} \tag{5.15}$$

where $\angle Y$ is the phase angle of Y.

As mentioned above, the dictionary $\mathrm{D_{log}}$ needs to be obtained before the sparse reconstruction stage. The training of the dictionary $\mathrm{D_{log}}$ can be carried out by solving the following optimization problem:

$$\min_{\mathrm{D,C_{log}}}\left\|\mathrm{S}_{log}-\mathrm{D_{log}C_{log}}\right\|_{\mathrm{F}}^{2} \quad \text{Subject To } \forall_{\mathrm{i}} \quad \left\|\mathrm{c_{log_i}}\right\|_{0}\leq K \tag{5.16}$$

where $\|\cdot\|_{\mathrm{F}}^{2}$ denotes the Frobenius norm; c_{log_i} is the *i*-th row of C_{log}, and K is the desired sparsity. S_{log} is the log true power spectrum of the clean speeches. In practice, we approximate them by their log periodogram. The approximate K-SVD method is directly adopted to train the dictionary $\mathrm{D_{log}}$. The complete algorithm is shown in Algorithm 2.

Algorithm 2: Approximate K-SVD on log power spectra

1. Input: Signal set $S_x \in \mathbb{R}^{NxframeNum}$;
   initial dictionary $D_{\log_0} \in \mathbb{R}^{NxP}$;
   target sparsity *K*, and number of iterations *n*
2. Set $S_{log} \leftarrow \log(S_x)$; $D_{\log} \leftarrow D_{\log_0}$
3. Normalize $X_N \leftarrow \{S_{log} - mean(S_{log})\}/std(S_{log})$
4. For *iter =1* to *n* do
5. $\forall_i: C_{\log_i} \leftarrow \arg\min_{c_{\log}} \|X_{Ni} - D_{\log} C_{\log}\|_2^2$ Subject To $\|c_{\log}\|_0 \leq K$
6. For *j=1* to *P* do
7. $D_{\log_{(:,j)}} \leftarrow 0$
8. $I \leftarrow \{\text{indexs of the signals in } S_{log} \textit{ whose representations use } d_j\}$
9. $g \leftarrow C_{\log_{j,I}}^T$
10. $d \leftarrow X_{N_I} g - D_{\log} C_{\log_I} g$
11. $d \leftarrow d/\|d\|_2$
12. $g \leftarrow X_{N_I}^T d - (D_{\log} C_{\log_I})^T d$
13. $D_{\log_{(:,j)}} \leftarrow d$
14. $C_{\log_{j,I}} \leftarrow g^T$
15. end for
16. end for
17. Output Dictionary $D_{\log} \in \mathbb{R}^{NxP}$

5.4.1 Adaptive residual coherence threshold

Recall that for the LARC algorithm as described in Algorithm 1, a residual coherence threshold τ_l is used to define the stopping criterion of the iterative sparse coding process (see step 11 in Algorithm 1). Basically its selection is not critical if the background noise is incoherent to the speech dictionary (such as white noise) [136]. Originally it is the case when applying to the proposed algorithm since the true power spectrum of a speech is different from its periodogram by an *i.i.d.* error function (see Eqn.(5.7)). Nevertheless, since the initial estimate of $\tilde{S}_x$ actually comes from the TCS. Some of the background noise, which can be structural, may still remain in the initial estimate of $\tilde{S}_x$. Such residual background noise components will thus be coherent with the speech dictionary. The selection of τ_l becomes

critical in this case as it can significantly affect the final speech enhancement performance. To illustrate this, a comparison of the enhancement performances of the proposed speech enhancement algorithm with various residual coherence thresholds τ_l is shown in Fig. 5.1. More specifically, we compare the PESQ scores (the Perceptual Evaluation of Speech Quality) of the enhanced speeches generated by the proposed method with different τ_l. PESQ is an ITU standard for evaluating speech quality [58]. Our results show that when the residual coherence thresholds τ_l is too high (e.g. $\tau_l = 0.4$), poor performance is resulted for all types of background noise as distortion occurs during reconstruction. For white noises which are incoherent to the speech dictionary, better results are obtained with lower threshold (e.g. τ_l=0.1) value. For babble noises which are highly coherent to the speech dictionary, higher threshold (e.g. τ_l=0.3) value gives better performance. For buccaneer noises which are partially coherent to the speech dictionary, lower threshold values are more favorable when the input SNR is low (0 and 10 dB). But when the input SNR is higher, a higher threshold value is preferred (e.g. τ_l=0.3). The above shows that the speech enhancement performance is not only affected by the coherence between the background noise and the dictionary, but also the input SNR.

To deal with the problem, the traditional approach uses a composite dictionary which consists of both the speech and background noise dictionary. However, it is difficult to train the background noise dictionary as the noise property is not known in advance for most applications. We propose to use an adaptive residual coherence threshold such that the threshold value can be adjusted automatically for different kinds of noise and also SNRs.

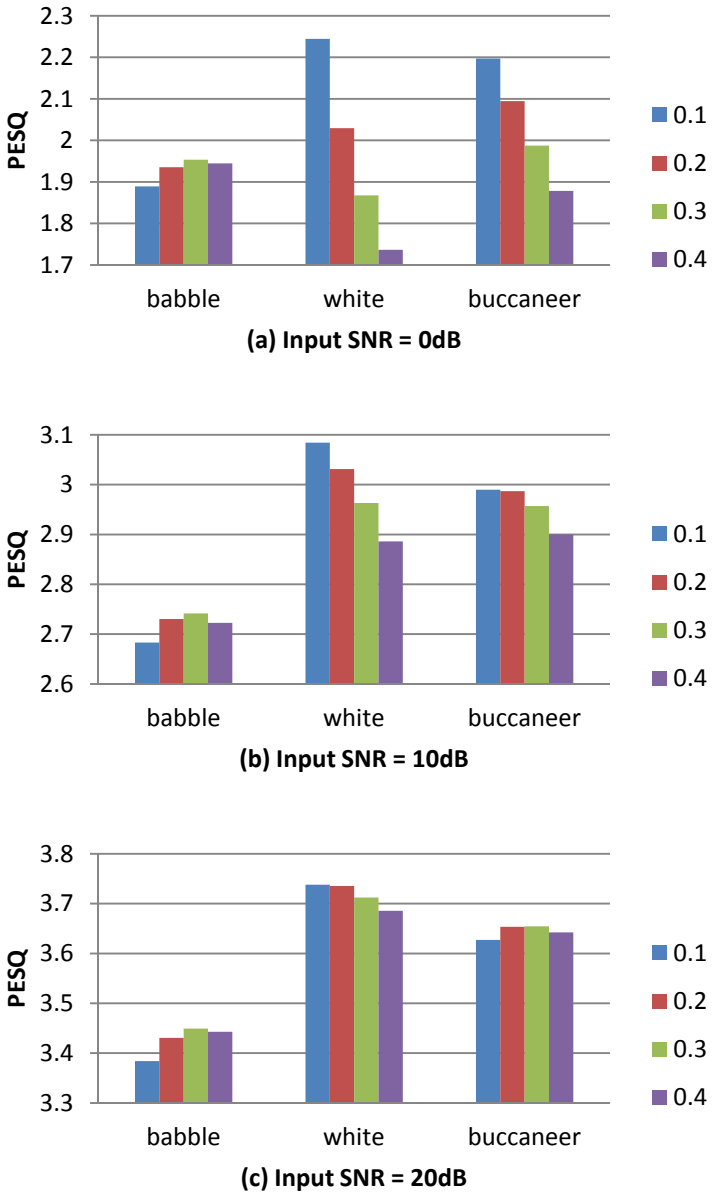


Fig. 5.1 – Enhancement performance of the proposed speech enhancement algorithm SRLPS-TSL with various residual coherence thresholds τ_l = 0.1, 0.2, 0.3 and 0.4.

By closely examining Algorithm 1, we notice that the parameter K_A (see step 5 of Algorithm 1) can give a good indication about the coherence between the input signal and the speech dictionary. So if we input the power spectrum of a noise signal to Algorithm 1, we can use K_A to indicate the coherence between the noise and the speech dictionary. To illustrate this, an experiment was done to investigate the change in the value of K_A when feeding the power spectrum of different kinds of noise to Algorithm 1. Table 5.1 shows the average value of K_A across frames for different kinds of background noise obtained from the NOISEX-92 database [193]. For the background noises which are less coherent to the speech dictionary (such as the white noise), the average K_A is lower. For the background noises which are highly coherent to the speech dictionary (such as the babble noise), the average K_A is also higher. Hence the parameter K_A in Algorithm 1 can be used as an indicator of the coherence between the background noise signal and the speech dictionary. In practice, we first use a VAD [53] to detect the noise frames in a noisy speech signal. Then the power spectrum of the noise is estimated by taking the average of the periodogram of all the detected noise frames. Algorithm 1 is then called with the estimated noise power spectrum to obtain the parameter K_A. We shall use such parameter to evaluate an adaptive residual coherence threshold to be used in the sparse coding process performed in the noisy speech frames.

