Tuan Van Pham

Wavelet Analysis For Robust Speech Processing and Applications

Dissertation

vorgelegt an der
Technischen Universität Graz

zur Erlangung des akademischen Grades
Doktor der Technischen Wissenschaften

durchgeführt am
Institut für Signalverarbeitung und Sprachkommunikation

First Advisor:
Prof. Dr. Gernot Kubin, Graz University of Technology, Austria
Second Advisor:
Prof. Dr. Zdravko Kačič, University of Maribor, Slovenia

Graz, February 2007

Abstract

In this work, we study the application of wavelet analysis for robust speech processing.

Reliable time-scale features (TS) which characterize the relevant phonetic classes such as voiced (V), unvoiced (UV), silence (S), mixed-excitation, and stop sounds are extracted. By training neural and Bayesian networks, the classification rates provided by only 7 TS features are mostly similar to the ones obtained by 13 MFCC features.

The TS features are further enhanced to design a reliable and low-complexity V/UV/S classifier. Quantile filtering and slope tracking are used for deriving adaptive thresholds. A robust voice activity detector is then built and used as a pre-processing stage to improve the performance of a speaker verification system.

Based on wavelet shrinkage, a statistical wavelet filtering (SWF) method is designed for speech enhancement. Non-stationary and colored noise is handled by employing quantile filtering and time-frequency adaptive weighting. A newly proposed comparison diagnostic test and other subjective tests show improvements compared with other denoising methods.

The SWF is further optimized to enhance speech quality for robust ASR. By changing the shape of the frequency weighting and estimating perceptual noise thresholds in critical subbands, the perceptual SWF method provides almost equal performance compared with the ETSI baseline for car noise and significant improvements compared with other methods in aircraft maintenance factory conditions.

Kurzfassung

Diese Arbeit beschäftigt sich mit der Anwendung der Wavelet Analyse zur robusten Sprachverarbeitung. Schlagwrter So genannte time-scale (TS) Merkmale, die phonetische Klassen (stimmhaft, stimmlos, gemischte Anregung, Verschlusslaut) charakterisieren, werden aus dem Sprachsignal extrahiert. Mittels Neuronaler und Bayes'scher Netze werden unter Verwendung von nur 7 TS Merkmalen gleiche Resultate in der Klassifikation wie mit 13 MFCC Merkmalen erreicht.

Um Robustheit gegenüber sich ändernden Umgebungsbedingungen zu gewährleisten, werden adaptive Schwellen unter der Verwendung von Methoden der Quantilen-Filterung oder der Verfolgung des Kurvenzuges abgeleitet. Diese Wavelet Methode wurde zur Detektierung von Sprachaktivität weiterentwickelt und reduziert als Vorverarbeitungseinheit eines Sprecherverifikationssystems dessen Fehlerrate.

Weiters wurde diese Methode zur Sprachverbesserung eingesetzt. Dafür wurde die statistische Wavelet Filterung angewandt, die auf der "wavelet shrinkage" Methode basiert. Um robust gegenber nicht-stationärem und nicht-weiem Rauschen zu sein, wird zusätzlich zur Quantilen-Filtering eine adaptive Zeit-Frequenz Gewichtung angewandt.

Nach der Optimierung der Gewichtungsfunktionen für die Spracherkennung konnten vergleichbare Ergebnisse wie mit der von ETSI entwickelten Methode, für in Fahrzeugen gemachte Aufnahmen erreicht werden sowie Verbesserungen in Aufnahmen, die bei der Fluzeugwartung gemacht wurden.

To my parents and brother, Pham V.Hai, Truong T.T.Thanh, Pham V.Trung,
and to my lovely wife Vu T.A.Nguyet and our little beautiful son Pham V.Kha.

Acknowledgments

It is much pleasure to acknowledge the support I received over my Ph.D study.

First of all, I would greatly appreciate my supervisor, Prof. Gernot Kubin, for his excellent guidance during my study. I always admired his deeply erudition, professionally researching methodology and ability of very simple and logical explanation. I did learn a lot from him not only in study but also in life.

I am thankful to Dr. Franz Pernkopf for his enthusiastic collaboration. I am very happy to work with him, my "consultant" in machine learning field. I specifically thank his support for learning Bayesian network. I never forget his help in the beginning period of my living in Austria.

I would express my gratefulness to Dr. Erhard Rank for his wonderful co-operation. Doing research with him is my absolute pleasure. His smart advices mostly lead to successful and promising outcomes.

Thanks to Michael Neffe for his indispensable speaker verification system. His patient and hard working helps our sharing research run smoothly. Thanks to Dr. Marian Kepesi and Dr. Dimitry Shutin for interesting discussions.

I would like to thank Prof. Zdravko Kačič for giving me an opportunity to visit his Digital Signal Processing Laboratory. I would take this chance to thank Dr. Bojan Kotnik for his valuable experiences in wavelet denoising topic.

Many thanks to all my colleagues at Signal Processing and Speech Communication Laboratory who have helped and influenced me throughout my doctoral study. My family and I would thank for your supports to our life in Graz.

Finally, I would very much thank to Österreichisher Austauschdienst for giving me a chance to study Ph.D in Austria.

Tuan Van Pham
Graz, February 2007.

Contents

Contents x

List of Figures xiv

List of Tables xvii

Abbreviations xxi

1 Introduction 1

1.1 Robust speech processing and applications 2

1.2 Wavelet transform 4

1.3 Why wavelets for speech processing ? 8

1.4 Outline of the thesis 10

1.5 Work contributions 12

2 Time-scale features for phonetic classification 15

2.1 Introduction 15

2.2 Time-scale features for speech classification 18

2.2.1 Acoustic-phonetic properties 18

2.2.2 Voiced/unvoiced/mixed-excitation/silence classification 19

2.2.3 Transient detection 20

2.2.4 Feature extraction 21

2.3 Joint neural network classifier 25

2.3.1 Network initialization 26

2.3.2 Network learning 27

2.3.3 Selection of parameters and structure 29

2.3.4 Joint classifier 30

2.4 Bayesian network classifier 32

2.4.1 Classifier structures 33

2.4.2 Network learning 33
2.5 Evaluation and discussions 35
2.5.1 Data setup 35
2.5.2 Performance of joint classifier 35
2.5.3 Performance of the Bayesian network 39
2.6 Conclusions 41

3 Robust speech classifiers for adverse environments 43
3.1 Introduction 43
3.2 Robust feature extraction 46
3.2.1 Teager Energy Operator 46
3.2.2 Sigmoidal delta feature 48
3.3 Threshold adaptation by quantile filtering 49
3.3.1 Quantile filtering 50
3.3.2 Threshold adaptation algorithm 51
3.4 Threshold adaptation by slope tracking 51
3.4.1 Slope generation 52
3.4.2 Slope detection 53
3.5 Evaluation of classification rate 56
3.5.1 Experimental setup 56
3.5.2 Results and discussions 57
3.6 Evaluation in application domain 61
3.6.1 VAD as a pre-processing stage 61
3.6.2 Experiments and evaluations 64
3.7 Conclusions 68

4 Statistical wavelet filtering for speech enhancement 71
4.1 Introduction 71
4.1.1 Optimal estimation for frequency-domain denoising 73
4.1.2 Wavelet shrinking estimator 74
4.2 State-of-the-art speech enhancement 76
4.2.1 Spectral subtraction and its variants 77
4.2.2 Wiener filtering 78
4.2.3 Optimal non-linear estimator 79
4.2.4 Noise estimation 81
4.3 Wavelet denoising approach 81

4.3.1 Hard and soft thresholding functions 82

4.3.2 Shrinking functions 84

4.3.3 Optimal shrinkage: Smoothed hard thresholding 85

4.4 Estimate of thresholds 88

4.4.1 Bias and variance 88

4.4.2 Universal threshold procedure 90

4.4.3 White and non-white noise 90

4.5 Statistical quantile filtering 92

4.5.1 Quantile filtering and recursive buffer 92

4.5.2 Noise threshold estimation 92

4.6 Time-frequency weighted quantile threshold 94

4.6.1 Non-linear frequency weighting 95

4.6.2 Adaptive temporal weighting 96

4.6.3 Adaptive factor for smoothed hard shrinking 97

4.7 Speech enhancement evaluation 99

4.7.1 Experimental setup 99

4.7.2 Objective test 101

4.7.3 Segmental evaluation: Comparison Diagnostic Test 101

4.7.4 Overall evaluation: Comparison Category Rating 103

4.7.5 Overall evaluation: ITU-T Standard P.85 103

4.8 Conclusion 104

5 Noise reduction for robust speech recognition 107

5.1 Introduction 107

5.2 Why noise reduction for robust ASR? 110

5.3 Robust front-ends 111

5.3.1 Standard front-end parameterization 112

5.3.2 Advanced front-end parameterization 112

5.4 Two-stage Mel-warped Wiener filtering 113

5.5 Enhancement of statistical wavelet filtering for ASR 115

5.5.1 Frequency weighting shape for ASR 116

5.5.2 Perceptual threshold estimation 118

5.6 Experiments and evaluations 123

5.6.1 Speech databases and recognizers 123

5.6.2 Test results with the SWF 124

5.6.3 Test results with perceptual SWF 126

5.7 Conclusion 128

6 Conclusions and perspectives **129**

A Wavelet properties **133**

B Multi-threshold decision classifier **135**

C Evaluating hypotheses **137**

D Subjective tests for speech enhancement **139**

D.1 Comparison Diagnostic Test 139

D.2 Overall quality test ITU-T P.85 140

E Robust speech recognition **143**

E.1 Frequency weighting for hearing aid 143

E.2 Parametric turning of the time-frequency weighting for ASR 144

E.3 Recognition performance 146

Bibliography **151**

List of Figures

2.1 Classification diagram for phonetic groups. 19

2.2 Voiced segment, approximation and detail 20

2.3 Unvoiced segment, approximation and detail 21

2.4 Voiced frame, detail variation and spectral tilt 22

2.5 Unvoiced frame, detail variation and spectral tilt 23

2.6 Speech signal, power variation of detail coefficients 24

2.7 Transient frame, detail distribution across scales 25

2.8 Evolutions of the four extracted features 26

2.9 Three extracted transient features 27

2.10 Feedforward neural network configuration 28

2.11 Performance of the learning algorithms at $\gamma = 0.5$. 30

2.12 The best performance ratios γ for LM algorithm. 31

2.13 Block scheme of combined classifier. 32

2.14 Bayesian Network: (a) NB, (b) TAN [PP06a]. 34

3.1 Block diagram of the proposed speech classifier. 46

3.2 Noise-free unvoiced frame, approximation and detail oefficients 47

3.3 Noisy unvoiced frame, wavelet and TEO coefficients 48

3.4 Feature extraction and enhancement 50

3.5 Estimation of quantile threshold 52

3.6 The filtered parameter D_f, clean speech 53

3.7 The filtered parameter D_f, noisy speech 54

3.8 Slope detection. 55

3.9 Noisy signal, V/U/S labels by quantile filtering and slope tracking . . . 56

3.10 VAD as pre-processing stage in the SV system [NPHKon]. 62

3.11 Design VAD from outputs of the phonetic classifier 62

3.12 Adaptive quantile factors for different buffers 63

3.13 VAD log-energy with constant quantile filtering 64

3.14 VAD ETSI with updated SNR-based threshold 65
3.15 VAD wavelet sigmoidal detal 66
3.16 DET-curves and EERs for the SV system 67

4.1 Block scheme of standard single-channel noise reduction approach. . . . 72
4.2 Input-output characteristic of hard thresholding 83
4.3 Input-output characteristic of soft thresholding 84
4.4 Enhanced wavelet shrinking by smoothed hard-thresholding. 86
4.5 Wavelet coefficients denoised by smoothed hard thresholding 87
4.6 Unvoiced segment denoised by smoothed hard thresholding 88
4.7 Voiced segment denoised by smoothed hard-thresholding 89
4.8 Quantiles of sorted threshold values 93
4.9 Recursive scheme. 94
4.10 Frequency non-linear weighting function. 96
4.11 Temporal adaptive weighting based on TDC 97
4.12 Frequency weighting in 3D 98
4.13 Time-frequency weighting in 3D 99
4.14 Smoothed time-frequency weighting in 3D 100
4.15 Input-ouput SegSNR for different algorithms. 101
4.16 Comparison Diagnostic Test results. 102
4.17 Overall test ITU-T (P.85) Aurora3 104
4.18 Overall test ITU-T (P.85) NTIMIT 105

5.1 Noise reduction approach for robust ASR. 109
5.2 Feature extraction scheme of standard front-end 112
5.3 Block scheme of advanced front-end 113
5.4 Two-stage Mel-warped Wiener filtering noise reduction scheme 113
5.5 Noise spectrum estimation for the construction of Wiener filter 114
5.6 Waveform of car-noisy recording and its enhancements 116
5.7 Spectrogram of car-noisy recording and its enhancements 117
5.8 Designed weighting function for speech recognition application. 118
5.9 Waveform of siren noise recording and its enhancements 119
5.10 Spectrogram of siren noise recording and its enhancement 120
5.11 Perceptual statistical wavelet filtering 120
5.12 Tree structure of the perceptual wavelet packet decomposition. 121

B.1 Scatter plot for features 136

B.2 The flow chart of the multi-threshold linear classifier. 136

E.1 Frequency non-linear weighting function. 144
E.2 Tendencies of D/S/I obtained with SFE in without retraining mode . . 146
E.3 Tendencies of D/S/I obtained with SFE in retraining mode 147
E.4 Tendencies of D/S/I obtained with AFE in without retraining mode . . 148
E.5 Tendencies of D/S/I obtained with AFE in retraining mode 149

List of Tables

2.1 Performance ratio and number of the hidden units 30

2.2 Joint neural networks classifier using time-scale features. 36

2.3 Neural network classifier using MFCC features. 36

2.4 Confusion matrix of joint FNN using time-scale features 37

2.5 Confusion matrix of FNN using MFCC 37

2.6 Recall and precision of time-scale features and MFCC features 39

2.7 Bayesian network using time-scale features. 39

2.8 Bayesian network using MFCC features. 40

2.9 Confusion matrix of BN (TAN-OMI-CL) using time-scale features . . . 41

2.10 Confusion matrix of BN (TAN-OMI-CL) using MFCC 41

3.1 Average classification rates (%) in case of white noise. 58

3.2 Average classification rates (%) in case of car noise. 58

3.3 Average classification rates (%) in case of factory noise. 58

3.4 Confusion matrix in case of additive white noise 59

3.5 Confusion matrix in case of additive car noise 60

3.6 Confusion matrix in case of additive factory noise 60

3.7 EER without using VAD and with proposed VAD 67

4.1 The setting of frame and buffer sizes 92

4.2 Means and standard deviations of overall evaluation based on CCR. . . 103

5.1 Build critical subbands 122

5.2 WRR/ACC derived by different noise reduction algorithms 124

5.3 Recognition performance as WRR/ACC using PSWF. 126

D.1 Assigned phonetic classes for evaluating with CDT. 139

D.2 Comparison Category Rating (ITU-T P.800) 140

D.3 Overall quality test ITU-T P.85 141

E.1 Best configurations of factors for different front-ends 145
E.2 Recognition performance obtained from the baseline. 147
E.3 Recognition performance obtained from standard front-end SFE. 148
E.4 Recognition performance obtained from advanced front-end AFE. . . . 149

Abbreviations

AFE	Advanced distributed speech recognition front-end
AMR	Adaptive multirate
AQF	Adaptive quantile filtering
ASR	Automatic speech recognition
AST	Adaptive slope tracking
ATC	Air traffic control
CBW	Critical bandwidth
CCR	Comparison category rating
CDT	Comparison diagnostic test
CL	Conditional likelihood
CMI	Conditional mutual information
CWT	Continuous wavelet transform
DET	Detection error trade-off
DFT	Discrete Fourier transform
DSR	Distributed speech recognition
DWT	Discrete wavelet transform
EER	Equal error rate
ETSI	European Telecommunications Standards Institute

FNNs	Feed-forward neural networks
GCI	Glottal closure indices
GMMs	Gaussian mixture models
HA	Hearing aid
HE	Histogram envelope
HMMs	Hidden Markov models
LPC	Linear prediction coefficients
MAP	Maximum a posteriori
Mel IDCT	Mel-warped inverse cosine transform
MFCC	Mel-frequency cepstral coefficients
ML	Maximum likelihood
MMSE	Minimum mean square error
MRA	Multiresolution analysis
MTD	Multi-threshold decision
NB	Naive Bayes
NSS	Nonlinear spectral subtraction
NSS-MM	Nonlinear spectral subtraction-Minimum statistics
OMI	Order mutual information
PSD	Power spectral density
PSWF	Perceptual statistical wavelet filtering
PWPD	Perceptual wavelet packet decomposition
SegSNR	Segmental signal-to-noise ratio
SFE	Standard front-end
SNRs	Signal-to-noise ratios

SR	Speech recognition
SS	Spectral subtraction
STFT	Short-time Fourier transform
SV	Speaker verification
SWF	Statistical wavelet filtering
TAN	Tree augmented naive Bayes
TDC	Threshold dependent curve
TEO	Teager energy operator
TF	Time-frequency
TSF	Time-scale features
UBM	Universal background model
VAD	Voice activity detection
WPD	Wavelet packet decomposition
WT	Wavelet transform

Chapter 1

Introduction

Speech communication is the interdisciplinary subject of describing the information transfer from one person to another via speech signals. Modern speech processing technology is usually considered to comprise several sub-fields of speech analysis, speech synthesis, speech recognition, speech coding, and speech enhancement. To be considered as an essential framework for coding, synthesis and recognition of speech, speech analysis methods study speech production, process the acoustic waveform, and extract interesting acoustic features. Speech synthesis and speech recognition establish a two-way communication between human beings and machines. They have been more and more widely used in many useful applications such as dialog, user interfaces, security systems, machine translation and understanding based on voice communication. While speech coding techniques are the most efficient processes for transmission and storage of speech signals for communication between humans, speech enhancement is motivated by the complicated and significant impacts of realistic environments which may distort speech quality and lead to system performance degradation.