Table 5.1 - Summary of the average number of active set K_A versus different background noises. ($\boldsymbol{\tau_l}$ = 0.5)

Noise	Average K_A
White	1
Speech babble	14.05
Destroyer engine room	9.45
F16 cockpit	8.99
Pink	3.51
Buccaneer cockpit	6.03
M109 Tank	7.50

As it is mentioned above, the input SNR can also affect the selection of the residual coherence threshold τ_l. However, we notice that its effect is not linear but similar to the

Wiener gain function, i.e. it will flat out when the noise level is sufficiently small. Since the Wiener filter gain function $G_{Wiener}(k)$ in Eqn. (2.1) could be updated in every frame, we propose to choose the mean of the Wiener filter gain function $G_{Wiener}(k)$ as a parameter to compute the residual coherence threshold. More specifically, we define a parameter h_i such that for a noisy speech frame,

$$h_i = \frac{1}{N}\sum_{k=0}^{N-1} G_{Wiener}(k) \tag{5.17}$$

Then, the residual coherence threshold for a specific noise τ_n could be obtained as follows:

$$\tau_n = \min\{\max\{\max\{(K_A - b_1)/b_2, 0\} + b_3 h_i, \tau_{\min}\}, \tau_{\max}\} \tag{5.18}$$

where τ_{min} = 0.1 and τ_{max} = 0.3; three parameters $b_1 = 20$, $b_2 = 25$ *and* $b_3 = 0.5$ are set empirically. Eqn. (5.18) is formulated based on our observation (explained earlier) that the residual coherence threshold τ_n should be proportional to the coherence between the residual background noise and the speech dictionary (represented by the parameter K_A), and the input SNR (represented by the parameter h_i). The parameter b_1 is the offset, b_2 and b_3 are the weighting factors. They are selected by fitting the PESQ results of the proposed algorithm using Eqn. (5.18) over 80 test speeches of different genders, noise types and noise levels (in fact, their selection is not sensitive to these factors). To reduce the fluctuation between frames, we further smooth τ_n by taking a weighted average with the estimate in the previous frame as follows:

$$\tau_a(i) = \alpha_a \tau_a(i-1) + (1-\alpha_a)\tau_n(i) \tag{5.19}$$

where the smoothing factor α_a is set to 0.8. τ_a is thus the proposed adaptive residual coherence threshold. To summarize, the proposed speech enhancement algorithm based on the new framework of sparse reconstruction on log power spectra with two-step MMSE-LSA filtering (SRLPS-TSL) can be described as follows:

1. Use Algorithm 2 to train D_{log} from a set of training clean speeches.

2. For the noisy speech input, compute the initial guess of the log periodogram of the clean speech $\hat{\mathrm{S}}_{log}$ using the TCS method (Eqn. (4.26) in Chapter 4), i.e. $\hat{\mathrm{S}}_{log} = \log(\hat{\mathrm{S}}_{\mathrm{x}}^{\mathrm{TCS}})$.
3. Compute the adaptive residual coherence threshold τ_a by Eqn. (5.19) and Algorithm 1.
4. Estimate the parameter $\hat{\mathrm{c}}_{\log}$ using Algorithm 1 with $\tau_l = \tau_a$.
5. Use Eqn. (5.11) and (5.12) to obtain the estimated log power spectrum $\bar{\mathrm{S}}_{log}$ of the clean speech.
6. First step of TSL - Estimate the *a-priori* SNR $\hat{\xi}^{\mathrm{TSL}}(k)$ by Eqn. (2.19) and (5.13).
7. Second step of TSL - Obtain the enhanced speech spectrum $\tilde{X}$ by Eqn. (5.14) and (5.15).
8. Obtain the enhanced speech signal by IDFT, overlap and add.

The operation of the proposed algorithm is also described in Fig. 5.2.

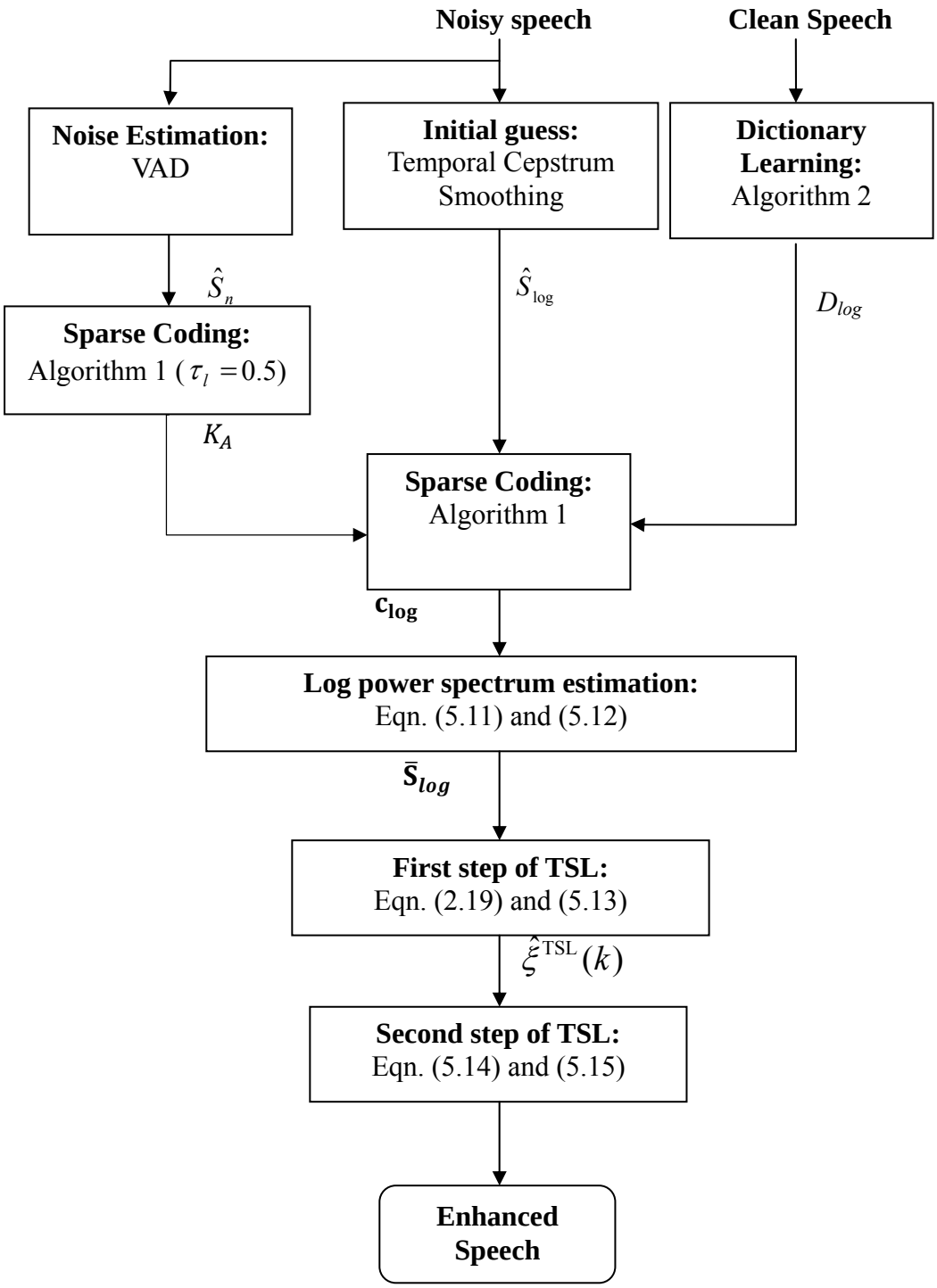


Fig. 5.2 – The operation of the proposed speech enhancement algorithm SRLPS-TSL

5.5 Simulations and Results

In this section, the performance of the proposed algorithm is shown and compared with those achieved by the state-of-the-art speech enhancement methods. To start with, we use an example to illustrate the importance for the proposed algorithm to work on the log power spectra instead of the normal power spectra. Fig. 5.3 shows a segment of a typical noisy speech periodogram (dash-dot line), its original clean speech periodogram (dotted line), the enhanced speech periodogram using the proposed speech enhancement algorithm on power spectrum (PS) (dashed line) and the proposed speech enhancement algorithm on log power spectrum (LPS) (solid line). It can be seen that both the reconstructed periodograms (PS and LPS) can restore the spectral peak of original speech. However, the LPS method restores the spectral valley much better than the PS method. This example demonstrates that it is more appropriate to apply the sparse reconstruction method to log-spectra for speech enhancement.

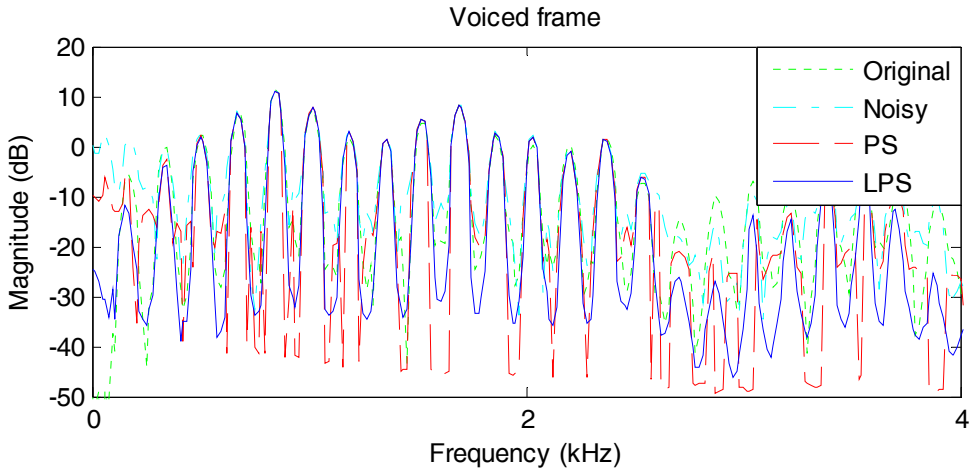


Fig. 5.3 – Effect of the noise reduction with sparse reconstruction and dictionary learning on a voiced frame (pink noise, input segSNR = -3.69dB). Reconstruction on power spectrum (dashed line), Reconstruction on log power spectrum (solid line), noisy speech spectrum (dash-dot line) and original speech spectrum (dotted line).

A comparison of the spectrogram of the enhanced speeches generated using different algorithms is shown in Fig. 5.4. Table 5.2 gives a summary of the algorithms that have been

compared. The speech sampling rate is 16kHz. Simulation details are listed as follows: frame size – 512 samples (~32ms), FFT size – 1024 samples (zeros padded each frame with 512 samples), window shift step size – 128 samples (75% overlap). The log power spectrum dictionary of the clean speech is trained by the approximate K-SVD MATLAB Toolbox [128]. The sparsity target K as mentioned in Eqn. (5.6) is set to 10 and the number of iterations is set to 30. The dictionary D_{log} is learned using 100 sentences extracted from the training portion of the TIMIT database, which is different to the testing portion. No sentence should appear in both the training and testing portions. The size of dictionary D_{log} is 513 x 1024 and is initialized by randomly taking the training data. For all algorithms, the noise power spectrum is estimated by first using the initial frames that are assumed to have no speech energy; then updated whenever a frame is detected to have no speech energy by using a VAD [53]. Fig. 5.4a shows the clean speech spectrogram of a female speech selected from the TIMIT database [192] saying the following sentence: "Cliff was soothed by the luxurious massage". Fig. 5.4b shows the result when the the military vehicle (Leopard) noise (from the NOISEX-92 database [193]) is added to the speech with input segSNR about -5.79 dB. Fig. 5.4c depicts the spectrogram using the traditional MMSE-LSA method. It can be seen that although some of the background noise is suppressed but the speech signal is distorted. Fig. 5.4d shows the spectrogram using the Harmonic Regeneration Noise Reduction Algorithm (HRNR) [37]. Although it can recover much speech content, its noise control is not sufficient and strong background residue noise remains in the enhanced speech. Fig. 5.4e shows the spectrogram given by the MMSE-LSA CEM algorithm [182] that we proposed in Chapter 4. We have seen that the MMSE-LSA CEM algorithm performs very well for buccaneer noises in Chapter 4. However, it is not the case for the Leopard noises as shown in Fig. 5.4e. Much residual noise remains at the low frequency part of the spectrum. It may be due to the strong spectral coefficients of the Leopard noise at low frequencies, where strong speech spectral

coefficients can also be found. They are then mixed up in the MMSE-LSA CEM algorithm. Fig. 5.4f shows the spectrogram using the proposed algorithm (SRLPS-TSL). It has very well background noise control (as indicated in the circled areas) and the speech content is also preserved. It can be seen that many of the low frequency noise coefficients are suppressed while those of the speech remain intact. The result demonstrates the ability of the proposed sparse coding method in differentiating the speech and noise coefficients.

Table 5.2 - Summary of the algorithms compared in the simulations.

Method	**Description**
MMSE-LSA	Minimum mean-square error log-spectral amplitude estimator [15]
HRNR	Harmonic regeneration noise reduction algorithm [37]
MMSE-LSA CEM	EM based cepstrum smoothing using MMSE-LSA filter [182]
SRLPS-TSL	The proposed algorithm based on the sparse reconstruction on log power spectra with two-step MMSE-LSA filtering

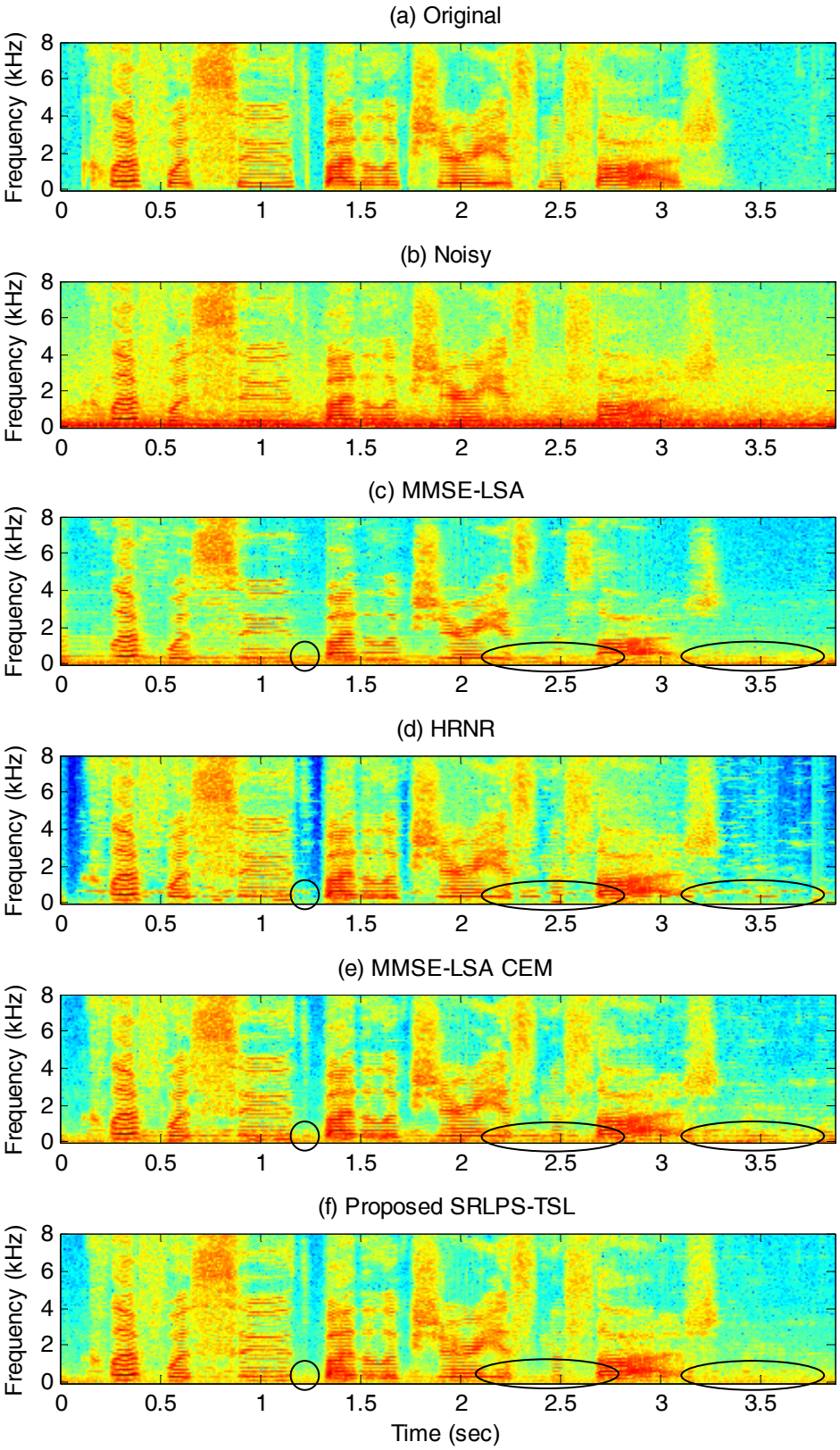


Fig. 5.4 – Spectrogram of the original, noisy and enhanced speeches (Leopard noise, segSNR = -5.79 dB)

The performance of the proposed SRLPS-TSL algorithm is further evaluated using the standard objective evaluation measures and compared with the following approaches: MMSE-LSA [15], Two Step Noise Reduction Algorithm (TSNR) [37], HRNR [37], MMSE-LSA plus SPP using TCS method [140], and the MMSE-LSA CEM [182] that we proposed in Chapter 4.

In the simulation, 40 male and 40 female test speeches were arbitrarily selected from the TIMIT database [192]. The noise signals were adopted from the NOISEX-92 database [193] and added to the speeches with input signal-to-noise ratio ranging from about 0dB to +20dB. To predict the quality of the enhanced speeches, three composite objective metrics [1] are used, which include (a) C_{sig}: Signal distortion (SIG) formed by linearly combining the LLR, PESQ, and WSS measures; (b) C_{bak}: Noise distortion (BAK) formed by linearly combining the segSNR, PESQ, and WSS measures; and (c) C_{ovl}: Overall quality (OVL) formed by linearly combining the PESQ, LLR, and WSS measures.

Table 5.3 lists the performance of 6 different speech enhancement algorithms in the cases of 6 different kinds of background noise, both stationary and non-stationary. As can be seen in the table, the proposed SRLPS-TSL algorithm always outperforms the other 5; and in many cases, the improvement is significant. Its performance is consistent for speeches made by people of different genders and at different noise levels. These results have demonstrated the robustness of the proposed algorithm.

Table 5.3 - Composite measurement comparison of different algorithms.