All addressed topics make robustness to acoustic background noise be highly challenging in speech communications. A comprehensive state-of-the-art research of the techniques in robust speech processing have been motivated by the increase of the need for low-complexity and efficient speech feature extraction methods, the need for enhancing the naturalness, acceptability and intelligibility of the received speech signal corrupted by environmental noise, and the need of reducing noise for robust speech recognition systems to achieve high recognition rate in harsh environments. In this dissertation, these challenges are studied by novel methods which are based on the wavelet transform. They are designed for robust speech processing and applications.

The next section is used to review briefly state-of-the-art speech technology, thereby addressing growing challenges in the field. Advantages of the wavelet transform as well as interpretative explanations for its application in robust speech processing are presented in the next two sections. Finally, the thesis contents and work contributions are presented.

1.1 Robust speech processing and applications

In last two decades, speech coding was developed following two different approaches as vocoders and waveform coders. They both shows the trade-off between speech quality and bit rate. Code-excited linear prediction [SA85], multi-pulse excitation [AR82] algorithms are some of the current generation coders in the time domain. The multiband excited coders [GL88], sinusoidal/harmonic [AS84, MQ86] can be considered as the ones in the frequency domain. The application of these speech coders, however, are limited due to the presence of acoustic background noise which can substantially degrade the performance of them in real communication environments. To deal with the problem, besides bandwidth extension and multi-channel speech enhancement approaches, many single-channel speech enhancement techniques were proposed to enhance the recorded noisy speech signal. The most conventional procedure proposed in some patents [HC95, ENS02] is the pre-processing of the input signal by applying noise reduction methods in order to enhance the speech quality before the speech coding stage [BSN03].

The usage of voice activity detection (VAD) is very necessary and is considered a classic approach to estimate the noise level [Coh03] in most of the noise reduction methods. The automatic speech/non-speech distinction and phonetic classification are the most crucial topics in speech processing and applications. In some speech coding systems, the optimal bit allocation is dependent on the different phonetic types of speech frames as voiced sound or unvoiced sound [KAK93, Kle93]. The discrimination between phonetic classes improves quality and performance of data-driven speech synthesizers by adjusting non-uniform scaling factors of each phonetic class in time-scale modification algorithms [KK94, DJC03]. Besides, the phonetic alignment of huge databases can be performed faster by applying the phonetic classifier as a pre-classification step. Finally, there is a need for a phonetic classifier in automatic speech recognition to improve the performance of end-point detection [RSB$^+$05] in order to increase the recognition rate.

The categorization of various speech enhancement systems is based on different approaches of deriving information about the speech and the noise. Some methods employ basic underlying principle that waveforms of voiced sounds are almost periodic. Then adaptive comb filtering [FSBO76] is applied on the noisy speech signal to eliminate the non-harmonic components which are considered as noise. Other methods rely on concepts of human perceptual criteria [CO91, JC03], and speech production [LO79, YMPSR02]. Some systems are based on estimation of speech characteristics in the time and frequency domains derived from the short-time Fourier transform (STFT) with various methods such as spectral subtraction [Bol79], nonlinear spectral subtraction [BSM79, MM80, EM83], Wiener filtering [LO79, SF96, AC99], and subspace-based method [ET95]. Other systems use statistical models to exploit a-priori signal information such as [Dru68, EM84] with the assumption that the Fourier coefficients of speech signals follow a Gaussian distribution, or [LV03, Mar02, Mar05] with the prior assumption of different super-Gaussians such as Gamma distribution for speech, and Laplace or Gaussian distribution for noise. The more sophisticated statistical models such as Gaussian mixture models (GMMs) [BG02], hidden Markov models (HMMs) [Eph92a, SSDB98], and codebooks [Sri05] which are trained on selected databases increase accuracy at the cost of a high computational complexity. The noise removal takes palce not only in the time-frequency domain but also in the time-scale domain by applying the wavelet transform. A survey on this later approach will be presented in the next sections. Thanks to the assumption of uncorrelatedness between speech and noise, and to the assumption that noise is more stationary than speech, the higher precision we model the speech/noise, the more ability to separate noise from speech. However, overmodeling also makes the speech enhancement algorithms inaccurate due to the wide variation of speech signals and noise signals in the real word. The assumptions do not hold in case of many real noise types, which are non-white and very non-stationary, such as machine, car, babble, cafeteria noise or human conversation interference. This trade-off can be addressed by designing the speech enhancement system for specific environments and applications. More precisely, the speech enhancement system can be optimized based on perceptual criteria, or more on mathematical criteria, or somewhere in the middle which is motivated by machine learning applications such as speech recognition.

With the growing demands of automatic human-machine interaction applications, many automatic speech recognition (ASR) systems such as HTK [YEG+05], CMU SPHINX [LHR90, CMU06] have been developed for large vocabulary continuous speech. Together with PC-based applications, speech recognition has become widely used for mobile communications. Due to the nature of mobile technology, the communication processes are often carried out in real environments such as office, car, constructions, etc., which exhibit a wide range of potential sources of noise and distortion that can degrade the quality of the speech signal. This makes robustness to environments for ASR a highly important research topic. There may be three different approaches towards robust ASR such as noise suppression [Ace90], robust feature extraction [Ohs93] and acoustic model adaptation [GY96]. The Aurora Distributed Speech Recognition (DSR) working group [Pea00] of the The European Telecommunications Standards Institute (ETSI) works on the development and standardization of algorithms to parameterize a representation of speech which is suitable for distributed speech recognition. This project can be considered as the integration of the first and second approaches to make ASR environmentally robust: the advanced distributed speech recognition front-end (AFE) [ETS03]. To overcome the addressed problems of the ASR systems using a single microphone, the approach based on single-channel noise reduction is considered low-complexity and effective method to clean up speech recorded in noisy environments, and also compensates the mismatch between speech recognition training and testing conditions.

1.2 Wavelet transform

The development of the wavelet transform (WT) is considered a revolution of modern signal processing techniques and most widely used over the past two decades. Significant contributions of wavelet analysis to different signal processing techniques have been developed for signal, image and speech processing as well as applications. Let $\psi_{s,u}(t) = (1/\sqrt{s})\psi((t-u)/s)$ be a set of continuous-time wavelet basis functions which is generated by scaling the mother wavelet $\psi(t)$ by s and translating it by u. The continuous wavelet transform (CWT) of any signal $x(t)$ is defined as:

$$CWT_x(s,u) = \int_{-\infty}^{+\infty} x(t)\psi_{s,u}(t)dt = \int_{-\infty}^{+\infty} x(t)\frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)dt, \tag{1.1}$$

The continuous wavelet transform has redundancy due to the continuous values of s and u. By sampling them as $s = s_0^m$ and $u = nu_0 s_0^m$, $m, n \in \mathbb{Z}$, we obtain a set of discrete-parameter continuous-time wavelet basis functions $\psi_{m,n}(t) = s_0^{-m/2}\psi(s_0^{-m}t - nu_0)$. If this set is complete in $L^2(R)$ for some choice of $\psi(t), s_0, u_0$, then any continuous-time signal $x(t) \in L^2(R)$ can be represented as the following superposition:

$$x(t) = \sum_m \sum_n d_{m,n}\psi_{m,n}(t), \tag{1.2}$$

where $d_{m,n}$ are the wavelet coefficients for all $m, n \in \mathbb{Z}$ and estimated by:

$$d_{m,n} = \int_{-\infty}^{\infty} x(t) s_0^{-m/2}\psi(s_0^{-m}t - nu_0)dt. \tag{1.3}$$

With $s_0 = 2$ and $u_0 = 1$, we obtain the dyadic-parameter wavelet basis functions $\psi_{m,n}(t) = 2^{-m/2}\psi(2^{-m}t - n)$ which are considered in our research [AH01]. As we see, the pure wavelet expansion in Equation 1.2 requests an infinite number of scales or resolutions m to represent the signal $x(t)$ completely. This is impractical. If the expansion is known only for certain scale $m < M$, we need a complement component to present information of expansion for $m > M$. This is done by introducing a scaling function $\phi(t)$ such that, $\forall m \in \mathbb{Z}$, the set $\phi_{m,n}(t) = 2^{-m/2}\phi(2^{-m}t - n)$ is an orthonormal basis for subsapce V_m of $L^2(R)$. With the introduced component, the signal $x(t) \in L^2(R)$ can be represented as a limit of successive approximations corresponding to different resolutions [AH01]. This presentation is named as multiresolution analysis (MRA) [Mal89, VK95]. In other words, the signal $x(t)$ is presented as the sum of an approximation plus M details at the M^{th} decomposed resolution:

$$\begin{aligned} x(t) &= \sum_n a_{M,n}\phi_{M,n}(t) + \sum_{m=1}^{M}\sum_n d_{m,n}\psi_{m,n}(t) \\ &= \sum_n a_{M,n}2^{-M/2}\phi\left(\frac{t}{2^M} - n\right) + \sum_{m=1}^{M}\sum_n d_{m,n}2^{-m/2}\psi\left(\frac{t}{2^m} - n\right), \end{aligned} \tag{1.4}$$

where M represents the number of scales. $a_{m,n}$ and $d_{m,n}$ are the approximation or scaling coefficients and the detail or wavelet coefficients. In discrete-time domain, the set of discrete-time scaling functions and wavelet functions can be constructed from filter bank as:

$$\phi_{m,n}(l) = \sum_p h(p - 2n)\phi_{m-1,n}(2^{-(m-1)}l - n), \tag{1.5}$$

$$\psi_{m,n}(l) = \sum_p g(p - 2n)\phi_{m-1,n}(2^{-(m-1)}l - n), \tag{1.6}$$

where $h(n)$ and $g(n)$ form a pair of conjugate mirror filters used at the analysis stage with $g(n) = (-1)^{1-n}h(1-n)$ [Mal99]. Based on these bases, the scaling coefficients $a_{m,n}$ and wavelet coefficients $d_{m,n}$ are derived by the discrete convolutions:

$$a_{m,n} = \sum_p a_{m-1,p}h(p-2n) = a_{m-1} * \overline{h}(2n), \tag{1.7}$$

$$d_{m,n} = \sum_p d_{m-1,p}g(p-2n) = d_{m-1} * \overline{g}(2n), \tag{1.8}$$

where $h(-2n) = \overline{h}(2n)$ and $g(-2n) = \overline{g}(2n)$ are synthesis filters. As a generalization of discussed wavelet decomposition, wavelet packet expands a range of signal analysis by doing additional implementations of wavelet decomposition on detail coefficients. In this study, we only consider the binary wavelet packet decomposition (WPD). Each packet node (m, k) corresponds to a space W_0^0 which is spanned by an orthonormal basis $\{\psi_m^k(2^{-m}l - n)\}_{n\in\mathbb{Z}}$, with $k = 1, \ldots, 2^m$ the packet channel index. Assuming that we already construct the basis at node (m, k), the two wavelet packet bases at the children nodes are calculated by:

$$\psi_{m,n}^{2k}(l) = \sum_p h(p)\psi_{m-1,n}^k(2^{-m}l - n), \tag{1.9}$$

$$\psi_{m,n}^{2k+1}(l) = \sum_p g(p)\psi_{m-1,n}^k(2^{-m}l - n), \tag{1.10}$$

The corresponding wavelet packet coefficients are derived as:

$$d_{m,n}^{2k} = \sum_p d_{m-1,p}^k h(p-2n) = d_{m-1}^k * \overline{h}(2n), \tag{1.11}$$

$$d_{m,n}^{2k+1} = \sum_p d_{m-1,p}^k g(p-2n) = d_{m-1}^k * \overline{g}(2n), \tag{1.12}$$

From now on, we call $X_{m,i}(n)$ the sequence of all wavelet coefficients (i.e. $a_{m,n}$ and $d_{m,n}$) which are derived by the DWT at the m^{th} scale of the i^{th} frame, n is the coefficient index. Let denote N_f be the number of samples in one speech frame, and also the number of coefficients in $X_{m,i}(n)$. Thus, $N_m = \dfrac{N_f}{2^m}$ is the number of coefficients in corresponding subbands at the m^{th} scale. In case of applying the WPD, $X_{m,i}^k(n)$ describes a sequence of wavelet packet coefficients (i.e. $d_{m,n}^{2k}$ and $d_{m,n}^{2k+1}$) derived at the m^{th} scale of the i^{th} frame. As studied later on in chapter 4, the WPD is implemented at a fixed decomposition scale $m = 7$ for an application, so m is discarded in the notation, and superscript k is become supscript k to simplify the notation of wavelet packet coefficients as $X_{k,i}(n)$.

Implementation of the wavelet transform by the way of multiresolution analysis is applied successfully in image compression as JPEG 2000 [Gro00], and considered as potential technique for speech coding and speech compression. It is known that the WT is suitable to analyze speech signals which are considered as non-stationary signals. In addition to the MRA ability, wavelet shrinkage with plentiful mathematical advantages has become a powerful and promising technique of removing noise for image and speech signals. The decomposition and reconstruction of the signal can be implemented efficiently by using convolutions with quadrature mirror filters in a pyramidal algorithm [Mal89]. With the wavelet representation, there is no redundant information because of the orthogonality of the wavelet basis. It is essential to select wavelet basis with desired properties such as *vanishing moments* and *compact support*. Number of vanishing moments P and compact support size of a wavelet basis ψ influent to the sparsity of the wavelet representation which is essential to the performance of noise removal and data compression.

If a wavelet function ψ that its $\Psi(\omega)$ is P times continuously differentiable at $\omega = 0$, it has at least P vanishing moments, while the converse statement is not true. So we have a loose relationship between the smoothness or regularity of the wavelet functions with the number of vanishing moments [Mal99]. If the analyzed signal $x(l)$ is the sum of a polynomial and some localized singularities, the wavelet with suitable order of vanishing moments will decompose the signal into two different parts clearly: the details reflecting the singular components of the signal and the polynomial approximation. Conversely, a wavelet with lower order leads to a interference of the approximation into the details which results in worser ability of analyzing singularities [MH92]. In other words, the wavelet kills all polynomials of degree smaller than number of vanishing moments and results in a sparse representation for piecewise smooth signals. This means if the analyzed signal $x(l)$ is regular and ψ has enough number of vanishing moments, then the wavelet coefficients $d_{m,n}$ at fine scale 2^m are small. Of course, the price to pay for having a large number of vanishing moments is that the basis functions will be less localized. Some experimental research suggests that the number of vanishing moments required depends heavily on the application. A more information on the number of vanishing moments are presented in appendix A.

The wavelets having a compact support are useful in local analysis. Wavelets with small support size are good because they can pick up isolated singularities and are fast to compute. Moreover, by reducing the support size, we can minimize the number of large coefficients which makes more sparse presentation of the analyzed signal. Consequently, the narrowness in time domain results in a very low resolution in frequency domain. Conversely, wavelets with large compact support are more regularity or smoother due to high number of vanishing moments. By designing the synthesize conjugate mirror filters $h(n)$ with as few non-negligible coefficients as possible, the support size of wavelet basis $\psi_{m,n}(l)$ is minimized [Mal99].

Wavelet thresholding/shrinking is a fully promising tool to remove noise from an observed noisy signal. The principle is based on thresholding or shrinking the wavelet coefficients towards zero. Due to the decorrelation property of the DWT, the noise is spread out over all wavelet coefficients. This means the DWT leads to a sparse representation which allows to replace the noisy coefficients by zero. Hard and soft thresholding are proposed by [DJ94] as the simple and effective denoising functions. A modification of hard and soft thresholdings which provide a smoother function is studied in [ZD97, ZL99]. A so-called wavelet firm shrinkage which generalize hard and soft thresholding is proposed in [BG97]. There are many procedures to calculate thresholds related to noise levels as minimax threshold proposed in [DJ98], universal threshold [DJ94, Don95], or SURE (Stein's unbiased risk estimator) threshold [DJ95].

1.3 Why wavelets for speech processing ?

Based on the MRA, a signal is decomposed into an approximation and details at various scales. In other words, various information levels across successive resolutions can be extracted by decomposing the original signal using a wavelet orthonormal basis. As a conventional transform, the well-known short-time Fourier transform (STFT) is used widely in mathematics and engineering. A limitation of the STFT is that, because a single window is used for all frequencies, the resolution of the analysis is the same at all locations in the time-frequency plane [VH92]. This limitation represents a handicap in speech and audio signals since human hearing system uses a frequency-dependent resolution. The DWT can solve this drawback with the rectangular tiling of the time-frequency plane.

In mathematics, research on singularities and irregular structures is very necessary because they often carry the most useful information in signals (e.g. transient, discontinuous, and non-stationary sounds). The greatest challenge is the selection of appropriate techniques which are able to study irregularities of signal structures. Until now, the Fourier transform was the main mathematical tool for analyzing singularities. However, the Fourier transform with sinusoidal waveforms extending over a fixed window length provides only a description of the global regularity of signals without well adapting to the localization of singularities in the time-frequency domain. This motivates the study of the wavelet transform which can characterize the local regularity of signals by decomposing signals into well-localized time-frequency components. As proved in [MH92], the detection of all the singularities of the signal are based on the local maxima property which is measured from the evolution across scales of these local maxima. The detection of singularities with multiscale transforms has been studied not only in mathematics but also in signal processing and application domains.

The flexible analysis in the time-frequency plane of the DWT in comparison with the STFT [VH92, MH92] shows its advantages for speech processing. With Heisenberg's uncertainty principle, it is known that no transform can provide high resolution in both time and frequency domains at the same time. The useful locality property is exploited in this context. Because the wavelet basis functions are short waves and generated by scaling from the mother wavelet, they are well-localized in time and scale domains. This automatic behavior of wavelet decomposition is absolutely suitable for processing of speech signals which requires high frequency resolution to analyze low-frequency components (voiced sounds, formant frequencies), and high temporal resolution to analyze high-frequency components (mostly unvoiced sounds). This smart behavior of the DWT is employed for auditory representations of acoustic signals [YWS92], speech coding [Lit98, CD99] using wavelet packet representation in the context of auditory modeling, speech segmentation [TLS$^+$94] and phonetic classification [CG05, PK04, PK05a].

In recent years, the wavelet shrinking approach to speech enhancement has been developed rapidly, starting with the simple hard and soft thresholdings proposed in [Don95]. Many improvements of wavelet thresholding to enhance speech signal have been done as semisoft thresholding with selected threshold for unvoiced regions [SB97], efficient hard and soft thresholdings [SMM02], smooth hard thresholding function based on μ-law [SA01, CKYK02, PK05b]. In [LGG04], the combination of soft and hard thresholding is applied to adapt with different properties of the speech signal. Motivated by lower speech distortion, wavelet thresholding/shrinking methods are integrated with other techniques such as the Teager energy operator and masked adaptive threshold [BR01]. Critical-band wavelet decomposition is used with noise masking threshold in [LW03], and perceptual wavelet packet decomposition (PWPD) which simulates the critical bands of the psychoacoustic model is proposed in [JM03, CW04]. A blind adaptive filter of speech from noise is designed in the wavelet domain in [VG03].

Dealing with enhancement and feature extraction for robust ASR, several parameterization methods which are based on the DWT and wavelet packet decomposition (WPD) have been proposed in [GT00, GG01]. More sophisticated shrinking functions with better characteristics than soft and hard thresholding are optimized for speech enhancement and speech recognition [KKH03]. The usage of wavelet-based features extracted from the WPD leads to improvement of recognition rate compared with the well-known conventional feature Mel-frequency cepstral coefficients (MFCC) [SPH98, CKYK02, Kot04]. In addition to this approach, wavelet denoising is applied as pre-processing stage before feature extraction to compensate environmental mismatches [BE97, FD03].

1.4 Outline of the thesis

Motivated by challenging topics of modern speech technology such as speech classification, speech enhancement for hearing aids and noise reduction for speech recognition, wavelet analysis with its technically powerful characteristics is exploited to develop novel methods for improving the performance and robustness of the addressed systems. Analysis and design of such systems are discussed in detail in the following chapters:

Time-scale features for phonetic classification

In chapter 2 some common methods of phonetic classification are reviewed by concentrating on the ways of extracting features and of learning classification models. Based on the analysis of phonetic characteristics of voiced, unvoiced, silence, mixed-excitation, voiced closure, and release of plosive sounds, the DWT is applied to extract time-scale features which characterize the interesting phonetic classes. To learn the classifiers in feature space, the feed-forward neural network and the Bayesian network are selected for their potential ability for pattern recognition. The processes of designing, training and testing of the classifiers based on wavelet features are discussed in detail. This study shows the effectiveness of the wavelet features compared with the baseline features and opens an approach to the design of a robust phonetic classifier in the next chapter.

Robust speech classifiers for adverse environments

Unlike the study in chapter 2, which is applied for noise-free speech signals, the robustness of speech classification against environmental noise is addressed in chapter 3. In detail, robust detection of speech/non-speech as well as classification of voiced/ unvoiced/ silence periods is developed to meet the demands of realistic environments. The time-scale features proposed in chapter 2 are adjusted and further enhanced by the non-linear Teager energy operator. Quantile filtering and slope tracking are designed as two advanced methods for deriving adaptive decision threshold. The robustness of the novel phonetic methods is proved by the evaluation on a speech database artificially contaminated by different kinds of realistic noise. The results show better performance compared with other methods. A further evaluation is made in the application domain where a voice activity detection (VAD) is developed from the outputs of the phonetic classifier and applied as pre-processing stage of a speaker verification system. An improvement of verification rate confirms the robustness and efficiency of the proposed method in harsh environments.

Statistical wavelet filtering for speech enhancement

Besides the advantage of flexible time-frequency analysis which is employed for speech classification, the powerful ability of noise removal by wavelet shrinking is exploited for speech enhancement. In chapter 4, the demand of improving quality, acceptability and intelligibility of noisy speech is addressed by a novel single-channel speech enhancement method. Statistical wavelet filtering is proposed to eliminate musical noise as well as to handle colored and non-stationary noise. The denoising process is done in the wavelet domain by appyling the WPD on the noisy speech signal. The estimate of the threshold relating to the noise level is implemented by the quantile filtering technique over recursive buffers which results in a more accurate and adaptive threshold. Non-stationary and non-white noise is handled by the nonlinear adaptive weighting functions in time and frequency domains. The suppression rule is based on the smoothed hard shrinking function with an adaptive factor to remove noise effectively while maintaining the pleasantness of the processed speech signal. In addition to standard subjective tests, a comparison diagnostic test is proposed to obtain more specific insight from evaluation studies of speech enhancement systems.

Noise Reduction For Robust Speech Recognition

Chapter 5 deals with robust automatic speech recognition (ASR) in adverse environments. The proposed noise reduction method in chapter 4 is employed as pre-processing stage in the font-end unit. Thus, the quality of the recorded speech signal is enhanced to ensure that proper information will be extracted by the feature extraction stage after that, thereby increasing the recognition performance of the ASR system. The proposed statistical wavelet filtering method is further optimized for achieving robust word recognition performance by changing the shape of the frequency weighting function and estimating noise thresholds for critical subbands. By integrating the proposed denoising process into the ASR training phase, the retrained models provide higher recognition rates. As improving the speech recognition rate is not our only interest, the study also investigates whether an increase in perceptual quality of a speech signal leads to an increase in recognition rate of the ASR system or not. As a last study issue, the need to employ noise reduction to compensate the mismatch between the training and testing phases of the ASR system is examined by experiments in different training/testing conditions which include complex noise from harsh environments.

Conclusions and perspectives

Chapter 6 summarizes the proposed approaches with more discussion and conclusions. Finally, we address some open issues that should be studied further as well as possible applications of the proposed robust speech processing methodologies in the next generation of modern speech technology.

1.5 Work contributions

The thesis mainly deals with application of wavelet analysis to a wide range of speech applications such as speech classification, speech enhancement, and speech recognition. The scientific contributions are discussed in the following.

At the beginning of the doctoral research, a significant effort was spent for constructing phonetic classifiers. The goals were firstly to analyze acoustic-phonetic properties of interesting phonetic classes, and secondly to exploit advantageous characteristics of wavelet analysis for extracting time-scale features which represent the corresponding phonetic classes with high confidence. Finally methods from simple multi-threshold decision models to modern pattern recognition methods were studied in order to build the effective classifiers. This first work contributions were published in [PK04, PK05a, PP06a]. Currently, we are developing an advanced phonetic classifier by extracting a better time-scale feature set. The improved results are reported in a submitted journal article [PP06b].

As a second contribution, a robust and low-complexity speech classifier was designed and applied successfully as a pre-processing unit of a speaker verification system. By considering the acoustic properties and spectrogram of noisy voiced and unvoiced segments, we extracted a single time-scale feature which is very robust against background noise after enhancement by applying a hyperbolic tangent sigmoidal function and median filtering. As one of the significant inventions, the quantile filtering technique is designed to estimate an adaptive decision threshold accurately. The experimental results showed that the classification performance is high and robust due to this effective estimation. Its variant with adaptive quantile factor was used to design a robust voice activity detector which is integrated as a pre-processing stage of the speaker verification system. The improved verification rate of the system in a harsh environment that simulates air traffic communication confirms its robustness. To meet delay and memory requirements of real-time applications, the slope tracking method was proposed to classify phonetic classes almost online. Though this method has not been applied for any real-time applications yet, we definitely believe that it is feasible and beneficial. These achievements were reported in [PKW$^+$06, PK06b, NPK07], and were recently contributed as a part of a book chapter [NPHKon].

Combining the quantile filtering technic with optimal wavelet shrinkage to design a novel speech enhancement algorithm - statistical wavelet filtering (SWF)- is considered as a third contribution of the thesis. The superior performance obtained in subjective tests proved its effectiveness in comparison with state-of-the-art single channel speech enhancement methods operating in Fourier domain. Noise thresholds are estimated accurately and adaptively for every frequency channel by quantile filtering applied over recursive buffers. A lot of effort was also invested in designing the nonlinear adaptive weighting functions in both time and frequency domains to handle the non-stationary and non-white noise effectively. By introducing the adaptive factor, the smoothed hard shrinking gain function can remove noise effectively while maintaining the pleasantness of the enhanced speech signal. As an additional contribution, we built the Comparison Diagnostic Test to perfect the evaluation of speech quality. A detailed description and application of the proposed system were published in [PK05b]. The application of SWF in the denoising of an office meeting database helps to improve the performance of VAD and direction of arrival estimation. These works, which were implemented for the MISTRAL project [mis] are not presented in the thesis, but reported in [KPK+06, KPN06].

The final contribution is the optimization of the statistical wavelet filtering to enhance speech quality for speech recognition in adverse environments. The perceptual SWF method was developed by applying full WPD and estimating of noise thresholds at critical wavelet subbands. Besides a suitable frequency weighting function was designed, the parameters of time-frequency weighting functions were turned to optimize recognition performance. As presented in the thesis, the retrained speech recognizer which employs the optimized SWF as pre-processing stage before front-end extraction provides almost equal word recognition rate compared with the baseline ETSI advanced front-end for the car noise condition. The proposed system was applied in the SNOW project [sno] to improve the performance of a mobile speech recognizer operating in the Airbus maintenance factory. The very first results obtained by applying the SWF and testing with the SNOW database was published in [RPK06]. With the perceptual SWF method, latest achievements which were published in the final report of the SNOW project [SNO06] show that our proposed algorithm is the best candidate among others. This improved system will be described in an in preparation journal article [PRK07].

Chapter 2

Time-scale features for phonetic classification

2.1 Introduction

[1] Automatic speech classification is crucial for different speech processing methods and various speech applications. The performance of concatenative speech synthesis may be improved by selecting proper smoothing strategies at concatenation points [HB96]. Moreover, the discrimination between phonetic classes improves quality and performance of data-driven speech synthesizers by adjusting non-uniform scaling factors of each phonetic class in the time-scale modification algorithm to get better perceptual results [KK94, DJC03]. Some speech coding systems use phonetic classification to determine the optimal bit allocation for every different speech frame [KAK93, ZWC97, O'S00]. In Internet telephony applications, the adaptive loss concealment algorithm uses the voiced/unvoiced detector at the sender [San98]. This helps the receiver to conceal the loss of information based on the similarity between the lost segments and the adjacent segments. Besides, its application to real-time speech transmission on the internet are proposed in [SCJ$^+$02]. By using the phonetic classifier as a pre-classification step, the phonetic alignment of huge databases can be implemented faster, too.

The speech classification task has been studied in many articles by a variety of methods since the 1980's. In principle, the classification is performed by training a model to learn differences of statistical distributions of the acoustic features between

[1] This chapter is based on materials published earlier in [PK04, PK05a, PK06a, PP06a, PP06b].

different phonetic classes [AR76]. These features can be derived by three approaches:

- The first approach works in the time domain and uses statistical measurements. The common features are the zero crossing rate, relative energy level, autocorrelation coefficients, etc. [Ked86, CHL89, LG99]. The calculation of these features is quite simple but they may be damaged in adverse noise conditions.

- The second approach works in the frequency domain. Frequently used features are the spectrum [ZSW97, YVH99], optimal filters [NS02], cepstrum pitch detection [AS99]. As reported, the frequency-based features which carry important characteristics of the speech signal such as fundamental frequency and formants help to increase the performance of speech classifiers and speech recognizers. Mel frequency cepstral coefficients [XH02] which are standard features for speech recognition are commonly used due to their high performance. The DWT is recently applied for the phonetic classification task with promising results [LE03].

- The third approach combines both time and frequency domains by considering the changes in time of the spectrum such as spectral flux [SS97a]. The variation of other features over time were used as smoothing on raw clasification as the hangover scheme in [ETS03]. The 15ms/200ms rule in [Bra68] could be applied for VAD.

Based on the extracted features, the classification can be made by simply applying multi-threshold decision (MTD) classifier [PK04]. The operation of the MTD classifier is based on comparison between observed features of the input speech frames and fixed pre-determined thresholds. The best hard thresholds of the model are found by experimental pattern classification. As an advanced approach, pattern recognition is applied to train models of the feature space for every class statistically. Then the likelihood of the observed feature with respect to the trained models is measured to make the classifying decision. Since the development of the backpropagation learning algorithm [Mit97], feed-forward neural networks (FNNs) have been used widely in pattern recognition. In particular, it has been applied for speech classification with promising potential [MBB96, QH93, GCEJ91]. The usage of hidden Markov models as a representation for phonetic classification and recognition is studied in [DS94]. Considered as an interesting survey, a comparison between Gaussian mixture models (GMM) and FNN has been presented in [LCG93]. In [CG05], a relationship between wavelets and filter banks is employed to design filter banks for feature extraction and classification is done by training GMMs. Recently, application of Bayesian networks

(BNs) in classification task has been studied as another approach. Generative classifiers learn a model of the joint probability of the features and the corresponding class label and perform predictions by using Bayes' rule. The usual approach for learning a generative model is maximum likelihood (ML) estimation [Pea88]. Discriminative classifiers directly model the class posterior probability. Maximizing the conditional likelihood (CL) [GZ02] of the class given the attributes results in optimizing the ability to correctly predict the class. Discriminatively trained Bayesian networks proposed in [PB06] achieve promising results in general classification domains. Though they have not been used for phonetic classification so far.

In this chapter, phonetic classifiers will be designed to segment the input speech signal into phonetic groups which are defined as homogeneous sequences of speech sounds such as silence (S), voiced (V), unvoiced (U), mixed excitation (M) classes, and partially the plosive (P) class by detecting the voiced closure interval (VC) and transient (T) frames. The approach of multi-domain features is employed. Time-scale features are extracted by applying the DWT to every overlapped speech frame multiplied by a window. Phonetic decisions are made by employing different classifiers such as feedforward neural networks and Bayesian networks. To achieve robust classification, two FNNs are trained to learn the optimal thresholds. Two FNNs with two layers each operate on input features extracted from speech frames by DWT and statistical measurements in order to classify these frames as transient, voiced, unvoiced and mixed excitation classes. After that, a linear classifier for detecting the silence and voiced closure interval classes are jointly combined with the trained FNNs to build an efficient multi-threshold FNN classifier. As another method to obtain statistical classifiers, both discriminative parameter learning by optimizing the CL and generative parameter training (ML estimation) on both discriminatively and generatively structured Bayesian networks are applied. The naive Bayes (NB) and the tree augmented naive Bayes (TAN) classifiers are used. The proposed classification systems and some other methods are evaluated with the TIMIT database. The extracted wavelet feature sets are compared with the baseline MFCC in terms of classification rates. Gender dependent and gender independent classifiers are studied to assess the impact of gender dependency to classification rate. Further discussion on the effectiveness and limitations of each system is given finally.