Noise	Method \ Input SNR	C_{sig}					C_{bak}					C_{ovl}				
		0	5	10	15	20	0	5	10	15	20	0	5	10	15	20
White	Noisy	1.34	1.92	2.61	3.26	3.87	1.72	2.17	2.64	3.12	3.59	1.35	1.87	2.42	2.95	3.45
	MMSE-LSA	1.84	2.54	3.16	3.67	4.05	2.18	2.59	2.97	3.28	3.51	1.80	2.37	2.86	3.26	3.55
	TSNR	1.48	2.10	2.56	3.00	3.52	2.11	2.58	3.02	3.45	3.90	1.57	2.12	2.55	2.98	3.44
	HRNR	1.94	2.50	3.06	3.62	4.17	2.28	2.72	3.17	3.63	4.08	1.83	2.33	2.84	3.35	3.85
	MMSE-LSA TCS SPP	2.11	2.91	3.50	3.96	4.33	2.41	2.87	3.29	3.69	4.08	2.05	2.70	3.21	3.62	3.98
	MMSE-LSA CEM	**2.47**	3.17	3.71	4.17	4.56	**2.62**	**3.05**	3.43	3.82	4.23	**2.35**	2.94	3.38	3.79	4.17
	Proposed SRLPS-TSL	2.43	**3.20**	**3.79**	**4.25**	**4.64**	2.57	3.03	**3.44**	**3.84**	**4.24**	2.30	**2.94**	**3.44**	**3.85**	**4.22**
Speech babble	Noisy	2.52	3.09	3.63	4.10	4.50	1.68	2.15	2.64	3.13	3.61	1.96	2.47	2.95	3.39	3.78
	MMSE-LSA	2.53	3.11	3.60	3.97	4.22	1.87	2.35	2.79	3.16	3.44	2.04	2.56	3.01	3.36	3.59
	TSNR	1.67	2.51	3.23	3.82	4.32	1.46	2.06	2.64	3.19	3.72	1.37	2.09	2.72	3.27	3.74
	HRNR	1.86	2.73	3.46	4.06	4.52	1.46	2.09	2.70	3.26	3.78	1.48	2.23	2.88	3.43	3.88
	MMSE-LSA TCS SPP	2.21	2.94	3.56	4.07	4.49	1.86	2.39	2.90	3.39	3.85	1.84	2.47	3.03	3.50	3.91
	MMSE-LSA CEM	2.41	3.09	3.69	4.19	4.59	1.95	2.47	2.96	3.46	3.92	2.00	2.59	3.12	3.59	3.99
	Proposed SRLPS-TSL	**2.58**	**3.24**	**3.81**	**4.29**	**4.67**	**2.11**	**2.60**	**3.07**	**3.54**	**3.98**	**2.15**	**2.73**	**3.23**	**3.69**	**4.06**
Destroyer engine room	Noisy	2.00	2.59	3.19	3.75	4.25	1.52	1.98	2.46	2.96	3.47	1.65	2.14	2.64	3.12	3.57
	MMSE-LSA	2.45	3.02	3.51	3.89	4.15	1.96	2.39	2.79	3.14	3.41	2.04	2.53	2.95	3.29	3.55
	TSNR	1.89	2.48	3.05	3.64	4.19	2.03	2.48	2.89	3.34	3.81	1.75	2.28	2.76	3.26	3.74
	HRNR	2.20	2.92	3.56	4.12	4.60	2.09	2.60	3.05	3.50	3.96	1.90	2.54	3.10	3.59	4.03
	MMSE-LSA TCS SPP	2.57	3.20	3.67	4.09	4.49	2.19	2.71	3.16	3.59	4.00	2.19	2.78	3.24	3.64	4.01
	MMSE-LSA CEM	3.00	3.55	3.97	4.36	4.72	2.49	2.94	3.35	3.74	4.16	2.57	3.08	3.48	3.85	4.20
	Proposed SRLPS-TSL	**3.10**	**3.66**	**4.09**	**4.47**	**4.81**	**2.50**	**2.97**	**3.40**	**3.81**	**4.21**	**2.63**	**3.15**	**3.58**	**3.95**	**4.28**
F16 cockpit	Noisy	1.95	2.58	3.20	3.77	4.28	1.57	2.05	2.54	3.05	3.55	1.63	2.17	2.70	3.20	3.66
	MMSE-LSA	2.45	3.05	3.56	3.94	4.20	2.03	2.47	2.87	3.21	3.46	2.09	2.60	3.04	3.37	3.61
	TSNR	1.67	2.31	2.92	3.53	4.13	1.90	2.42	2.92	3.39	3.87	1.56	2.16	2.71	3.23	3.75
	HRNR	2.04	2.79	3.48	4.10	4.59	2.00	2.54	3.06	3.54	3.99	1.75	2.44	3.07	3.61	4.05
	MMSE-LSA TCS SPP	2.56	3.18	3.67	4.07	4.48	2.21	2.72	3.18	3.61	4.04	2.21	2.79	3.26	3.65	4.03
	MMSE-LSA CEM	2.89	3.45	3.88	4.30	4.68	2.43	2.89	3.32	3.74	4.16	2.50	3.01	3.43	3.82	4.19
	Proposed SRLPS-TSL	**2.99**	**3.56**	**4.00**	**4.39**	**4.75**	**2.46**	**2.93**	**3.35**	**3.77**	**4.18**	**2.57**	**3.09**	**3.52**	**3.90**	**4.25**
Leopard (Military vehicle)	Noisy	3.39	3.83	4.22	4.56	4.86	1.99	2.42	2.87	3.32	3.80	2.66	3.08	3.45	3.80	4.14
	MMSE-LSA	3.41	3.67	3.88	4.03	4.13	2.63	2.93	3.18	3.37	3.52	2.88	3.16	3.38	3.55	3.66
	TSNR	3.55	4.07	4.48	4.80	4.98	2.76	3.21	3.63	4.01	4.38	3.03	3.52	3.91	4.23	4.49
	HRNR	3.72	4.25	4.66	4.94	5.00	2.84	3.29	3.71	4.09	4.43	3.15	3.63	4.03	4.35	4.59
	MMSE-LSA TCS SPP	3.82	4.29	4.66	4.92	4.99	2.90	3.33	3.73	4.08	4.42	3.25	3.70	4.07	4.35	4.56
	MMSE-LSA CEM	3.92	4.36	4.71	4.94	5.00	2.89	3.33	3.73	4.10	4.46	3.29	3.73	4.09	4.38	4.60
	Proposed SRLPS-TSL	**4.16**	**4.52**	**4.81**	**4.97**	**5.00**	**3.13**	**3.50**	**3.85**	**4.17**	**4.49**	**3.53**	**3.90**	**4.20**	**4.44**	**4.63**
M109 (Tank)	Noisy	2.66	3.24	3.78	4.26	4.68	1.85	2.31	2.78	3.26	3.75	2.23	2.73	3.20	3.63	4.03
	MMSE-LSA	3.05	3.46	3.77	3.98	4.11	2.49	2.85	3.15	3.38	3.53	2.66	3.03	3.31	3.51	3.63
	TSNR	2.36	3.21	3.91	4.44	4.83	2.42	2.95	3.47	3.95	4.37	2.26	2.94	3.52	3.97	4.33
	HRNR	2.74	3.57	4.22	4.69	4.95	2.54	3.06	3.54	3.97	4.35	2.49	3.18	3.73	4.14	4.45
	MMSE-LSA TCS SPP	3.13	3.71	4.23	4.66	4.94	2.62	3.10	3.57	4.02	4.42	2.77	3.30	3.76	4.15	4.45
	MMSE-LSA CEM	3.37	3.93	4.41	4.80	4.98	2.76	3.22	3.67	4.10	4.49	2.95	3.45	3.89	4.26	4.54
	Proposed SRLPS-TSL	**3.55**	**4.08**	**4.53**	**4.87**	**4.99**	**2.85**	**3.30**	**3.73**	**4.14**	**4.52**	**3.09**	**3.57**	**3.99**	**4.32**	**4.58**

5.6 Chapter Summary

In this chapter, an improved speech enhancement algorithm based on the sparse coding on log-spectra is proposed. The proposed algorithm starts with the traditional cepstrum smoothing method which gives the initial guess of the log periodogram of the clean speech. The sparse coding is carried out by using the LARC algorithm with the speech dictionary trained using the K-SVD method. We improve the LARC algorithm by introducing a noise adaptive residual coherence threshold so that the stopping criterion will be adaptive to the noise type and the input SNR. The improved LARC algorithm gives a good estimate of the sparse coding coefficients of the clean speech's log power spectrum. Combining with the TSL method, an enhanced speech is obtained. The proposed algorithm does not only fully exploit the sparsity of speeches through the use of the sparse speech dictionary, but also reduces the confusion in the sparse coding process due to the coherence between the noise and the speech. As a result, the proposed algorithm works particularly well when the input speech is contaminated by noises that contain structure (they include most colored noises). Besides, due to the TSL process, the proposed algorithm has very good control of the residual background noises. They make it outperform the traditional methods. Simulation results have verified that the proposed algorithm improves over the traditional speech enhancement methods in almost all testing conditions using different evaluation measures. They have clearly shown the robustness of the proposed algorithms in general speech enhancement applications.

Chapter 6 Conclusions

In this thesis, we have investigated three different sparse representation methods for speech enhancement, namely, the discrete wavelet transform, the cepstral transform and the sparse coding based on dictionary learning. For each method, a new algorithm is proposed. In this chapter, we draw the conclusions of these works and suggest possible future works.

6.1 General Conclusions

In Chapter 3, we proposed a new algorithm for the estimation of speech presence probability (SPP) of a noisy speech signal based on the discrete wavelet transform. Although it is known that a good estimator of SPP can be obtained by smoothing the observed noisy speech power spectrum before using it in the estimation process, care must be taken to ensure the smoothing operation will not destroy the spectral peaks which are important to the intelligibility of the enhanced speech. The major contribution of this work is two-folded. First, we successfully developed a two-stage wavelet denoising algorithm that effectively removes the noise while preserving the spectral peaks in a noisy speech power spectrum. It outperforms the traditional approaches by combining the information of noise and spectral peaks in both the periodogram and the log MTS of a noisy speech. The denoised speech power spectrum in turn lets us generate a smooth *a-posteriori* SNR function. Second, we proposed a new method for estimating the generalized likelihood ratio (GLR). It is by directly estimating the *PDF* of the *a-posteriori* SNR under the hypothesis H_0, i.e. speech is absent, using the data in different noise frames. It simplifies the estimation process and avoids the use of many empirically selected parameters in traditional approaches. The new SPP estimator was then applied to the MMSE-LSA speech enhancement algorithm. Compared

with the traditional SPP estimators, up to 15% improvement was noted for different noises at different noise levels when measuring using the standard composite objective measures. When inspecting the spectrogram of the enhanced speeches using different approaches, the proposed algorithm in general preserves much better the speech contents while effectively removing the background noise.

While the proposed algorithm in Chapter 3 successfully makes use of the discrete wavelet transform to detect the spectral peaks of speeches, problem may arise for certain kinds of noise which also have spectral peaks similar to those of speeches. Further effort is needed to differentiate the speech from noise in a noisy speech. In Chapter 4, we presented a novel expectation-maximization (EM) framework for speech enhancement algorithm. Based on the sparsity of speeches in the cepstral domain, the new algorithm makes use of the EM algorithm to define a theoretical framework for the estimation of the true power spectrum of the original speech and its periodogram from a noisy observation. The proposed algorithm starts with the traditional cepstrum smoothing method which gives the initial guess of the periodogram of the clean speech. It is applied to an L_1 norm regularizer in the M-step of the EM framework to estimate the cepstral coefficients of the true speech power spectrum. It enables the estimation of the *a-priori* SNR and is used in the E-step, which is indeed an MMSE-LSA gain function, to refine the estimate of the clean speech periodogram. The M-step and E-step then iterate for 2 more times, with which we have shown to be sufficient in most cases to achieve good result. The proposed algorithm fully utilizes the sparsity of speeches in the cepstral domain by adopting the L_1 norm regularizer. It enables the optimization process to be carried out on coefficients with improved SNR and hence reduces the effect due to the estimation error of the non-stationary noise characteristics. As a result, the proposed algorithm works particularly well when the input speech is contaminated by non-stationary noises. Besides, due to the iterative process, the proposed algorithm has very good control of the residue background

noises which makes it outperform the traditional methods. Simulation results have verified that the proposed algorithm improves over the competing speech enhancement methods in almost all testing conditions, such as different kinds of noise at different noise levels using different evaluation measures

It is shown in Chapter 4 that the knowledge of the sparse representation of speeches is one of the key factors when developing a speech enhancement algorithm. A question naturally arises if the cepstral representation that is used in Chapter 4 is the best as far as the sparsity is concerned. In Chapter 5, we investigated using the sparse coding technique with the dictionary learnt from a speech database. We further propose an improved speech enhancement algorithm based on the framework of sparse reconstruction on log power spectrum with two-step MMSE-LSA filtering (SRLPS-TSL). The new algorithm makes use of sparse coding technique to efficiently estimate the true log power spectrum of the clean speech from a noisy observation. Similar to that in Chapter 4, the proposed algorithm starts with the traditional temporal cepstrum smoothing method which gives the initial guess of the periodogram of the clean speech. The batch LARC algorithm is then used to perform the sparse coding with the speech dictionary trained by the approximate K-SVD method. We improve the batch LARC algorithm by introducing a noise adaptive residual coherence threshold that allows the sparse coding process to be adaptive to the noise types and the input SNR. Combining with the TSL method, the enhanced speech is obtained. The proposed algorithm does not only fully exploit the sparsity of speeches on log-spectra but also reduce the confusion in the sparse coding process due to the coherence of the residual noise and the speech dictionary. As a result, the proposed algorithm works particularly well when the input speech is contaminated by noises which contain structure (it is the case for most colored noises). Besides, due to the TSL process, the proposed algorithm has very good control of the residual background noises which makes it outperform the traditional methods. Compared

with the traditional speech enhancement algorithms, significant improvement is noted for different kinds of noise at different noise levels when evaluating using the three composite objective metrics. Informal subjective listening tests also indicate that the proposed algorithm is generally more preferred to other competing methods. They have clearly demonstrated the robustness of the proposed algorithms in general speech enhancement applications. When comparing with the proposed Logmmse-L1-EM method in Chapter 4, the proposed SRLPS-TSL algorithm works particularly well when the input speech is contaminated by noises that contain structure, although the Logmmse-L1-EM algorithm works well when the input speech is contaminated by noise that is incoherent to the speech. We notice that the proposed method based on the sparse coding technique can better identify speech spectral components from those of noises, although a higher computational complexity may incur due to the iterative sparse coding process.

In conclusion, we have demonstrated in this thesis that the acquisition of an efficient sparse representation of speeches is one of the key factors to the success of speech enhancement. In this work, a number of sparse representation methods have been employed that enhance a noisy speech in different environments. We have shown that the proposed algorithms outperform the traditional methods when evaluating by different objective measures and informal subjective listening tests. We believe that the results obtained in this work have contributed significantly to the field of study.

6.2 Future Works

There are several potential future works to be explored in the area of speech enhancement. First of all, we believe that the methods for enhancing the unvoiced speech deserve further consideration. It is known that on average over 20% of a normal speech is

unvoiced. However, both the cepstral method and the MMSE method may not be the best candidates for their enhancement. It is because unvoiced speeches do not have a harmonic structure and they have a noise-like statistical property which can easily be classified as noises by the MMSE based methods and are thus suppressed. A different algorithm is needed to work together with the proposed algorithm in order to take care of both the voiced and unvoiced parts of a speech in an enhancement process.

Another important direction is the development of algorithms for fast and accurate tracking of noise power spectral density (PSD), which is another important parameter of most speech enhancement algorithms. As mentioned in Section 1.5, the noise PSD has to be estimated from the noisy speech because it is unknown in advance in most applications. Traditional approaches (e.g. minimum statistics, VAD, etc.) can give a good estimate of the noise PSD for stationary noises. However, their performance becomes unsatisfactory when the noise source tends to change rapidly. While there were works from time-to-time claimed to provide accurate PSD estimate for non-stationary noises (such as [194] and [195]), their effectiveness remains to be verified. We believe a good noise PSD estimator will be another key factor to the success of speech enhancement.

Another interesting direction for future development is the estimation of the clean speech phase. Although it is assumed that the enhancement of the noisy spectral amplitude is more important than the enhancement of the spectral phase [25], it was recently shown [196] that employing the clean speech phase can further improve noise reduction algorithms with a more redundant spectral representation in the frequency domain. From that work, it follows that we can further improve the proposed speech enhancement algorithms if we have a good estimate of the clean speech phase.

As to the applications of the speech enhancement algorithms, we suggest that further investigations can be made on their applications in biomedical diagnostic systems. While

speeches are some natural signals generated by human being, they also carry information about the health condition of the human subject. However, due to the various constraints in the speech acquisition process (some of them are clinical constraints), the speeches obtained are contaminated with different kinds of noise such that a speech enhancement algorithm that tailors for the clinical environments will greatly enhance the diagnostic results. One of the possible applications is in the diagnosis of the obstructive sleep apnea (OSA). It is known that sound recorded during sleep (i.e. the sound of snoring) has been used to diagnose OSA. It is characterized by repetitive complete (apnea) or partial (htpopnea) cessation of breathing during sleep for at least 10 seconds as a consequence of complete or partial collapse of the upper airway, respectively. However, additive background acoustical noises received during the acquisition process can degrade the quality of the recorded snoring sound and ends up with inaccurate diagnostic results. In order to improve the recorded signal quality, it is shown in [198] that the background noise embedded in the acquired signals could be effectively removed by a translation-invariant wavelet denoising scheme. Since the snoring sound is also a structured human generated signal similar to speeches, it is possible that the snoring sound will also have a sparse representation when analyzing by a suitably chosen dictionary. Therefore, a more advanced snoring activity detector can be developed based on the enhancement methods proposed in Chapter 3 to Chapter 5.

References

[1] P.C. Loizou, *Speech enhancement: theory and practice*, CRC Press, 2007.

[2] L.R. Rabiner and R.W. Schafer, *Theory and Application of Digital Speech Processing,* 2011.

[3] B. Gold, N. Morgan and D. Ellis. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Wiley, Berkeley, California., 2011

[4] J. S. Lim and A. V. Oppenheim, "Enhancement and band-width compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586 -1604, 1979.

[5] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp.197 -210. 1978.

[6] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition", *IEEE Trans. Signal Processing*, vol. 39, pp.795 -805, 1991.

[7] M. Dendrinos, S. Bakamidis, and G. Garayannis, "Speech enhancement from noise: A regenerative approach", *Speech Commun.*, vol. 10, pp.45 -57, 1991.

[8] Y. Ephraim and H.L.V. Trees, "A Signal Subspace Approach for Speech Enhancement", *IEEE Trans. Speech and Audio Process.*, vol.3, no.4, pp.251-266, Jul. 1995.

[9] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise", *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp.104 -106 2003.

[10] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement", *IEEE Trans. Speech and Audio Process.*, vol.11, no.6, pp.700-708, 2003.

[11] M. R. Weiss, E. Aschkenasy and T. W. Parsons, “Study and development of the INTEL technique for improving speech intelligibility”, Technical Report NSC-FR/4023, Nicolet Scientific Corporation, Northvale, NJ, 1974.

[12] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp.113 - 120, 1979.

[13] R. J. McAulay and N. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp.137 - 145, 1980.

[14] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error short time spectral amplitude estimator”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp.1109-1121, Dec. 1984.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Trans. Acoust., Speech and Signal Process.*, vol.33, no.2, pp.443-445, Apr. 1985.