The chapter is organized as follows: the next section describes the speech characteristics and the application of the DWT for the speech classification task. The two following sections show the processes of designing, training and testing of two classifiers based on the FNN and the BN. The evaluations and discussion are reported after that, and the final section presents a conclusion and perspectives for future research.

2.2 Time-scale features for speech classification

2.2.1 Acoustic-phonetic properties

Speech is the audible representation of the language system and composed of a sequence of sounds. In order to understand how sound is generated, the study of acoustic properties is necessary [OGC93]. One can classify speech signals into five acoustic classes:

- Silence shows up as a blank section in spectrograms with very low amplitude levels.
- Periodic sound refers to the repeating structure of the speech wave caused by the release of pressure pulses at regular intervals. Each pulse has approximately the same amplitude, and the same time period.
- Noise-like sound refers to the irregular structure of the pressure wave caused by turbulent airflow. The spectrogram appears as random over a wide range of frequencies.
- Mixed-excitation sound is created from both sources of periodic and noise-like pressure waves simultaneously.
- Transient sound is formed when a high-pressure jet of air releases into lower pressure areas, it creates a single shock-wave that travels outwards.

In this research, we want to classify five types of phonetic groups which are homogeneous frame sequences having the same phonetic characteristics as classified in Fig. 2.1. We do not classify other non-continuant such as affricates. We consider a stop as a group consisting of all parts of a plosive: closure interval, transient and release. Besides the silence group, other phonetic groups are classified as follows:

- The voiced group includes vowels, semivowels, diphthongs and nasals which have repetitive time-domain structure and low-frequency voiced striations in the spectrogram.

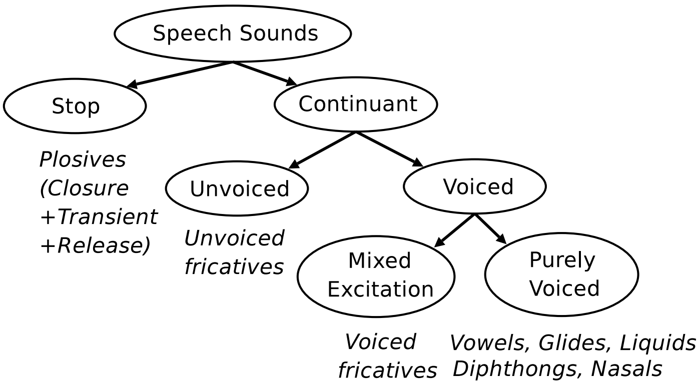

Figure 2.1: Classification diagram for phonetic groups.

- The unvoiced group includes unvoiced fricatives and release part of plosives which have irregular time-domain structure and high frequencies in the spectrogram

- The mixed group includes voiced and glottal fricatives which have both periodic and noise-like properties.

- The stop group contains closure intervals, transient and releases. There are voiced plosives with voiced closure and unvoiced plosives that is closure is similar with silence. Only voiced closure and transient frames are detected in this research.

2.2.2 Voiced/unvoiced/mixed-excitation/silence classification

Depending on the phonetic properties of the input speech frames, the power distribution in different subbands varies. As a primitive analysis, DWT at the 1^{st} decomposition scale is applied on voiced, unvoiced, silence and mixed-excitation speech frames. We observe that the power of wavelet coefficients derived from voiced frames is mostly contained in the approximation part and much less in the detail part as depicted in Fig. 2.2. This is reversed for the unvoiced frames as shown in Fig. 2.3. A relatively equal power distribution occurs for the mixed-excitation and silence frames.

An analysis on intra-scale relationship is done by considering the power variation of different details at different scales. The power of detail coefficients extracted from voiced frames increases from scale 1 to scale 4 as shown in Fig. 2.4 (b). However, this power order occurs vice versa for unvoiced frames which is depicted in Fig. 2.5 (b). There is almost no power change over various scales for mixed-excitation and silence frames. Fig. 2.6 shows the power variation of the detail coefficients which are derived from a speech segment consisting of a sequence of voiced-silence-unvoiced frames. These

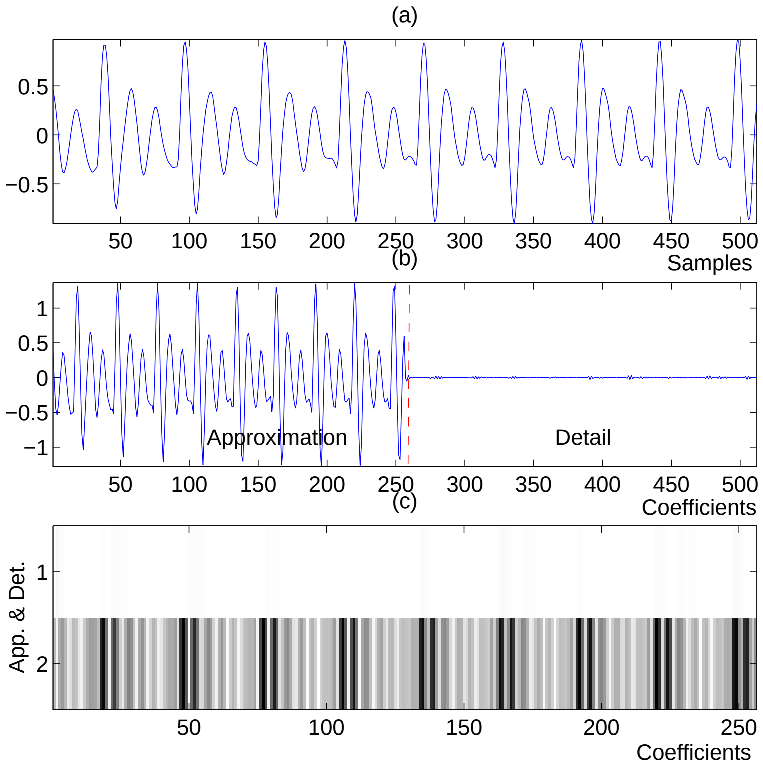

Figure 2.2: (a) Waveform of a voiced segment, (b) Approximation coefficients $a_{1,n}$ and detail coefficients $d_{1,n}$ derived from the DWT of the segment at the 1^{st} scale, (c) $a_{1,n}$ and $d_{1,n}$ in time-scale plane.

behaviors are similar with spectral tilts of the voiced and unvoiced segments as shown in Fig. 2.4 (c) and Fig. 2.5 (c). By comparing the frame-based short-term power values as well, mixed-excitation frames can be distinguished from silence frames. So, these properties can be used in the specific representation to distinguish between voiced frames, unvoiced frames, mixed-excitation frames and silence frames.

2.2.3 Transient detection

From the observation of the first three levels of the wavelet analysis in Fig. 2.6, we can also apply the power variation of detail coefficients for detecting transient frames when considering two neighboring frames. A transient frame always has higher absolute power in its details than a closure interval frame which may be silence or periodic. This characteristic is used to define the closure-transient detail power ratio and it is combined with statistical features to detect transient frames.

In addition to wavelet features, statistical measures in the time domain are further

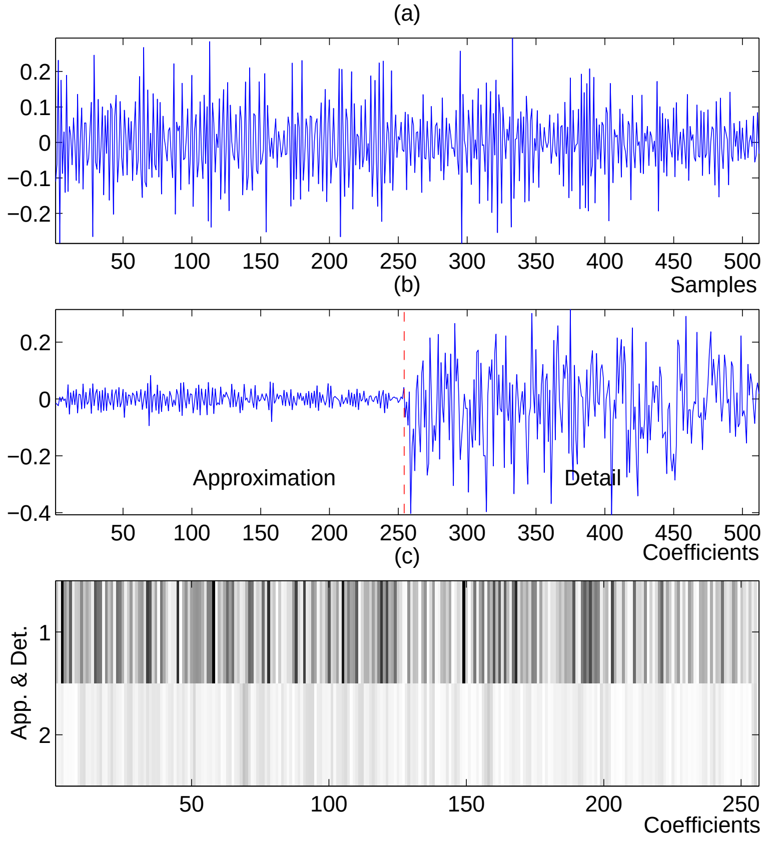

Figure 2.3: (a) Waveform of an unvoiced segment, (b) Approximation coefficients $a_{1,n}$ and detail coefficients $d_{1,n}$ derived from the DWT of the segment at the 1^{st} scale, (c) $a_{1,n}$ and $d_{1,n}$ in time-scale plane.

applied to provide more representative information of every phonetic class. Besides the short-term energy feature, the zero crossing rate (ZCR) which is counted by the number of sign changes between successive samples in a speech frame is also exploited. Clearly, because voiced speech does not change so fast, unvoiced speech always has higher ZCR than voiced speech [Ked86], and this relation occurs between voiced speech and silence, too.

2.2.4 Feature extraction

By applying the DWT at the scale $M = 3$ on each i^{th} speech frame of 16ms frame length and 4ms frame rate, we obtain one approximation subband and three detail subbands which form the sequence of wavelet coefficients $X_{3,i}(n) = \{a_{3,p}, d_{3,p}, d_{2,p}, d_{1,p}\}$, where $m = 3, 2, 1$. The numbers of coefficients in the approximation subband and the three following detail subbands are denoted as

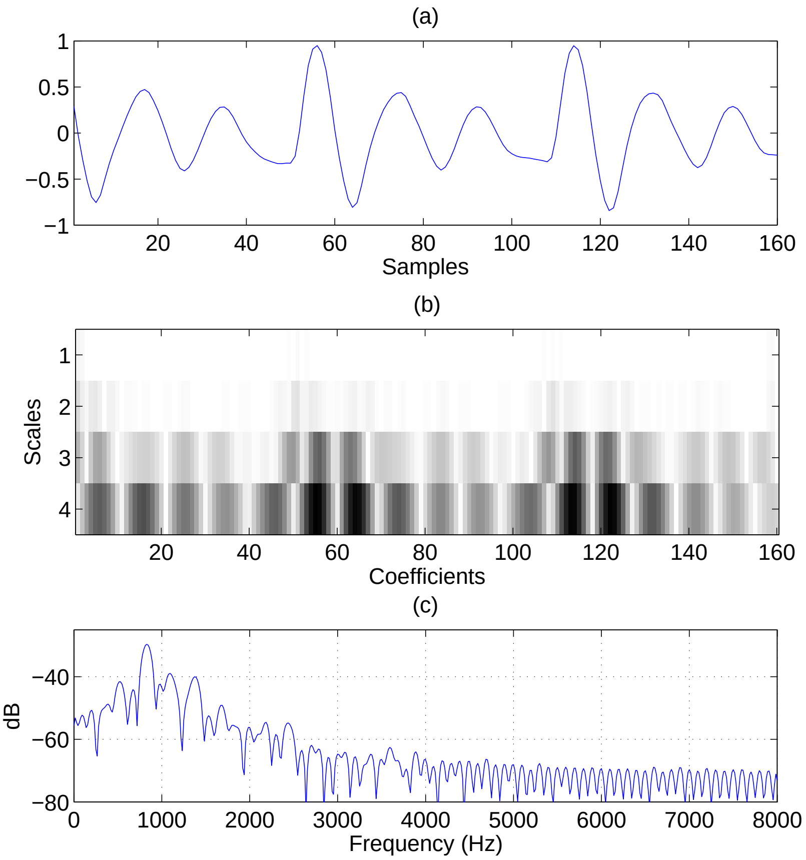

Figure 2.4: (a) Waveform of a voiced frame, (b) Power variation over different scales, (c) Spectral tilt illustrated by DFT.

$\left\{N_3 = \frac{N_f}{8}, N_3 = \frac{N_f}{8}, N_2 = \frac{N_f}{4}, N_1 = \frac{N_f}{2}\right\}$, respectively. The set of features is calculated as follows:

- ***Wavelet power ratio (WPR)*** is the ratio between the power of the approximation coefficients at the 1^{st} scale and the power of all wavelet coefficients:

$$WPR(i) = \frac{2N_3 + N_2 + N_1}{2N_3 + N_2} \frac{\sum_{n=1}^{N_1} X_{3,i}^2(n)}{\sum_{n=1}^{N_1} X_{3,i}^2(n) + \sum_{n=N_1+1}^{N_f} X_{3,i}^2(n)} \tag{2.1}$$

- ***The power variation of detail coefficient (PVD)*** is defined by the difference between the power of the 1^{st} and the 3^{rd} detail coefficients:

$$PVD(i) = \frac{1}{N_1} \sum_{n=N_1+1}^{N_f} X_{3,i}^2(n) - \frac{1}{N_3} \sum_{n=N_3+1}^{N_2} X_{3,i}^2(n) \tag{2.2}$$

- ***Short-term logarithmic average energy (SAE)*** is calculated for every i^{th}

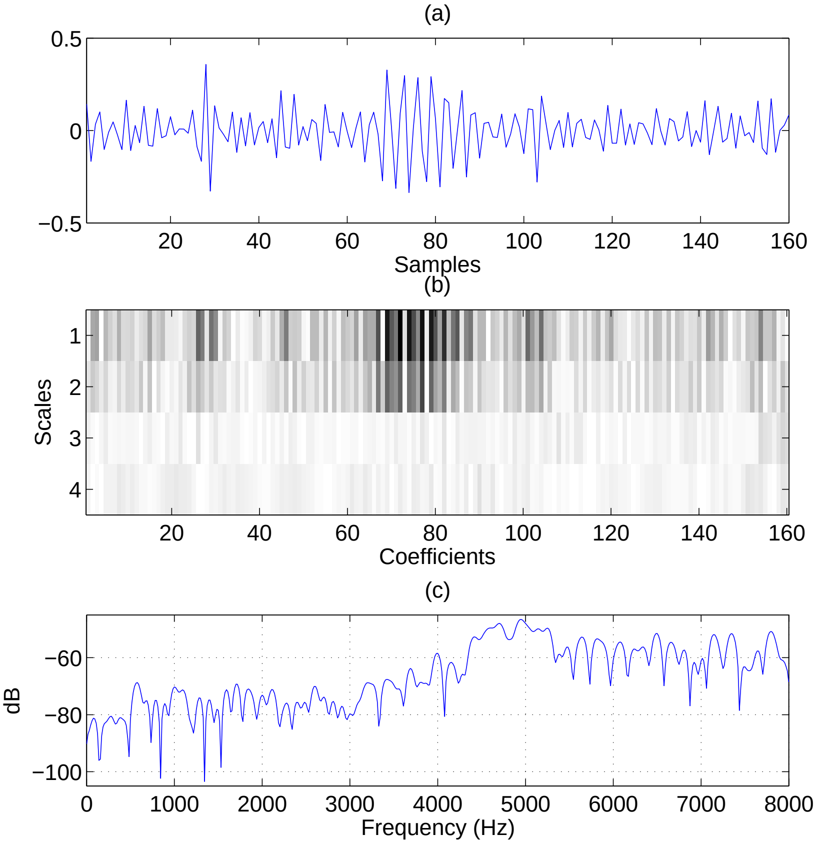

Figure 2.5: (a) Waveform of an unvoiced frame, (b) Power variation over different scales, (c) Spectral tilt illustrated by DFT.

speech frame :

$$SAE(i) = 10 \log \left(\frac{1}{N_f} \cdot \sum_{l=1}^{N_f} x(l)^2 \right) \tag{2.3}$$

- ***Zero crossing rate (ZCR)*** is the number of sign changes of successive samples in i^{th} speech frame:

$$ZCR(i) = \sum_{l=1}^{N_f} |\mathrm{sgn}(x(l)) - \mathrm{sgn}(x(l-1))| \tag{2.4}$$

- ***The closure interval-transient detail ratio (CTDR)*** is the ratio of the detail energy at the same decomposed scale, computed for the current i^{th} frame and the previous $(i-1)^{th}$ frame as shown in Fig. 2.7:

$$CTDR_m(i) = \frac{\sum_{n=N_m+1}^{N_{m-1}} X_{3,i}^2(n)}{\sum_{n=N_m+1}^{N_{m-1}} X_{3,i-1}^2(n)} \tag{2.5}$$

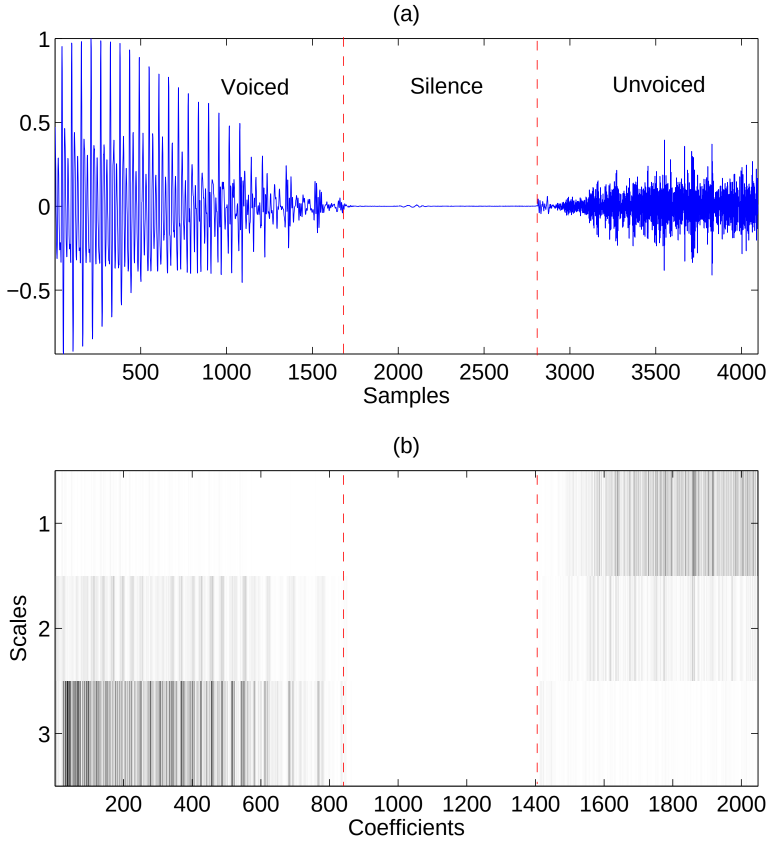

Figure 2.6: (a) A speech segment consisting of voiced, unvoiced and silence frames, (b) Power variation of detail coefficients.

where $m = 1, \cdots, M$, and $M = 3$ is the decomposition scale. When $m = 1$, the upper index becomes $N_{m-1} = N_0 \equiv N_f$.