[16] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors", *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845-856, 2005.

[17] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model", EURASIP J. Appl. Signal Process., vol. 7, pp. 1110-1126, 2005.

[18] C. H. You, S. N. Koh and R. Susanto, “β-order MMSE spectral amplitude estimation for speech enhancement”, *IEEE Trans. Speech Audio Process*., vol. 13, no. 4, pp. 457-486, 2005.

[19] I.Y. Soon and S.N. Koh, "Speech Enhancement Using Two Dimensional Fourier Transform", *IEEE Trans. Speech and Audio Processing*, vol 11, No. 6, pp 717-724, Nov 2003

[20] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression", *IEEE Trans. Speech Audio Processing*, vol. 5, pp.479 -514, 1997.

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

[22] M. D. Plumbley , T. Blumensath , L. Daudet , R. Gribonval and M. Davies "Sparse representations in audio and music: From coding to source separation", *Proc. IEEE*, vol. 98, no. 6, pp.995 -1005 2010

[23] M. Elad and M. Aharon, “Image denoising via sparse redundant representations over learned dictionaries,” *IEEE Trans. Image Process*., vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[24] Ö. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process*., vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[25] D.L. Wang and J.S. Lim, "The Unimportance of Phase in Speech Enhancement", *IEEE Trans. On Acoustics, Speech, and Signal Processing*, vol. 30, no.4, pp. 679-681, Aug. 1982.

[26] K.K. Paliwal and L.D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153-170, Feb. 2005.

[27] R. E. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis", IEEE Trans. Acoust., Speech, Signal Processing, vol. 28, pp.99 - 102, 1980

[28] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp.208 - 211, 1979.

[29] P. Scalart and J. V. Filho, "Speech enhancement based on A Priori Signal to Noise Estimation", *Proc. ICASSP*, vol. 2, pp.629 - 632, 1996.

[30] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor", *IEEE Trans. Speech Audio Process.*, vol.2, no.2, pp.345-349, Apr. 1994.

[31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 2000. :Academic

[32] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation", *IEEE Trans. Speech Audio Process.*, vol.13, no.5, pp.870-881, 2005.

[33] J. S. Erkelens, J. Jensen and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria", *Speech Commun.,* vol. 49, pp.530 -541, 2007.

[34] M. K. Hasan, S. Salahuddin and M. R. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules", *IEEE Signal Process. Lett.*, vol. 11, no. 4, pp.450-453, 2004.

[35] I.Y. Soon and S.N. Koh, "Low Distortion Speech Enhancement", *IEE Proceedings on Vision, Image and Signal Processing*, Vol. 147, Issue 3, pp 247-253, June 2000.

[36] I. Cohen, "Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation", *Speech Commun.*, vol. 47, no. 3, pp.336 -350, 2005.

[37] C. Plapous, C. Marro and P. Scalart, “Improved Signal-to-Noise Ratio Estimation for Speech Enhancement”, *IEEE Trans. Acoust., Speech and Signal Process*., vol.14, no.6, pp.2098-2108, Nov. 2006.

[38] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement", *Proc. IEEE Workshop Statistical Signal Processing*, pp.496 -499, 2001.

[39] C. H. You , S. N. Koh and S. Rahardja, " β-order MMSE spectral amplitude estimation for speech enhancement", *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp.475 -486, 2005.

[40] R. Martin, "Speech enhancement using MMSE short-time spectral estimation with gamma speech prior", *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. I, pp.253-256, 2002.

[41] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model", *EURASIP J. Appl. Signal Process.*, vol. 7, pp.1110-1126, 2005.

[42] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors", *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp.845 -856, 2005.

[43] J. S. Erkelens , R. C. Hendriks , R. Heusdens and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors", *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 6, pp.1741-1752, 2007.

[44] C. Bin and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement", *Speech Commun.*, pp.134-143, 2007.

[45] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments", *Proc. 24th IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'99)*, pp.789 - 792, 1999.

[46] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments", *Signal Process.*, vol. 81, no. 11, pp.2403–2418, Nov. 2001.

[47] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator", *IEEE Signal Process. Lett.*, vol.9, no. 4, pp.113-116, Apr. 2002.

[48] I.Y. Soon, S.N. Koh and C.K. Yeo, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence", *Signal Processing*, vol.75, pp.151-159, 1999.

[49] T. Gerkmann, C. Breithaupt and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors", *IEEE Trans. Audio, Speech and Language Processing*, vol.16, no.5, July 2008.

[50] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Trans. Speech Audio Processing*, vol. 9, pp.504 -512, 2001.

[51] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detector", *IEEE Signal Processing Lett.* vol. 6, pp.1 - 3, 1999.

[52] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging", *IEEE Trans. Speech Audio Processing*, vol. 11, pp.466 -475, 2003.

[53] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 412-424, 2006.

[54] “Subjective performance evaluation of telephone band and wideband codecs”, ITU, ITU-T Rec. P. 830, 1998.

[55] M. Brookes, N. D. Gaubitch, M. Huckvale, and P. A. Naylor, “Speech cleaning literature review”, Tech. Rep. CTR-2, Feb. 2008. [Online]. Available: www.clear-labs.com/Tutorial-LitReview/index.html

[56] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms", *Proc. Int. Conf. Spoken Lang. Process.*, vol. 7, pp.2819-2822, 1998.

[57] S. Quackenbush, T. Barnwell and M. Clements, *Objective Measures of Speech Quality*, 1988 :Prentice-Hall.

[58] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, ITU-T Rec. P. 862, 2000.

[59] A. Rix, J. Beerends, M. Hollier and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, pp.749-752, 2001.

[60] N. Kitawaki and T. Tamada, "Subjective and Objective Quality Assessment for Noise Reduced Speech," ETSI Workshop on Speech and Noise in Wideband Communication, May 2007.

[61] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement", *Proc. Interspeech*, pp.1447-1450, 2006.

[62] “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm”, ITU-T, ITU-T Rec.P.835, 2003

[63] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement", *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp.229-238, 2008.

[64] R. Gribonval and K. Schnass, “Some recovery conditions for basis learning by L_1-minimization,” in *Proc. Int. Symp. Commun., Control, Signal Process. (ISCCSP),* 2008, pp. 768–733.

[65] M. G. Jafari and M. D. Plumbley "Fast dictionary learning for sparse representations of speech signals", *IEEE J. Sel. Topics Signal Process.,* vol. 5, no. 5, pp.1025 -1031, 2011.

[66] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries", *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp.3397-3415, 1993.

[67] H. Rauhut, K. Schnass and P. Vandergheynst, "Compressed sensing and redundant dictionaries", *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp.2210-2219, 2008.

[68] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage", *Biometrika*, vol. 81, pp.425-455, 1994.

[69] L. Rebollo-Neira, “Dictionary redundancy elimination,” *IEE Proc.—Vis., Image, Signal Process.*, vol. 151, pp. 31–34, 2004.

[70] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis", *IEEE Trans. Inform. Theory*, vol. 36, no. 5, pp.961 - 1005 , 1990.

[71] S.G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp.674 - 693, 1989.

[72] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Wavelet for speech denoising", *TENCON Proc.*, vol. 2, pp.479 - 482, 1997.

[73] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets", *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp.617-643, 1992.

[74] D. L. Donoho, "Nonlinear wavelet methods for recovery of signals, images, and densities from noisy and incomplete data", *Amer. Math. Soc. Different Perspectives Wavelets*, vol. 1, pp.173-205, 1993.

[75] Y. Xu, J. Weaver, D. Healy, J. Lu, "Wavelet Transform Domain Filters: A Spatially Selective Noise Filtration Technique," *IEEE Trans. Image Processing,* vol. 3, pp. 747-758, Nov. 1994.

[76] Q. Pan, L. Zhang, G. Dai, H. Zhang, "Two denoising methods by wavelet transform", *IEEE Trans. Signal Processing*, vol. 47, pp.3401-3406, 1999.

[77] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[78] I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise", *J. R. Statist. Soc.*, vol. 59, 1997.

[79] D.L. Donoho and I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage", *Journal of the American Statistical Association*, vol.90, pp.1200-1224, 1995.

[80] X.-P. Zhang and M. D. Desai, "Adaptive denoising based on SURE risk", *IEEE Signal Process. Lett.*, pp.265-267, 1998.

[81] T. Gulzow, A. Engelsberg, and U. Heute, "Comparison of a discrete wavelet transformation and nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement", *Signal Process.*, vol. 64, pp.5-19, 1998.

[82] D. Mahmoudi, "A microphone array for speech enhancement using multiresolution wavelet transform", *Proc. Eurospeech',* pp.339-342, 1997.

[83] J. Sika and V. Davidek, "Multi-channel noise reduction using wavelet filter bank", *EuroSpeech',* pp.2595-2598, 1997.

[84] D. Mahmoudi and A. Drygajlo, "Combined wiener and coherence filtering in wavelet domain for microphone array speech enhancement", *ICASSP*, pp.358-388, 1998.

[85] Y. Ghanbari, M.R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets", *Speech Communication*, vol. 48, pp. 927–940, 2006.

[86] M.T. Johnson, X. Yuan, Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding", *Speech Communication*, vol. 49, pp. 123–133, 2007.

[87] M. Bahoura, J. Rouat, "Wavelet speech enhancement based on time-scale adaptation", *Speech Communication*, 48, 1620-1637, 2006.

[88] Y. Hu and P.C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum", *IEEE Trans. Speech Audio Process*., vol.12, no.1, pp.59-67, Jan. 2004.

[89] C.-T. Lu, -H.C. Wang, "Speech enhancement using perceptually-constrained gain factors in critical-band-wavelet-packet transform", *Electronic Letter*, vol. 40 (6), pp. 394–396. 2004.

[90] C.-T. Lu, K.-F. Tseng, "A gain factor adapted by masking property and SNR variation for speech enhancement in colored-noise corruptions", *Computer Speech & Lang.*, vol. 24, pp. 632–647, 2010.

[91] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Time Series Analysis*, M. Rosenblatt, Ed. Ch. 15, pp. 209–243, 1963.

[92] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Englewood, NJ: Prentice-Hall, 1975.

[93] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp.357-366, 1980.

[94] C. R. Jankowski, H. H. Vo, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition", *IEEE Trans. Speech Audio Processing*, vol. 3, pp.286-293, 1995.

[95] R. Hagen, "Spectral quantization of cepstral coefficients", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp.II13-II16, 1994.

[96] R. W. Schafer, "Homomorphic systems and cepstrum analysis of speech," *Springer Handbook of Speech Processing and Communication*, Springer, 2008.

[97] L. Rabiner and R. Schafer. *Introduction to digital speech processing. Foundations and Trends in Signal Processing*, 1(1/2):1-194, 2007.

[98] J. D. Markel and A. H. Gray Jr. *Linear Prediction of Speech*, Berlin, Germany: Springer-Verlag, 1976.

[99] M. B. Priestley, *Spectral Analysis and Time Series*, New York: Academic, 1992.

[100] R. C. S. Lai, T. C. M. Lee, R. K. W. Wong, and F. Yao, "Nonparametric cepstrum estimation via optimal risk smoothing", *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1507-1514, 2010.

[101] P. Stoica and N. Sandgren, "Smoothed non parametric spectral estimation via cepstrum thresholding," *IEEE Signal Process. Mag*., vol. 23, no. 6, pp. 34–45, 2006.

[102] P. Stoica and N. Sandgren, "Total variance reduction via thresholding: Application to cepstral analysis," *IEEE Trans. Signal Process.*, vol. 55, no. 1, pp. 66-72, 2007.

[103] E. J. Hannan and D. F. Nicholls, "The estimation of the prediction error variance", *J. Amer. Statist. Assoc.*, vol. 72, no. 360, pp.834-840, 1977.

[104] M. Taniguchi, "On estimation of parameters of Gaussian stationary processes", *J. Appl. Prob.*, vol. 16, pp.575-591, 1979.

[105] Y. Ephraim and M. Rahim, "On second-order statistics and linear estimation of cepstral coefficients", *IEEE Trans. Speech Audio Process.*, vol.7, no.2, pp.162-176, March 1999.

[106] E. Gudmundson, N. Sandgren and P. Stoica, "Automatic smoothing of periodograms", *Proc. 31st ICASSP*, vol. 3, pp.III: 504 -III: 507, 2006.

[107] I.M. Johnstone and B.W. Silverman, "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possible Sparse Sequences", *Annals of Statistics*, vol. 32, pp.1594-1649, 2004.

[108] I.M. Johnstone and B.W Silverman, "EbayesThresh: R Programs for Empirical Bayes Thresholding", *Journal of Statistical Software*, vol. 12, issue 8, pp1-38, 2005.

[109] T. Sreenivasuly Reddy and G. Ramachandra Reddy, "MST Radar Signal Processing Using Cepstral Thresholding", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 6, pp. 2704-2710, 2010.

[110] K. Linhard and T. Haulick, "Noise subtraction with parametric recursive gain curves", *Proc. Eurospeech—Eur. Conf. Speech Communication and Technology*, pp.2611-2614, 1999.

[111] Z. Goh , K.-C. Tan and B. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction", *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp.287-292, 1998.

[112] H. Gustafsson , S. E. Nordholm and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging", *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp.799-807, 2001.

[113] R. Martin and T. Lotter, "Optimal recursive smoothing of non-stationary periodograms", *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, pp.167-170, 2001.

[114] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*, New York: Wiley, 2006.

[115] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 1036-1039, 2007.

[116] C. Breithaupt, T. Gerkmann and R. Martin, "A novel a prior SNR estimation approach based on selective cepstro-temporal smoothing", *Proceedings, IEEE Int. Conf. Acoustics, Speech and Signal Processing*, March 2008, pp. 4897-4900.

[117] A. M. Noll, "Cepstrum pitch estimation", *J. Acoust. Soc. Amer.*, vol. 41, pp.293-309, 1967.

[118] J. Wang, H. Liu, C. Zheng, X. Li, "Spectral subtraction based on two-stage spectral estimation modified cepstrum thresholding", *Applied Acoustics*, vol. 74, pp. 450–458, 2013.

[119] T. Gerkmann, R. C. Hendriks, "Improved MMS-based noise PSD tracking using temporal cepstrum smoothing," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., March 2012, pp. 105-108.

[120] X. Hu, S. Wang, C. Zheng and X. Li, "A cepstrum-based preprocessing and postprocessing for speech enhancement in adverse environments", *Applied Acoustics*, vol. 74, no. 12, pp.1458-1462, 2013.

[121] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images", *Nature*, vol. 381, pp.607-609, 1996.

[122] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations", *Neural*

Comput., vol. 12, pp.337-365, 2000.

[123] K. Kreutz-Delgado, J. Murray, D. Rao, K. Engan, T. Lee and T. Sejnowski, "Dictionary learning algorithms for sparse representations", *Neural Comput.*, vol. 15, pp.349-396, 2003.

[124] I. Gorodnitsky, J. George and B. Rao, "Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm", *J. Electroencephalography Clinical Neurophysiol.*, vol. 95, pp.231-251, 1995.

[125] K. Engan, S. O. Aase and J. Hakon Husoy, "Method of optimal directions for frame design", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, pp.2443-2446, 1999.

[126] K. Engan, B. D. Rao and K. Kreutz-Delgado, "Frame design using FOCUSS with method of optimal directions (MOD)", *Proc. Norwegian Signal Process. Symp.*, pp.65-69, 1999.

[127] M. Aharon, M. Elad and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representations", *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp.4311-4322, 2006.

[128] R. Rubinstein, M. Zibulevsky and M. Elad, “Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit,” technical report, Technion-computer Science Department, Haifa, 2008.

[129] M. Aharon and M. Elad, "Sparse and redundant modeling of image content using an image-signature-dictionary", *SIAM J. Imaging Sci.*, vol. 1, no. 3, pp.228-247, 2008.

[130] M. Yaghoobi, L. Daudet and M. E. Davies, "Parametric dictionary design for sparse coding", *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp.3311-3332, 2009.

[131] J. Mairal , F. Bach, J. Ponce and G. Sapiro, "Online learning for matrix factorization and sparse coding", *J. Mach. Learn. Res.*, vol. 11, pp.19-60, 2010.

[132] S. S. Chen, D. L. Donoho, and M. A. Sauders, “Atomic decomposition by basis pursuit”, *SIAM Review*, vol.43(1), pp.129–159, Mar. 2001.

[133] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill, “Sparse linear regression with structured priors and application to denoising of musical audio”, *IEEE Trans. on Audio, Speech and Language Processing*, vol.16(1), pp.174–185, 2008.

[134] M.N. Schmidt, J. Larsen and F.T. Hsiao, “Wind noise reduction using non-negative sparse coding,” *in Proc., IEEE Workshop on Machine Learning for Signal Processing*, 2007, pp. 431-436.

[135] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” *in Proc., IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Mar. 2008, pp. 4029-4032.

[136] C. Sigg, T. Dikk and J. Buhmann, “Speech enhancement using generative dictionary Learning. “*IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no.67, pp. 1698-1712, 2012.

[137] Y. Zhao, X. Zhao, B. Wang, “A speech enhancement method based on sparse reconstruction of power spectral density,” *Computers & Electrical Engineering*, vol. 40, issue 4, pp. 1080–1089, 2014.

[138] T. Hasan and M.K. Hasan, “MMSE estimator for speech enhancement considering the constructive and destructive interference of noise”, *IET Signal Processing*, Vol 4, Iss. 1, pp.1–11, 2010.

[139] Y. Shao and C. H. Chang, “A generalized time-frequency subtraction method for robust speech enhancement based on wavelet filter bank modeling of human auditory system,” *IEEE Trans. Systems, Man and Cybernetics, Part B: Cybernetics,* vol. 37, no. 4, pp. 877- 889, August 2007.

[140] T. Gerkmann, M. Krawczyk and R. Martin, “Speech Presence probability estimation based on temporal cepstrum smoothing”, *Proceedings, IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp.4254-4257, March 2010.

[141] H. Tasmaz, E. Ercelebi, "Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE-STSA estimation in various noise environments", *Digital Signal Processing*, vol.18, pp.797-812, 2008.

[142] C.T. Lu and H.C. Wang, "Speech enhancement using hybrid gain factor incritical-band-wavelet-packet transform", *Digital Signal Processing*, vol.17, pp.172-188, 2007.

[143] R. M. Uderea, N. D. Vizireanu, and S. Ciochina, "An improved spectral subtraction method for speech enhancement using a perceptual weighting filter", *Digital Signal Processing*, vol.18, 2008, pp.581-587.

[144] J. M. Kum, Y. S. Park, J. H. Chang, "Improved minima controlled recursive averaging technique using conditional maximum a posteriori criterion for speech enhancement", *Digital Signal Processing*, vol. 20, 2010, pp. 1572-1578.