In total, seven features are estimated for the classification task. The different behavior of the feature values with respect to different phonetic classes of the speech frames is shown in Fig. 2.8, which visualizes the evolution of the first 4 features (the mixed sounds and their features are not shown in this figure). We see that the PVD values are larger than zero for unvoiced frames but less or equal zero for voiced and silence frames. Besides, the energy of the voiced frames is often higher than the one of unvoiced frames, and both of them are higher than that of silence frames. Moreover, the values of WPR are low while the ZCR values are high for unvoiced frames in comparison with voiced and silence frames. The ZCR values are smoothed by 3-point median filtering to reduce the fluctuations of continuous speech. From the analysis of the DWT of speech frames for different phonetic classes, we see that silence, voiced, unvoiced and mixed-excitation groups can be classified directly by measuring the listed

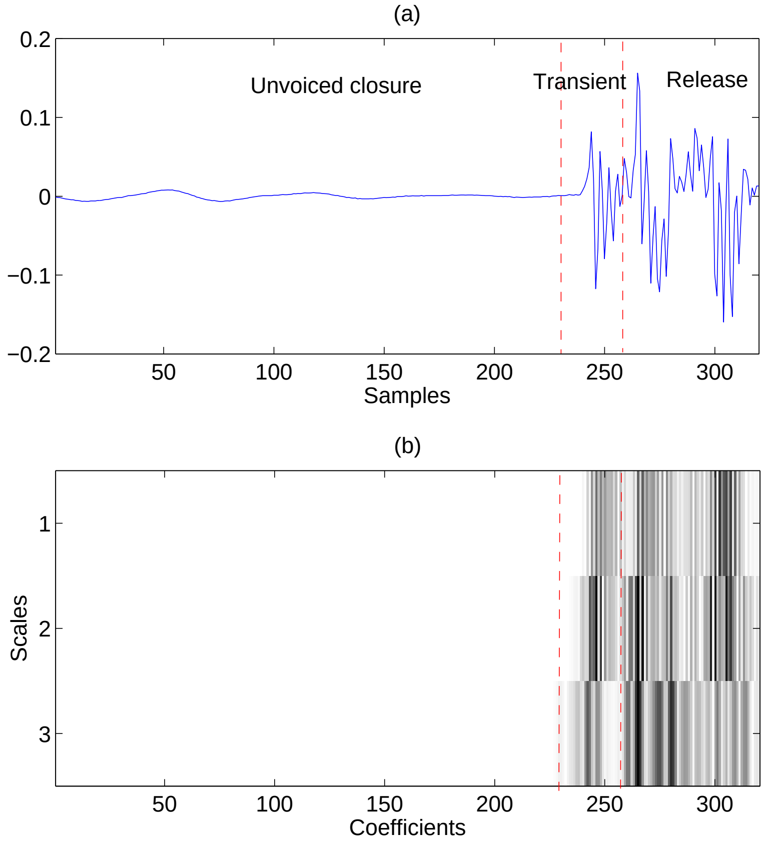

Figure 2.7: (a) Closure interval frames followed by a transient and release frames, (b) Detail coefficients decomposed at 3 scales.

features. Fig. 2.9 shows CTDR peaks detected at different decomposed scales which correspond to the transient frames of stops. To detect a plosive group, we need to combine the classificatie results of voiced/unvoiced closure frames, a transient frame, and unvoiced frames by an interpolator which will be explained in the next section.

2.3 Joint neural network classifier

A novel and efficient phonetic classifier is developed by jointly combining the FNNs and a multi-threshold classifier. As reported in [LCG93] and other studies [MBB96, QH93, GCEJ91], application of neural networks helps to improve the performance of the classifier. Two FNNs operate on input features to classify the input frames as voiced, unvoiced, mixed excitation and transient frames, respectively. Silence frames and voiced closure interval frames are detected by employing one part of the MTD classifier proposed in [PK04]. While the optimal thresholds which are found automatically

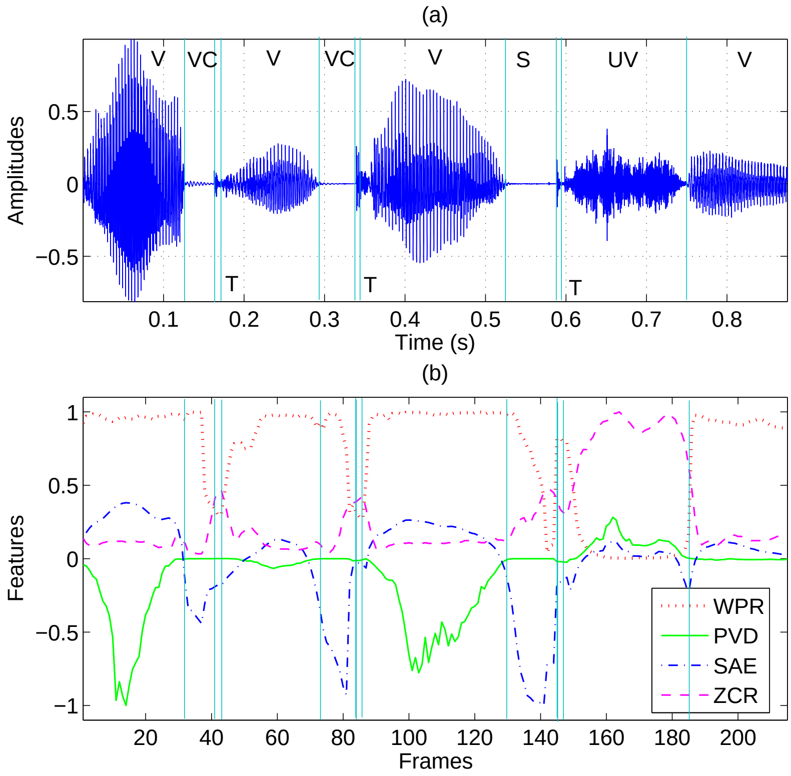

Figure 2.8: (a) A speech utterance "h**ad your dark su**it" and its phonetic classes V/VC/S/UV/T, (b) Evolution of the extracted features and segmentation of phonetic regions.

by training FNNs, can improve the performance of the joint classifier, its complexity is lowered due to the simplicity of the linear thresholds derived from the MTD classifier.

2.3.1 Network initialization

A supervised learning algorithm is used to train the FNNs (depicted in Fig. 2.10). So, the network weights and biases are adjusted to minimize the mean-square error between the network outputs and real target outputs. To detect transient frames, a first two-layer network with 5-dimensional input vector and 1-dimensional output is trained. A second network is configured with 4-dimensional input vector and 3-dimensional output to perform the three-way classification: voiced, mixed excitation, and unvoiced frames. The targets are labeled with 1 for the desired class and 0 for the other classes. As a preprocessing step, all elements of the input and output vectors are normalized to get zero mean and unity standard deviation over the training set. To compute error

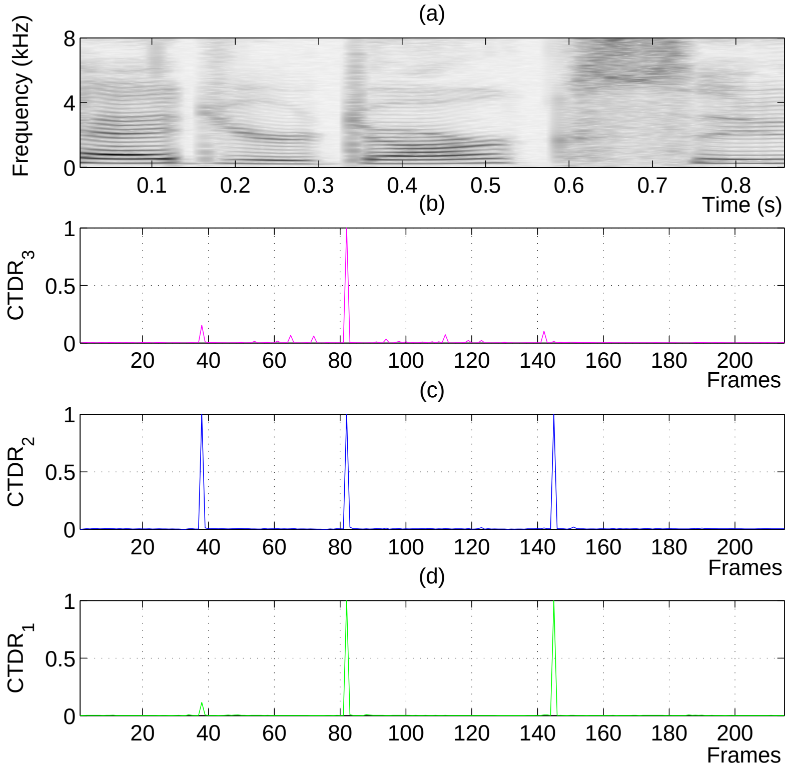

Figure 2.9: (a) Spectrogram of the same speech utterance, (b) The closure interval-transient detail ratio $CTDR_3$, (c) $CTDR_2$, (d) $CTDR_1$.

percentages, the outputs will be postprocessed to convert them back into the same units that were used for the original target data. The feedforward networks use log-sigmoid transfer functions for all hidden units in their hidden layer and linear transfer functions at the output layer. Biases and weights of each unit are initialized to very small random values.

2.3.2 Network learning

A weight decay heuristic or regularization is used to avoid overfitting in the backpropagation learning algorithm. This approach results in smaller weights and biases and forces the network response to be smoother over its complex decision surface [Mit97]. In detail, the weights are decreased by some small factor during each iteration. This means that the standard cost function is modified by adding a penalty term corre-

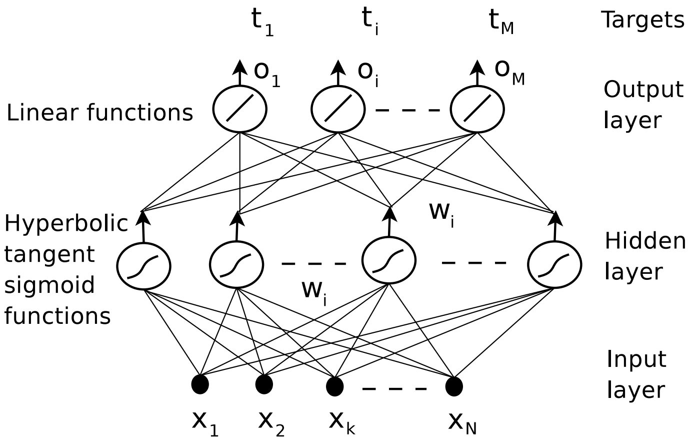

Figure 2.10: Feedforward neural network configuration with N inputs x_k, M outputs o_i, and M targets t_i, w_j are the weighting factors between units of different layers.

sponding to the sum of squares of the network weights:

$$Ereg = \gamma \frac{1}{N} \sum_{i=1}^{N} (t_i - o_i)^2 + (1-\gamma) \frac{1}{M} \sum_{j=1}^{M} (w_j)^2 \qquad (2.6)$$

where γ is a performance ratio, w_j are all weights of the NN, and t_i and o_i are the output and target values, respectively. Some common parameters of the learning algorithms are set up to initialize the neural network configurations as follows:

- The learning rate $lr = 0.05$.
- The number of iterations $epochs = 1000$.
- The training performance $goal = 1e - 5$

There are several learning algorithms for learning the neural network such as: momentum, variable learning rate, Levenberg-Marquardt and BFGS Quasi-Newton algorithms. Their characteristics [DB02] are explained briefly as follows:

- Adding momentum: the weight update on the current iteration depends partially on the update that occurred during the last iteration. This helps the gradient descent-based algorithm pass some local minima or flat region in the error surface.
- Variable learning rate: the performance of the steepest descent algorithm can be improved if the learning rate is changed during the training process. An adaptive learning rate algorithm tries to keep the learning step size as large as possible while maintaining stability.

- Levenberg-Marquardt and BFGS Quasi-Newton algorithms are improvements over the steepest descent algorithms and generally achieve faster convergence.

2.3.3 Selection of parameters and structure

It is difficult to determine the optimum value for the performance ratio γ and the number of hidden units nH which result in overfitting or underfitting. The best values can be chosen experimentally by a procedure proposed in [PK05a]. By varying the performance ratio γ and number of hidden units nH over the ranges $[0.3, 0.35, ..., 0.8]$ and $[5, 10, ..., 140]$, respectively, several experiments were implemented to select the best parameters and structure. The optimal values will be the ones for which the sum of training and testing error rates is smallest. The datasets containing the extracted time-scale features used for learning the network structure are constructed from the dialect speaking region 1 of the TIMIT database [GLF+93]. There are 110 utterances uttered by four female speaker and seven male speaker are collected separately to investigate the gender-dependency of the phonetic classifiers. Another dataset containing both genders is used to design a gender-independent phonetic classifier. Each dataset is divided into 70% for training and 30% for testing.

Before finding the optimal parameters and structure of the FNNs, the investigation of selecting the best learning algorithm is done. By fixing the performance ratio $\gamma = 0.5$ and varying number of hidden units nH over the ranges $[5, 10, ..., 140]$, four listed learning algorithms are applied to train the classifiers. For each dataset, the results of the training phase shows that the Levenberg - Marquardt (LM) algorithm generally gives the highest network performance. Fig. 2.11 shows the changes of classification performance obtained from every learning algorithm over the defined range for the number of hidden units nH for the training dataset spoken by female speakers. For this learning algorithm, the optimal choices of the performance ratio γ and the number of hidden units nH are found out by trying many different combinations of γ and nH from the defined ranges. Figure 2.12 shows the performance error rates derived from two different combinations of γ and nH for the training and testing female datasets. The best combinations which provide the lowest error percentage on the training and test sets of the one-class FNN classifier and of the three-classes FNN classifier are presented in Table 2.1. These configuration will be used for evaluation of the designed classifiers in section 2.5.

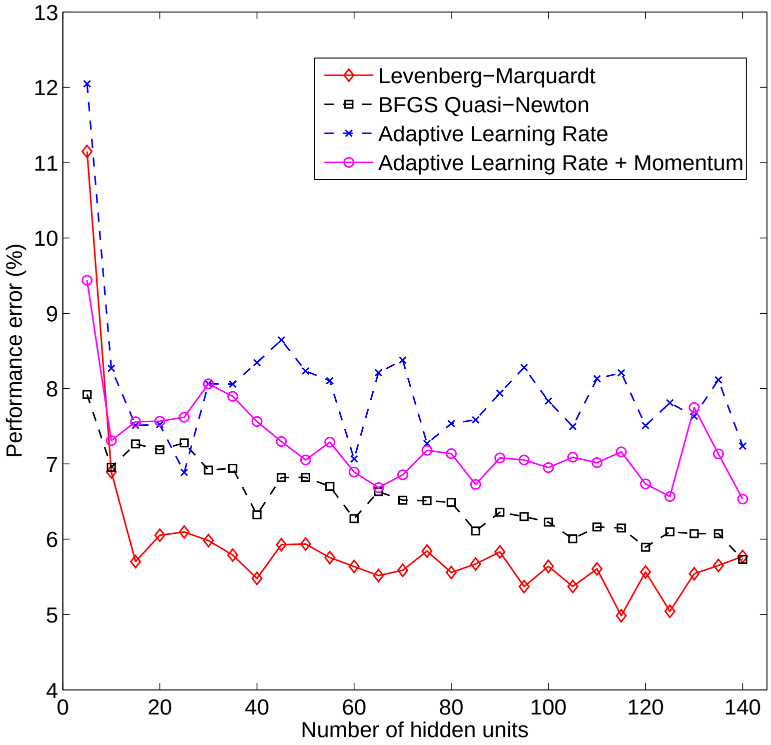

Figure 2.11: Performance of the learning algorithms at $\gamma = 0.5$.

Table 2.1: Optimal choices for the performance ratio and the number of hidden units.

Gender	Female		Male		Mixed	
Parameters	γ	nH	γ	nH	γ	nH
Neural network						
One-class FNN	0.75	60	0.55	55	0.55	70
Three-classes NN	0.40	100	0.50	125	0.65	110

2.3.4 Joint classifier

The MTD classifier proposed in [PK04] considers all seven input features to classify every speech frame into different phonetic classes. It works by comparing determined thresholds with the extracted features. Based on this comparison, the classification decision is carried out. The hard thresholds which are employed by the model are chosen via experimental pattern classification which is described in more detail in appendix B. Because the fixed thresholds had been optimized experimentally for a specific data set, they are not robust when applied for other databases recorded in different conditions.

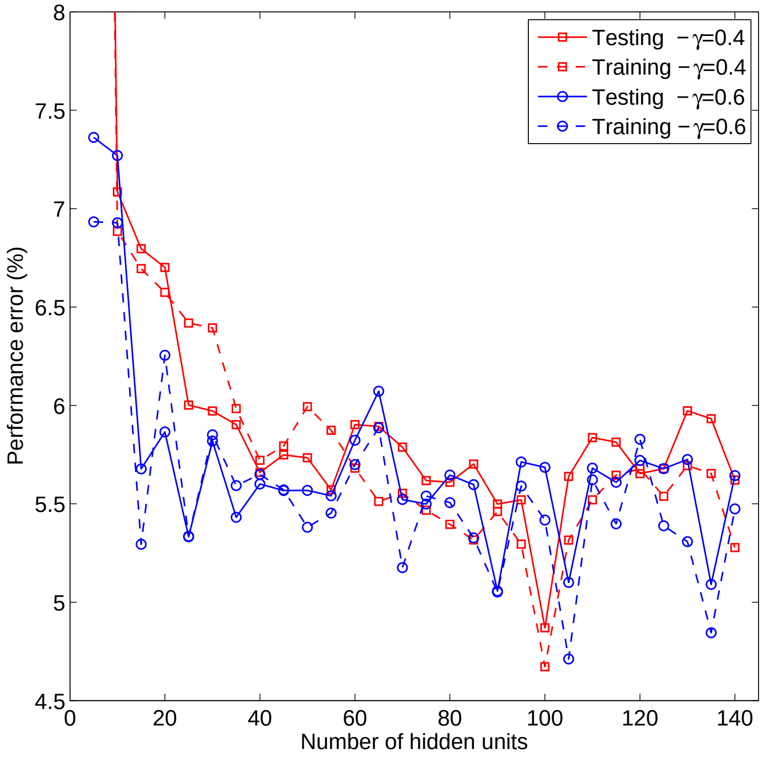

Figure 2.12: The best performance ratios γ for LM algorithm.

That is why the FNN is employed as well. However, the characteristics of the silence and voiced closure interval classes are quite simple and can be detected accurately by applying the multi-threshold decision classifier. This property certainly allows to lower the required number of inputs and outputs of the FNNs and, thereby to increase the classification performance of the network classifiers. Thus, to exploit the simplicity of the MTD classifier and the optimality of FNNs, a joint phonetic classifier is proposed with the following scheme:

The algorithm for classification into different phonetic groups is implemented in four sequential steps as follows:

- First, silence and voiced closure interval frames are detected by a linear classifier which is based on the comparison between extracted feature values and predetermined hard thresholds. The linear classifier is one of the modules of the MTD classifier. A procedure for designing the MTD and a flow chart of the linear classifier are described in appendix B.

- Second, transient (T) frames are detected by considering frames which immedi-

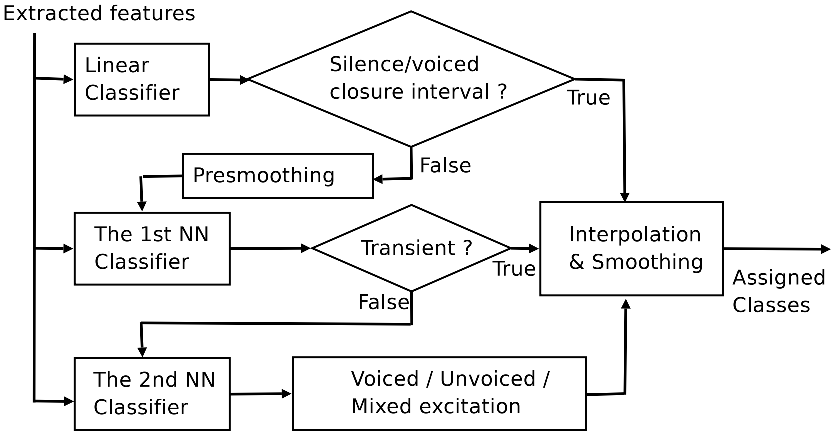

Figure 2.13: Block scheme of combined classifier.

ately follow silence or voiced closure interval frames (VC). Five parameters (i.e. WPR, ZCR, $CTDR_1$, $CTDR_2$, and $CTDR_3$) are computed for these frames and classified by the first NN. The network distinguishes between transient frames and other frames. Then, all silence and voiced closure interval frames occurring before transient frames are marked as closure interval frames.

- Third, four parameters (i.e. WPR, SAE, ZCR and PVD) of every not yet classified frame are calculated to build the input vectors for the second NN. The three classes voiced, unvoiced and mixed-excitation are recognized based on the output values of the second neural network.

- Finally, an interpolation method relying on phonemic features is implemented to determine the temporal boundary of plosives which are formed by closure interval + transient + release frames (unvoiced frames following a transient frame). The operation of the interpolation method relies on phonemic features and the sequential structure of speech sounds in American English. Then, some remaining incorrect decisions are repaired by a smoothing method based on the sequential consistency of speech sounds. For example: VVSVV → VVVVV, VVUVV → VVVVV.

2.4 Bayesian network classifier

[2] In this section, the application of Bayesian networks for the phonetic classification task is studied. A Bayesian network [Pea88] $\mathcal{B} = \langle \mathcal{G}, \Theta \rangle$ is a directed acyclic

[2] Thanks to F. Pernkopf for designing the Bayesian classifiers and assisting the text material published in [PP06a].

graph $\mathcal{G} = (\mathbf{Z}, \mathbf{E})$ consisting of a set of nodes $\mathbf{Z}$ and a set of directed edges $\mathbf{E} = \left\{E_{Z_i,Z_j}, E_{Z_i,Z_k}, \ldots\right\}$ connecting the nodes where E_{Z_i,Z_j} is an edge from Z_i to Z_j. This graph represents factorization properties of the distribution of a set of random variables $\mathbf{Z} = \{C, X_1, \ldots, X_N\} = \{Z_1, \ldots, Z_{N+1}\}$, where each variable in $\mathbf{Z}$ has values denoted by lower case letters $\{c, x_1, \ldots, x_N\}$. We use boldface capital letters, e.g. $\mathbf{Z}$, to denote a set of random variables and correspondingly boldface lower-case letters denote a set of instantiations (values). The random variable $C \in \{1, \ldots, |C|\}$ represents the classes, $|C|$ is the cardinality of C, $\mathbf{X}_{1:N} = \{X_1, \ldots, X_N\}$ denote the set of random variables of the N attributes of the classifier. Each graph node represents a random variable, while the lack of edges specifies independences. Specifically, in a Bayesian network each node is independent of its non-descendants given its parents. These conditional independence relationships reduce both the number of parameters and the required computational effort. The symbol Θ represents the set of parameters which quantify the network. Each node Z_j is represented as a local conditional probability distribution given its parents Z_{Π_j}. The joint probability distribution of the network is determined by the local conditional probability distributions as $P_\Theta(\mathbf{Z}) = \prod_{j=1}^{N+1} P_\Theta\left(Z_j | Z_{\Pi_j}\right)$.

2.4.1 Classifier structures

The naive Bayes (NB) and tree augmented naive (TAN) structures are used in our research. For TAN structures we have to learn a 1-tree. The NB network assumes that all the attributes are conditionally independent given the class label [FGG97]. The structure of the naive Bayes classifier is illustrated in Figure 2.14a. In order to correct some of the limitations of the NB classifier, Friedman et al. [FGG97] introduced the TAN classifier. A TAN is based on structural augmentations of the NB network, where additional edges are added between attributes in order to relax some of the most flagrant conditional independence properties of NB.

2.4.2 Network learning

We apply both generative and discriminative parameter learning. The generative parameter learning which is based on ML estimation [Pea88] to maximize the likelihood function:

$$LL(\mathcal{B}|\mathcal{S}) = \sum_{m=1}^{M} \log P_\Theta\left(Z = z^m\right) \tag{2.7}$$

with a fixed structure of $\mathcal{B}$, $\mathcal{S}$ is the data set containing M samples. Discriminative parameter learning presented in [GZ02] tries to optimize the conditional likelihood

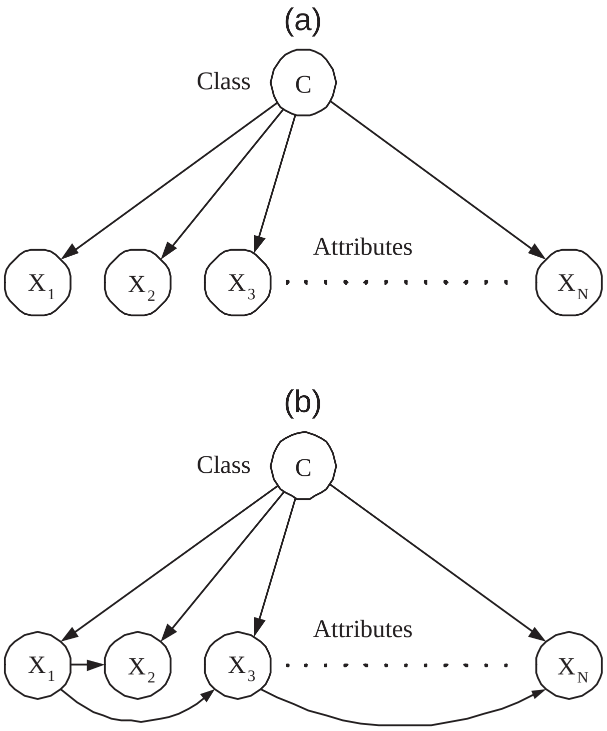

Figure 2.14: Bayesian Network: (a) NB, (b) TAN [PP06a].

(CL):

$$CLL\left(B|\mathcal{S}\right) = \sum_{m=1}^{M} \log P_{\Theta}\left(C = c^m | X_{1:N} = x_{1:N}^m\right) \tag{2.8}$$

For learning structures of the Bayes networks, in the first approach, we employ generative structure learning where the conditional mutual information (CMI) $I\left(X_i; X_j|C\right)$ between the attributes given the class variable is used as score. Friedman et al. [FGG97] give an algorithm for constructing a TAN network using this measure. In the second approach, we use an order-based greedy search heuristic, in fact order mutual information (OMI) heuristic, for efficient learning of the discriminative structure of a Bayesian network classifier which is proposed by Pernkopf at al. [PB06]. It consists of two steps:

- First establish an ordering of nodes $\mathbf{X}_{\prec}^{1:N} = \left\{X_{\prec}^{1}, X_{\prec}^{2}, \ldots, X_{\prec}^{N}\right\}$

- Second select the parent of $X_{\prec}^{j} \in \mathbf{X}_{\prec}^{1:j-1}$ by maximizing the classification rate.

2.5 Evaluation and discussions

2.5.1 Data setup

The speech data used to build the experimental datasets are general data including transition frames between different classes extracted from dialect speaking region 4 (DR4) of the TIMIT database [GLF+93] which has a balance distribution between the number of male and female speakers. All of the speech sounds were digitized at a sampling rate of 20 kHz with an anti-aliasing filter at 10 kHz, then downsampled to F_s=16kHz. The data have been split into 2 mutually exclusive subsets of $\mathcal{D} \in \{\mathcal{S}_1, \mathcal{S}_2\}$ where the size of the training data $\mathcal{S}_1$ is 70% and of the test data $\mathcal{S}_2$ is 30% of $\mathcal{D}$. Throughout the experiments, we use exactly the same data partitioning. The dialect region 4 consists of 16 male and 16 female speakers, 320 utterances, and 121629 frames in total. We perform classification experiments on data of male speakers (Ma), female speakers (Fe), and both (Ma+Fe) genders. Besides the proposed features, baseline features, i.e. 12 MFCCs + Log-energy are used for comparison. The values of attributes in the data sets are continuous.

The distributions of six phonetic classes V/UV/S/M/VC/T which are extracted from the transcriptions of the DR4, TIMIT database are 20.9%, 54.66%, 12.26%, 2.74%, 6.08%, 3.36%, respectively. The performance of the proposed classifiers is assessed by several measurements. We see that silence, voiced, unvoiced and mixed-excitation classes are evaluated directly at the outputs of the linear or non-linear threshold classifiers. As discussed at the end of the section 2.2, plosive class consists of voiced closure interval or unvoiced closure interval frames with transient frame and release frames. So, its performance can be assessed indirectly via the classification rate of the VC class and the T class, or directly at the output of the interpolator in section 2.3.4 which is based on phoneme structure.

2.5.2 Performance of joint classifier

The configurations of the parameters and structure of the neural networks which are reported in section 2.3 are re-used to train and test the joint classifiers with the current datasets of dialect speaking region 4. Because these configurations were experimentally optimized for specific dataset, some other configurations are also tried in this experiment. The combined linear classifier is the same as designed in [PK05a].

In order to compare the performance of time-scale features with MFCC features,

Table 2.2: Joint neural networks classifier using time-scale features.

Experiments		No. 1		No. 2		No. 3	
Parameters		γ	nH	γ	nH	γ	nH
One-class NN		0.55	70	0.55	55	0.75	60
Three-class NN		0.65	110	0.50	125	0.40	100
Data set	Features						
Ma+Fe	TSF	**83.02** ± **0.20**		82.93 ± 0.20		81.87 ± 0.20	
Ma	TSF	83.76 ± 0.27		**83.67** ± **0.27**		82.23 ± 0.28	
Fe	TSF	80.97 ± 0.29		79.84 ± 0.30		**81.04** ± **0.29**	

Table 2.3: Neural network classifier using MFCC features.

Experiments		No. 1		No. 2		No. 3	
Parameters		γ	nH	γ	nH	γ	nH
		0.50	100	0.5	200	0.5	400
Data set	Features						
Ma+Fe	MFCC	81.40 ± 0.20		**82.78** ± **0.20**		80.24 ± 0.21	
Ma	MFCC	**84.41** ± **0.27**		83.19 ± 0.28		83.27 ± 0.28	
Fe	MFCC	79.26 ± 0.30		**82.85** ± **0.28**		80.13 ± 0.30	

a two-layer FNN classifier which has 13 inputs and 6 outputs is trained from MFCC features. We repeat the procedure in section 2.3 to find out several good configurations for this FNN classifier. From our observation, it seems that the best performance ratio is around value of 0.5. Thus, we fix the performance ratio at $\gamma = 0.5$ and only vary the number of hidden units. Table 2.2 and table 2.3 present the average classification rates [%] derived from each three different configurations of parameters and structures of the neural networks using time-scale features (TSF) and MFCC features. Because the distributions of the six phonetic classes are not the same, the average classification rates is calculated as a ratio between the sum over all numbers of true acceptance frames and total number of tested frames. The claculation of standard deviation is based on evaluating hypothese that error rate is a random variable which obeys the Binomial distribution [Mit97]. The detailed explanation and formular are presented in appendix C.