[145] Md. J. Alam, D. O'Shaughnessy, "Perceptual improvement of Wiener filtering employing a post-filter", *Digital Signal Processing,* vol. 21, 2011, pp. 54-65.

[146] P. Moulin, “Wavelet thresholding techniques for power spectrum estimation”, *IEEE Trans. Signal Process*., vol.42, pp.3126-3136, Nov. 1994.

[147] A.T. Walden, D.B. Percival, and E.J. McCoy, “Spectrum estimation by wavelet thresholding of multitaper estimators”, *IEEE Trans. Signal Process*., vol.46, pp.3153-3165, Dec. 1998.

[148] P.M. Clarkson, *Optimal and adaptive Signal processing*, CRC Press, 1993.

[149] D.B. Percival and A.T. Walden, *Spectral analysis for physical applications: multitaper and conventional univariate techniques*, Cambridge Univ. Press, 1993.

[150] C. M. Stein, “Estimation of the mean of a multivariate normal distribution,” *Ann. Statist.,* vol.9, no.6, pp. 1135–1151, 1981.

[151] M. Jansen, M.Malfait and A.Bultheel, "Generalized Cross Validation for wavelet thresholding", *Signal Processing,* vol.56, no.1, pp.33-44, Jan. 1997.

[152] D. Leporini and J.C. Pesquet, “Bayesian wavelet denoising; Besov priors and non-Gaussian noises”, *Signal Processing,* vol.81, pp.55-67, 2001.

[153] L. Sendur and I.W. Selesnick, “Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency”, *IEEE Trans. Signal Process.,* vol.50, no.11, pp.2744-2756, Nov. 2002.

[154] R.D. Nowak, “Wavelet-based Rician noise removal for magnetic resonance imaging,” *IEEE Trans. Image Processing,* vol. 8, no. 10, October 1999, pp. 1408–1419.

[155] A. Davis, S. Nordholm and R. Togneri, “Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold”, *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 2, March 2006, pp.412-424.

[156] D.K. Kim and J.H. Chang, “A subspace approach based on embedded prewhitening for voice activity detection”, J. Acoust. Soc. Am. Volume 130, Issue 5, 2011, pp. EL304-EL310.

[157] P.J. Huber, *Robust Statistics*. New York: John Wiley and Sons, 1981.

[158] J. Picklands, “Maxima of stationary Gaussian processes”, *Probability Theory and Related Fields*, Vol.7, No.3, 1967, pp.190-223.

[159] I. Daubechies, *Ten lectures on wavelets,* Philadelphia, Pa.: SIAM, 1992.

[160] D.P.K. Lun, T.W. Shen, T.C. Hsung, D.K.C. Ho,” Wavelet based speech presence probability estimator for speech enhancement”, *Digital Signal Processing*, Vol.22, Issue 6, pp. 1161–1173, Dec. 2012.

[161] A. P. Dempster, N. M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete

data via the EM algorithm," *J. Royal Statiscal Soc.*, vol. 39, no. 1, pp. 1 -38, 1977.

[162] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imag.*, vol.1, no. 2, pp. 113-122, 1982.

[163] D. L. Snyder and D. G. Politte, "Image reconstruction from list-mode data in an emission tomography system having time-of-flight measurements," *IEEE Trans. Nucl. Sci.*, vol. 30, no. 3, pp. 1843-1849, 1983.

[164] J. Zhou, J. Coatrieux, A. Bouusse, H. Shu, and L. Luo, "A Bayesian MAP-EM algorithm for PET image reconstruction using wavelet transform," *IEEE Trans. Nucl. Sci.*, vol. 54, no. 5, pp. 1660-1669, 2007.

[165] J. Tang, T. S. Lee, X. He, W. P. Segars, and B. M. W. Tsui, "Comparison of 3D OS-EM and 4D MAP-RBI-EM reconstruction algorithms for cardiac motion abnormality classification using a motion observer," *IEEE Trans. Nucl. Sci.*, vol. 57, no. 5, pp. 2571-2577, 2010.

[166] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[167] L. R. Welch, "Hidden Markov Models and the Baum-Welch Algorithm," *IEEE Inf. Theory Society Newsletter*, vol. 53, no. 4, pp. 1–13, 2003.

[168] D. Y. Zhao and W. B. Kleijn "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 882-892, 2007.

[169] J. Hao, H. Attias, and S. Nagarajan, T.-W. Lee, and T. J. Sejnowski "Speech enhancement, gain, and noise spectrum adaptation using approximate Bayesian estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 1, pp. 24 -37, 2009.

[170] S. Gannot, D. Burshtein, and E. Weinstein, “Iterative and sequential Kalman filter-based speech enhancement algorithms,” *IEEE Trans. Speech, Audio Process.*, vol. 6, no. 4, pp. 373-385, 1998.

[171] S. Park and S. Choi, “A constrained sequential EM algorithm for speech enhancement,” *Neural Networks*, vol. 21, pp. 1401-1409, 2008.

[172] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 11, pp. 47 -60, 1996.

[173] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, New York, NY: John Wiley & Sons, 2nd Edition, 2008.

[174] Y. Ephraim and W. Roberts, “On second-order statistics of log-periodogram with correlated components,” *IEEE Signal Process. Lett.*, vol. 12, pp. 625–628, 2005.

[175] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, Walter de Gruyter, Berlin, 2008.

[176] M. Figueiredo, “An EM algorithm for wavelet-based image restoration,” *IEEE Trans. Image Process.*, vol. 12, no. 8 pp. 906 – 916, 2003.

[177] N. Cao , A. Nehorai, and M. Jacob, "Image reconstruction for diffuse optical tomography using sparsity regularization and expectation-maximization algorithm," *Opt. Express*, vol. 15, no. 21, pp. 13695 -13708, 2007.

[178] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with Generalized Gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, 2007.

[179] D.P.K. Lun and T.C. Hsung, “Improved Wavelet Based A-priori SNR Estimation for Speech Enhancement”, in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS),* Paris, France, pp.2382-2385, May 2010.

[180] D.P.K. Lun, T.W. Shen, T.C. Hsung, D.K.C. Ho," Improved speech presence probability estimation based on wavelet denoising", *Proceedings, IEEE International Symposium on Circuits and Systems (ISCAS'2012),* Seoul, Korea, May 2012, pp.1018-1021.

[181] T.W. Shen and D.P.K. Lun," Speech Enhancement Based on L1 Regularization in the Cepstral Domain", *Proceedings, IEEE International Symposium on Circuits and Systems (ISCAS'2014)*, Melbourne, Australia, June 2014, pp. 121-124.

[182] Daniel P.K. Lun, T.W. Shen and K.C. Ho, "A novel expectation-maximization framework for speech enhancement in non-stationary noise environments", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 22, Issue 2, pp.335-346, Feb 2014.

[183] T.W. Shen, D.P.K. Lun and T.C. Hsung, "Speech Enhancement Using Harmonic Regeneration With Improved Wavelet Based A-Priori Signal To Noise Ratio Estimator" *Proceedings, 2010 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS 2010)*, pp.225-228, Cheng Du, China, Dec 2010.

[184] J.F. Gemmeke, T. Virtanen, A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2067-2080, 2011.

[185] J. Mairal, M. Elad, G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53-69, 2008.

[186] S.W. Deng, J.Q. Han, "Statistical voice activity detection based on sparse representation over learned dictionary," *Digital Signal Processing*, vol. 23, issue 4, pp. 1228–1232, 2013.

[187] I. Naseem, R. Togneri, M. Bennamoun, “Sparse representation for speaker identification,” *in Proc., IEEE Int. Conf. Pattern Recognition*, Aug. 2010, pp. 4460-4463.

[188] R. Rubinstein, M. Zibulevsky, and M. Elad, “Double sparsity: Learning sparse dictionaries for sparse signal approximation,” *IEEE Trans. Signal Process*., vol. 58, no. 3, pp. 1553–1564, 2010.

[189] B. Efron , T. Hastie , I. Johnstone and R. Tibshirani, "Least angle regression", *Ann. Statist.*, vol. 32, pp.407-499, 2004.

[190] I. Tosic and P. Frossard, "Dictionary learning", *IEEE Signal Process. Mag*., vol. 28, no. 2, pp.27 -38, 2011.

[191] T.W. Shen and D.P.K. Lun, “A Speech Enhancement Method Based on Sparse Reconstruction on Log-Spectra”, under preparation.

[192] J.S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database", Nat. Inst. Standards Technol. (NIST), Gaithersburg, MD, prototype as of Dec.1988.

[193] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems", *Speech Communication*, vol.12, no.3, pp.247-251, July 1993.

[194] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proceedings of the IEEE ICASSP*, pp. 4266-4269, Mar. 2010.

[195] T. Gerkmann, R. C. Hendriks, "Improved MMS-based noise PSD tracking using temporal cepstrum smoothing," in *Proceedings of the IEEE ICASSP*, pp. 105-108, Mar 2012.

[196] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase inspeech enhancement", *ELSEVIER Speech Commun.*, vol. 53, no. 4, pp.465-494, 2011.

[197] U. R. Abeyratne, A. S. Wakwella, and C. Hukins, "Pitch jump probability measures for the analysis of snoring sounds in apnea, " *Physiological Measurement*, vol. 26, no. 5, pp. 779-98, 2005.

[198] A. K. Ng, T. S. Koh, K. Puvanendran and U. R. Abeyratne, "Snore signal enhancement and activity detection via translation-invariant wavelet transform", *IEEE Trans. Biomed. Eng.*, vol. 55, no. 10, pp.2332-2342 2008.