With the observed results, we see that the usage of TSF provides slightly lower classification rates than the usage of the MFCC features for Ma and Fe databases, while higher classification rates are obtained in case of Ma+Fe database. However, for TSF we used only 7 features compared to 13 MFCC features, that may lead to the simpler classifier. In comparison with the MTD classifier, the joint neural network improves significantly the average classification rates for three data sets Ma+Fe, Ma and Fe given as 83.02, 83.67 and 81.04 [%] compared to 54.32, 57.24 and 48.17 [%] of the MTD. The classification performance of the MTD classifier is rather low because its hard thresholds were manually optimized for another smaller data set. With the optimization constrained to a specific limited data set and the usage of re-labeled transcriptions at the acoustic level, high classification rates were originally achieved in [PK04], but they are now known to not generate well. This observation also implies that the MTD classifier suffers from the overfitting problem which is a consequence from its careful manual training.

Table 2.4: Confusion matrix in [%] for all six classes derived from joint FNN using time-scale features, for mixed data set.

Assignments	V	UV	S	M	VC	T
Transcriptions						
V	**84.63±0.58**	4.07±0.32	3.64±0.30	0.69±0.13	2.16±0.24	4.81±0.35
UV	0.95±0.10	**94.33±0.23**	0.87±0.09	1.16±0.11	1.78±0.13	0.91±0.10
S	5.83±0.50	4.16±0.42	**80.02±0.85**	0.58±0.16	9.14±0.61	0.27±0.11
M	16.58±1.66	11.04±1.40	13.89±1.55	**42.07±2.21**	13.62±1.54	2.80±0.74
VC	11.02±0.94	26.82±1.33	25.24±1.30	1.02±0.30	**34.77±1.43**	1.13±0.32
T	29.52±1.84	21.47±1.66	2.29±0.60	1.03±0.41	2.98±0.69	**42.72±2.00**

Table 2.5: Confusion matrix in [%] for all six classes derived from FNN using MFCC, for mixed data set.

Assignments	V	UV	S	M	VC	T
Transcriptions						
V	**83.22±0.61**	4.34±0.33	3.87±0.31	1.33±0.19	2.32±0.24	4.92±0.35
UV	0.87±0.09	**93.86±0.24**	1.24±0.11	1.20±0.11	1.28±0.11	1.55±0.12
S	5.65±0.49	4.41±0.43	**79.12±0.86**	0.66±0.17	9.81±0.63	0.35±0.12
M	16.98±1.68	11.42±1.42	12.84±1.50	**40.95±2.20**	13.88±1.55	3.93±0.87
VC	11.71±0.97	25.03±1.30	24.42±1.29	1.23±0.33	**36.39±1.44**	1.22±0.33
T	27.17±1.80	19.02±1.59	2.39±0.62	0.88±0.38	2.24±0.60	**43.30±2.00**

The confusion matrices show the classification performance of each phonetic class. The results in percentage obtained from evaluating the joint classifier using time-scale features and the neural network using MFCC features on the mixed (male and female) data set are reported in Table 2.4 and Table 2.5. The classification rates of the voiced, unvoiced and silence classes are higher than the rates of the remaining classes. As expected, the application of wavelet power distributions according to section 2.2.2 results in these interesting rates. This results from the flexibility of having short basis functions to analyze high-frequency speech components while long ones are applied on low-frequency speech components of the DWT. The mixed-excitation class has lower classification rates because it consists of acoustic properties of both voiced and unvoiced sounds. This makes it more dfficult to detect the class. As reported in Table 2.4 and Table 2.5, the classification rates of the transient class derived from the TSF is a little lower than the ones obtained by the MFCC features. By analyzing the classification outputs, we see that sometimes double transient frames are detected from a single plosive. This may come from the use of too long wavelet basis (dB20) with high number of vanishing moments to obtain good separation between voiced and unvoiced sounds.

Based on the classification performance of the voiced closure interval class and the transient class, the plosive detection rate can be calculated by applying the interpolation rule in section 2.3.4. With the designed joint neural network classifier, the plosive detection rates are 37.72, 41.86 and 40.17 [%] for the Ma+Fe, Ma and Fe data sets, respectively. By analyzing the errors, we observed that many misdetected plosives do not have a perfect acoustic properties as they would be expected while the measure of detection rates is based on the transcriptions performed manually by phoneticians. The same observation was reported by [NS02] in detail. As analyzed, the deviations may come from individual speaker characteristic and unsuitable phonetic labels. Speakers with high mis-classification rates have stop consonants that are always weakly visible in the speech signal.

From the confusion matrices, the recall (Re) and precision (Pr) measures which represent the classification rate and accuracy of the classifiers can be calculated as:

$$Re = \frac{TA}{TA + FR}, \quad Pr = \frac{TA}{TA + FA}, \tag{2.9}$$

where TA, FR and FA is the number of true acceptance, false rejection and false acceptance frames, respectively, derived from the confusion matrices (in number of frames rather than %). The measures are illustrated in table 2.6. In most cases, the precisions of every class are quite balance with the recalls with the average difference is about 3%. The unvoiced class has the highest recall and precision scores obtained by all classifiers. It is interesting that the accuracy of the mixed-excitation, voiced closure, and transient classes are better than their detection rates. This means there are lower probabilities of introducing inserted frames into these three classes.

Table 2.6: Recall and precision measured on mixed data set for all six classes derived from joint FNN using time-scale features and MFCC features.

Classes		V	UV	S	M	VC	T
Measures	Features						
Re	TSF	84.63	**94.33**	80.02	42.07	34.77	42.72
Pr	TSF	84.95	**92.86**	77.42	72.18	42.53	47.24
Re	MFCC	83.22	**93.86**	79.12	40.95	36.39	43.30
Pr	MFCC	85.12	**91.11**	76.34	58.38	44.79	36.22

2.5.3 Performance of the Bayesian network

We apply both discriminative parameter learning by optimizing the CL [GZ02] and generative parameter training (ML estimation) [Pea88] on both discriminatively (TAN-OMI) and generatively (TAN-CMI) structured Bayesian networks. We use the NB and the TAN classifier topology. Since the classifiers are constructed for multinomial attributes, the features have been discretized using the algorithm in [FI93] where the codebook is produced using only the training data. Zero probabilities in the conditional probability tables of the Bayesian networks are replaced with a small epsilon $\varepsilon = 0.00001$. Discriminative parameter learning is currently implemented in a naive way. We either perform 15 iterations of the gradient descent algorithm or prematurely terminate the parameter optimization in case of convergence.

Table 2.7: Bayesian network using time-scale features.

Classifier	NB		TAN		TAN	
Struct. Learn.	-		CMI		OMI	
Param. Learn.	ML	CL	ML	CL	ML	CL
Data set						
Ma+Fe	81.54 ± 0.20	81.63 ± 0.20	83.12 ± 0.20	83.15 ± 0.20	83.47 ± 0.19	**83.49 ± 0.19**
Ma	82.50 ± 0.28	82.60 ± 0.28	84.05 ± 0.27	84.06 ± 0.27	84.72 ± 0.27	**84.73 ± 0.27**
Fe	81.14 ± 0.29	81.20 ± 0.29	82.52 ± 0.28	82.58 ± 0.28	**82.70 ± 0.28**	82.69 ± 0.28

Table 2.8: Bayesian network using MFCC features.

Classifier	NB		TAN		TAN	
Struct. Learn.	-		CMI		OMI	
Param. Learn.	ML	CL	ML	CL	ML	CL
Data set						
Ma+Fe	82.08 ± 0.20	82.16 ± 0.20	82.89 ± 0.20	82.91 ± 0.20	83.39 ± 0.19	**83.40 ± 0.19**
Ma	82.35 ± 0.28	82.48 ± 0.28	83.87 ± 0.27	83.88 ± 0.27	84.28 ± 0.27	**84.31 ± 0.27**
Fe	81.99 ± 0.28	82.05 ± 0.28	82.63 ± 0.28	82.63 ± 0.28	83.03 ± 0.28	**83.06 ± 0.28**

Table 2.7 and table 2.8 present the classification rates [%] obtained from all different generative/discriminative classifiers for 6 phonetic classes using TSF features and MFCC features. As reported, the TAN classifier using discriminative structure and parameter learning (TAN-OMI-CL) outperforms the generative approaches for both feature sets by a small margin. However, the evaluation of the classification rate in the OMI algorithm is computationally expensive. Discriminative parameter learning (CL) produces mostly a better classification performance than generative parameter learning (ML).

In general, the obtained classification rates are higher than the results derived from the FNN classifiers. The improvement may come from the abilities of learning structure of BN by the generative structure learning [FGG97] and discriminative structure learning [PB06] algorithms. These automatic learning methods yields a better structure than the experimental structure selection method described in section 2.3.2. However, the use of neural networks requires much lower computational effort while providing approximately similar results. Just as in the case of neural networks, the proposed time-scale features outperform the baseline MFCC features in most classifiers (for mixed and male data sets). The confusion matrices in percentage obtained from evaluating the BN (TAN-OMI-CL) using time-scale features and MFCC features on the male data set are reported in table 2.9 and table 2.10, respectively.

Again, we see that the classification rates of the transient class derived from the TSF is lower than the ones obtained by the MFCC features. For all classifiers tested above, both the neural networks and the Bayesian networks, the classification performance achieved by the gender-independent classifiers are somehow lower than the one derived from the gender-dependent classifiers. The reason is the gender-independent classifiers have to learn more complex class borders from the bigger mixed dataset. The average difference of classification performance due to the gender factor approximates 1.50 % for

Table 2.9: Confusion matrix in [%] for all six classes derived from BN (TAN-OMI-CL) using time-scale features, male speaker data set.

Assignments	V	UV	S	M	VC	T
Transcriptions						
V	**85.27±0.57**	3.93±0.31	3.51±0.30	0.66±0.13	2.06±0.23	4.57±0.34
UV	0.84±0.09	**96.31±0.19**	0.56±0.08	0.17±0.04	1.31±0.11	0.81±0.09
S	5.75±0.49	3.80±0.40	**81.10±0.83**	0.24±0.10	8.93±0.60	0.19±0.09
M	16.25±1.65	10.91±1.40	13.45±1.53	**43.65±2.22**	13.46±1.53	2.28±0.67
VC	10.53±0.92	26.41±1.32	24.91±1.30	0.56±0.22	**36.84±1.45**	0.75±0.26
T	29.48±1.84	20.92±1.64	1.43±0.48	0.48±0.28	2.69±0.65	**45.01±2.01**

Table 2.10: Confusion matrix in [%] for all six classes derived from BN (TAN-OMI-CL) using MFCC, male speaker data set.

Assignments	V	UV	S	M	VC	T
Transcriptions						
V	**84.73±0.58**	4.59±0.34	3.68±0.30	0.69±0.13	2.99±0.28	3.31±0.29
UV	1.35±0.12	**95.82±0.20**	0.70±0.08	0.47±0.07	0.77±0.09	0.89±0.09
S	7.36±0.55	3.99±0.41	**78.25±0.87**	0.95±0.21	9.16±0.61	0.28±0.11
M	15.74±1.63	13.20±1.52	12.69±1.49	**40.10±2.19**	11.92±1.45	6.35±1.09
VC	10.90±0.94	26.79±1.33	22.56±1.26	1.79±0.40	**36.94±1.45**	1.03±0.30
T	21.55±1.66	21.08±1.65	1.90±0.55	0.48±0.28	4.28±0.82	**50.71±2.02**

most considered classifiers. The average difference of classification rates between male speakers and female speakers is also quite low. This achievement opens an approach for gender independent phonetic classification.

2.6 Conclusions

Two novel methods for phonetic classification have been developed. The discrete wavelet transform is exploited to extract useful features for classifying a sequence of input speech frames into six phonetic classes given as the voiced, unvoiced, silence, mixed-excitation, voiced closure, and transient classes. The classification decision is done by training feedforward neural networks and Bayesian network. The linear classifier using a multi-threshold decision scheme is combined with two trained FNNs in order to build an efficient joint classifier. An interpolation rule and phonetic smoothing helps to detect the plosive class. This joint classifier has optimal non-linear thresholds learned by FNN and low complexity due to the usage of the linear multi-threshold classifier. Furthermore, discriminative structure learning of Bayesian networks is superior to the generative approach. Discriminative parameter training improves the classification rate in most cases and gives slightly better classification performance than the one derived by the joint FNN classifiers. Both the joint FNN classifier and BN classifier provide quite good classification rate with high accuracy. The proposed time-scale features provide similar classification performance compared to the baseline MFCC features in most cases. The performance of gender-dependent/independent based classifiers suggest the viability of gender independent phonetic classification.

From our analysis, we see that the time-scale features can be improved further in order to improve the performance of the mixed-excitation, voiced closure, and transient classes. In addition, an adaptive selection of wavelet bases for different phonetic classes should be investigated. The wavelet bases with very short support would be suitable to capture transient frames exactly while the long ones are used to detect voiced/unvoiced frames. The usage of hidden Markov models with Viterbi decoding is a good candidate for replacing the phonetic interpolation rule in plosive detection. Besides, the optimal structure design of structure for neural networks should be investigated. Comparison between FNN-based classifiers and GMM-based classifiers in terms of classification rate is an interesting issue. From the reported results, we observe that the classification rates are just increased slightly from the use of the neural networks to Bayesian networks owning optimal trained structures. It might be wise to investigate on how to improve the feature extraction to gain significant performance than to train the more smart but complicated classifiers with lower gain. Gender mismatch between training and testing phase is an interesting open issue that should be tested. Finally, evaluation on other databases with more accurate transcriptions will be investigated. As an evaluation in the application domain, the outputs of the phonetic classifiers will be used in time-scale modification algorithms as extra input information to adjust a non-uniform scaling factor for each phonetic class in order to get better perceptual quality. Besides, the application of the proposed classifiers to phonetic pre-alignment with addressed phonetic classes for large databases is very useful.

Chapter 3

Robust speech classifiers for adverse environments

3.1 Introduction

[3] Nowadays, the demand of using speech technologies for human communication in real-world environments increases swiftly. Plentiful approaches of speech processing have been developed to achieve robustness of speech applications in real environments which exhibit different kinds of noise. Voice activity detection (VAD) which is employed by most speech applications is a common task of speech classification. VAD is used to estimate the noise level which is the critical point for many noise reduction methods [Coh03]. In automatic speech recognition, there is a need of VAD for a speech/non-speech detection to improve the recognition rate [ETS03, RSB$^+$05]. Application of VAD in speaker verification (SV) systems was introduced in [MCGB01] to achieve reliable model estimation. Study on influence of VAD to performance of SV systems under adverse conditions is an interesting topic. The classification of voiced/unvoiced/silence (V/U/S) as well as voice activity detection are not simple when they are applied in realistic environments.

Many studies on speech and phonetic classification have been proposed since in the last decade. Very common features are used for the task such as zero-crossing rate, autocorrelation coefficients [CHL89, TO00], periodicity estimation [Tuc92], short-term energy [ETS03]. These features may be affected by complex and strong noise. Recently, some novel methods have been proposed to deal with high noise conditions and differ-

[3] This chapter is based on materials published earlier in [NPHKon, PK06b, NPK07], and unpublished in [PNK07].

ent noise types which may appear in real-world environments. The features extracted from the frequency domain are the LPC distance feature [RS97], cepstrum [HM93], mel frequency cepstrum coefficients [XH02], instantaneous frequency amplitude spectrum [AK05] and pitch estimation [Che99]. Several standard algorithms were proposed by ITU-T and ETSI for different applications. ITU-T standard Rec. G.729 Annex B [BSS+97] is developed for fixed telephony and multimedia communications. It uses input speech frame of 10ms to extract four features, i.e. differential power in the $0-1$ kHz band, differential power over the whole band, differential zero crossing rate, and spectral distortion. ETSI [ets99] has two standards for the adaptive multirate (AMR) codec. They are developed for the third-generation mobile communication systems. The AMR standard 1 estimates the signal-to-noise ratios (SNRs) in nine subbands and compares them with adaptive thresholds which are different for each subband. The AMR standard 2 divides frames of 20ms into two subframes of 10 ms, and extracts three features for each of them, i.e. channel power, voice metrics, and noise power. The detection is done by comparing the voice metrics with a SNR-based threshold. If at least one subframe is detected as speech then a frame is labeled as speech frame. ETSI ES 202 050 [ETS03] has proposed VAD for the advanced front-end in distributed speech recognition. The VAD algorithm employs the short-term log-energy in comparison with an adaptive SNR-based threshold. A hangover scheme is used after that to smooth the output decisions. An adaptive forgetting factor is used to update the noise estimation.

The features extracted in the frequency domain by applying the short-time Fourier transform face some limitations due to a fixed absolute resolution of time-frequency (TF) analysis. The application of the discrete wavelet transform which provides a rectangular tiling of the TF plane with fixed relative resolution to design robust VAD has been investigated recently in [SS97b, CW02]. To be robust against noise, several VAD algorithms based on wavelet features have been designed by employing statistical models [LA06], or the auto-correlation function and the Teager energy operator (TEO) [WW06]. Other methods using subband order-statistics filters [RSB+04] and higher-order statistics in the LPC residual domain [NGM01] can improve VAD performance in non-stationary noise conditions. Nevertheless, more research on this and other TF analysis techniques for VAD is worth undertaking.

In this chapter, an efficient and robust speech classifier based on a single wavelet feature and quantile filtering is presented. The classifier is designed to classify every input speech frame into V/U/S classes and thereby detect speech frames out of non-speech frames. The DWT is applied on every speech frame extracted by overlapping windows. Then a frame-based delta feature is extracted from the calculated TEO of the wavelet coefficients. To be robust against noise, the extracted feature is enhanced by applying a sigmoid function and median filtering. After that, the enhanced feature is compared with a quantile-based threshold to classify each frame into V/U/S classes. Furthermore, by applying a slope tracking method on the processed feature instead of using the adaptive threshold, the V/U/S decisions can be obtained with lower delay and memory requirements. The VAD decision is done from this phonetic classification output. To obtain a better estimate of the noise threshold, the quantile filtering method of [PK06b] is further improved by an adaptive estimation of the quantile factor. Finally, by applying a hang-over scheme [PKW+06] over a buffer storing frame-based phonetic decisions, fluctuations resulting from strong non-stationary noise in the VAD outputs are further smoothed out. The proposed algorithms are tested with noisy speech databases with additive white, car, and factory noise over a wide range of signal-to-noise ratios. Separate evaluations are done to study the impact of gender dependence on the algorithm. Moreover, the performance of the proposed speech classifier is further assessed via its effect on the performance of a speaker verification system [NPHKon]. There, a voice activity detector, which is built from outputs of the speech classifier, is used in the front-end processing stage.

The remainder of this chapter is structured as follows: In the next section, the application of the TEO and other pre-processing methods to extract a wavelet feature are presented. In section 3.3 the quantile filtering method and the slope tracking method are designed to derive the adaptive noise threshold. The noise threshold estimate is improved by an adaptive selection of the quantile factor. In section 3.5, the performance of the proposed algorithms is evaluated and discussed. After that, section 3.6 presents the application of VAD in a speaker verification system. The performance of the proposed VAD is assessed via the recognition rate of the system. The conclusion finalizes the chapter.

3.2 Robust feature extraction

Based on the observed wavelet power distribution in section 2.2, an efficient wavelet feature is extracted to classify every noisy speech frame into one of three classes: Voiced/Unvoiced/Silence. An adaptive threshold is estimated for making the classifying decision. Thus, a block scheme of the proposed classifier is given as follows:

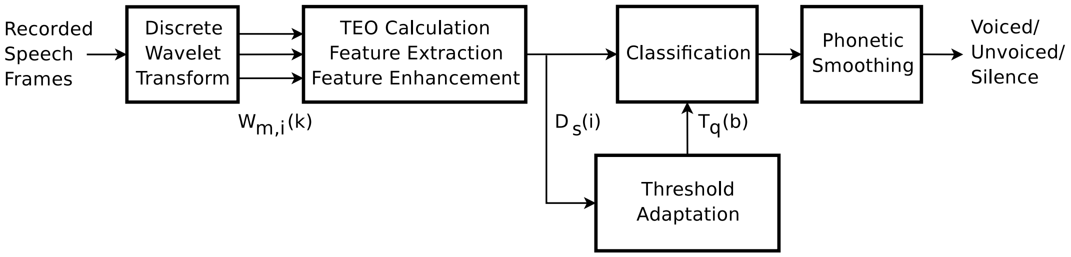

Figure 3.1: Block diagram of the proposed speech classifier.

3.2.1 Teager Energy Operator

Teager's studies in [HT90, Tea90] on nonlinear speech modeling pioneered the importance of analyzing speech signals from an energy point of view. A simple nonlinear energy tracking operator is derived, for a continuous-time signal $x(t)$ as:

$$\Psi_c[x(t)] = \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right). \tag{3.1}$$

and for a discrete-time signal $x(l)$ as:

$$\Psi[x(l)] = x^2(l) - x(l-1)x(l+1). \tag{3.2}$$

These operators were introduced systematically by Kaiser [Kai90a, Kai90b]. The TEO has been considered as an efficient nonlinear operator for many speech processing algorithms and speech applications. A study in [MKQ93] motivated the use of TEO for general speech modeling. After that, the discrete TEO has been applied for stress detection by analyzing nonlinear feature characteristics in specific frequency bands [RHM+02]. As reported by [BR01, CW04], application of the TEO can enhance the discriminability of speech components from noise.

Dealing with the speech classification task, after applying the DWT for every input frame, the TEO coefficients $E_{m,i}$ are calculated from the wavelet coefficients by the discrete form in 3.2 as follows:

$$E_{m,i}(n) = X_{m,i}^2(n) - X_{m,i}(n+1)X_{m,i}(n-1). \tag{3.3}$$

where $X_{m,i}$ are wavelet coefficients derived from the DWT at the m^{th} scale of the i^{th} frame and n is the coefficient index as explained in section 1.2. An output consisting of approximation and detail parts derived from the wavelet decomposition on an unvoiced frame at the 3^{rd} scale is shown in Fig. 3.2.

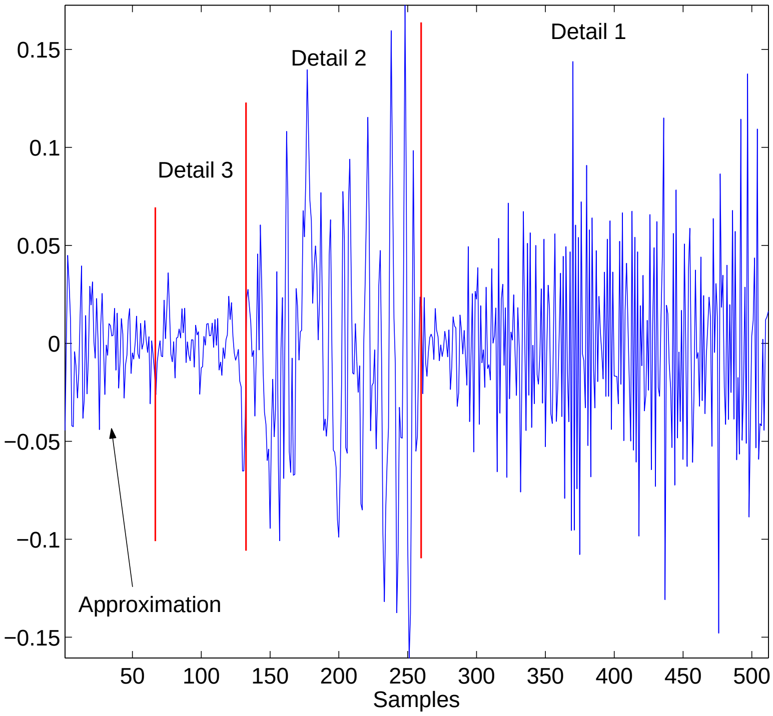

Figure 3.2: Approximation coefficients and detail coefficients derived by WD at 3^{rd} scale for a clean recording unvoiced frame.

As observed in chapter 2 for unvoiced frames, the signal energy is mostly distributed in the detail parts and less in the approximation part. However, it is hard to observe the same energy distribution at very low SNR conditions as shown in the upper diagram of Fig. 3.3. Very high energy is concentrated in the low-frequency subband. The benefit of the TEO is a magnification of the amplitude difference between the wavelet coefficients in the approximation subband and the ones in the detail subbands as depicted in the lower diagram of Fig.3.3. We see that the power difference D (which is calculated in Equation 3.4) from TEO coefficients is larger than the one of wavelet coefficients. This improvement is certainly useful to extract a robust feature for the classification task in harsh environments, especially for the unvoiced class.

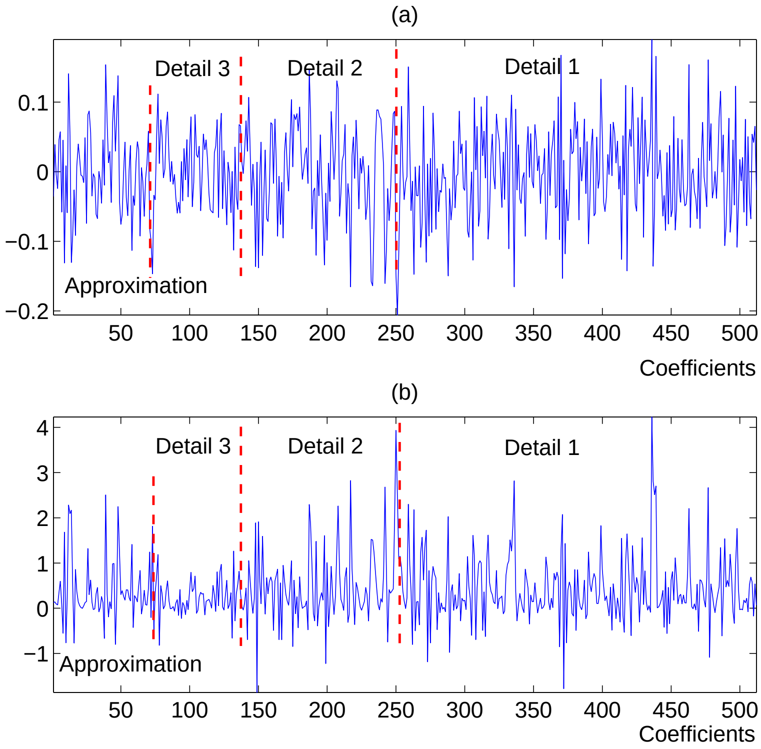

Figure 3.3: WD at 3^{rd} scale of an unvoiced frame with additive white noise at 5dB SNR: (a) Wavelet coefficients, (b) TEO coefficients.

3.2.2 Sigmoidal delta feature

As reported in Section 2.2.2, the power of the voiced frames is mostly contained in the approximation subband and much less in the detail subbands, and vice versa for the unvoiced frames. A relatively equal power distribution occurs for the silence frames. From the statistical properties of speech sounds, we observe that the spectrogram power in the range $[0.0 - 1.0]$ kHz is very low for unvoiced fricative frames in comparison with voiced frames consisting of vowels and voiced fricatives. In this study, the decomposition scale is chosen as $m = 3$ to consider the relation between the frequency band $0 - 1kHz$ and other higher frequency bands of the speech signal sampled at $f_s = 16$kHz. A delta parameter which is the power difference between the approximation subband

and the detail subbands is extracted as:

$$D(i) = \frac{1}{N_a}\sum_{n=1}^{N_a} E^2_{m,i}(n) - \frac{1}{N-N_a}\sum_{n=N_a+1}^{N} E^2_{m,i}(n). \tag{3.4}$$

where $N_a = \dfrac{N}{2^m}$ and $N - N_a$ are the length of the approximation subband and detail subbands, respectively. N is the number of samples in one speech frame. This delta feature differs from the feature PVD in Equation 2.2 which consider the difference between the 1^{st} detail and the 3^{rd} one.

Some voiced or unvoiced frames at low volume result in small values of D while, in general, the voiced and unvoiced frames give very high values of D with positive and negative sign, respectively. In order to balance the impact of the large range of values of D during processing, the hyperbolic tangent sigmoidal function is applied on $D(i)$ as:

$$D_s(i) = \frac{2}{1+e^{-2D(i)}} - 1. \tag{3.5}$$

With this operation, small values of the delta feature $D(i)$ resulting from weak speech frames are amplified. It also compresses very high values of the $D(i)$ feature resulting from loud speech frames. Because of the strong noise at low SNRs, the parameter D_s fluctuates with high variance even during silence segments. To make the classifier robust against noise, the parameter D_s is further smoothed by median filtering with a window length of 4 frames to keep a low delay. Median filtering is a non-linear signal enhancement technique for the smoothing of signals, the suppression of impulse noise while preserving edges. The figure 3.4 demonstrates the feature evolution which is extracted without using the TEO from speech with added car noise at SNR=5dB. Its enhanced version by calculating the TEO, applying the sigmoidal operation, and finally by median filtering shows less fluctuation and more robust behavior than the primitive feature.

3.3 Threshold adaptation by quantile filtering

To make the classifying decision, the extracted feature values need to be compared with the thresholds which are derived from speech signal data in advance. As a principle, the thresholds are set as constants which are found by doing experiments for certain conditions of recorded speech and noise sources. However, these constant threshold values will not be valid for other testing conditions. The need for threshold adaptation leads to a challenging task. The next two sections present two novel techniques for threshold adaptation.

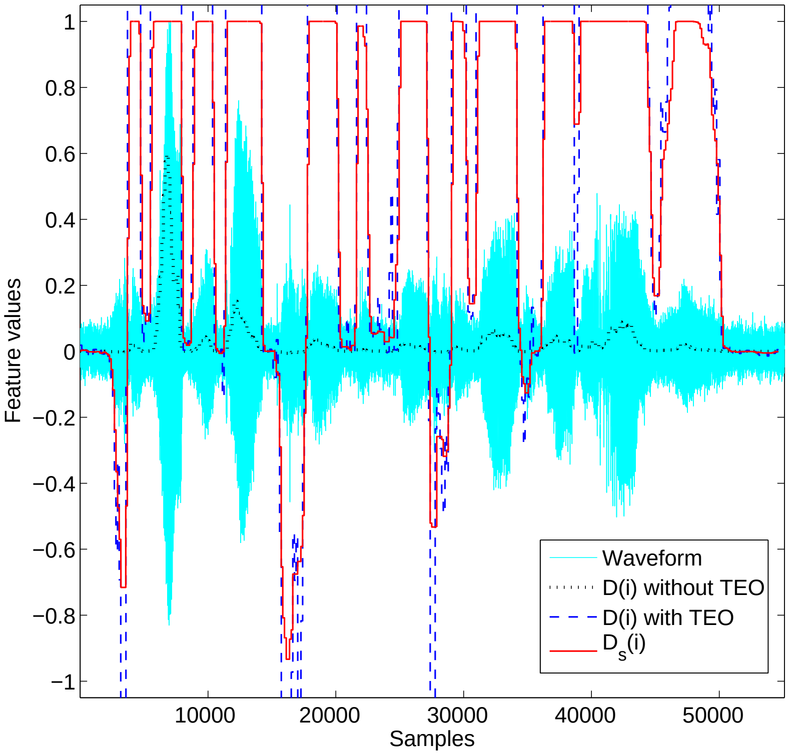

Figure 3.4: Signal waveform at SNR = 5dB and the feature extracted without applying TEO, with TEO and its enhanced version $D_s(i)$ including sigmoidal compressing and median filtering.

3.3.1 Quantile filtering

As reported in [SFB00] the signal energy in each frequency channel can be at the noise level over a significant part of the time, i.e. whenever there is no speech energy present in this band at this moment of time. Thus, the noise level can be estimated by observing the q^{th} quantile over the duration of the utterance in every channel. This proposal was applied and further developed in [PK05b] to estimate the noise level in the wavelet domain. There, pre-determined overlapped buffers, instead of the whole utterance which might continue indefinitely, are used to meet memory saving and delay requirements for the estimation algorithm.

This quantile approach is applied to threshold adaptation for the classification task. We observed that typically, at every wavelet subband, there is a high energy distribution of non-speech frames over each buffer of ten seconds in length. This means the energy in each wavelet subband stays at the noise level over a significant part of the buffer. Because the delta parameter D_s represents an energy relation between subbands, its values also stay at the noise level over a significant part of the buffer. If a threshold relating to the noise level can be determined, speech frames can be detected out of non-speech frames. In a second step, voiced and unvoiced frames will be classified due to specific properties of their delta feature values D_s.

3.3.2 Threshold adaptation algorithm

The threshold adaptation is implemented in two steps:

- First, the delta values $D_s(i)$ are sorted in ascending order over a buffer b of one second length storing N_f frames $\Rightarrow D_s(i'),\ i' = [1 \dots N_f]$.
- Second, the quantile threshold T_q is determined by taking the q^{th} quantile:

$$T_q(b) = D_s(i') \quad \Big| \quad_{i'=\lfloor qN_f \rfloor} \tag{3.6}$$

The quantile factor $q = 0.3$ shows in Fig. 3.5 has been selected experimentally over the range of values $q = 0.0, 0.1, \dots, 1.0$. Quantile filtering is actually a generalization of the minimum statistics approach [Mar94]. If the quantile factor $q = 0$, it coincides with the minimum statistics principle. The quantile threshold estimate is adaptive because it is updated for every buffer of one second in length. Thanks to this adaptability, the method is absolutely suitable to be used for estimation the noise threshold in case of non-stationary noise conditions. To make the V/U/S decisions, the enhanced delta parameter $D_s(i)$ of each input speech frame is calculated and compared with the adaptive threshold by the following rule:

$$D_s(i) = \begin{cases} \text{Voiced,} & \text{if } D_s(i) > T_q, \\ \text{Unvoiced,} & \text{if } D_s(i) < -T_q, \\ \text{Silence,} & \text{otherwise.} \end{cases} \quad . \tag{3.7}$$

3.4 Threshold adaptation by slope tracking

The proposed quantile method needs memory for storing delta values in the buffer and one frame delay. To lower the memory and delay requirements, a slope tracking method is proposed for the processed feature.

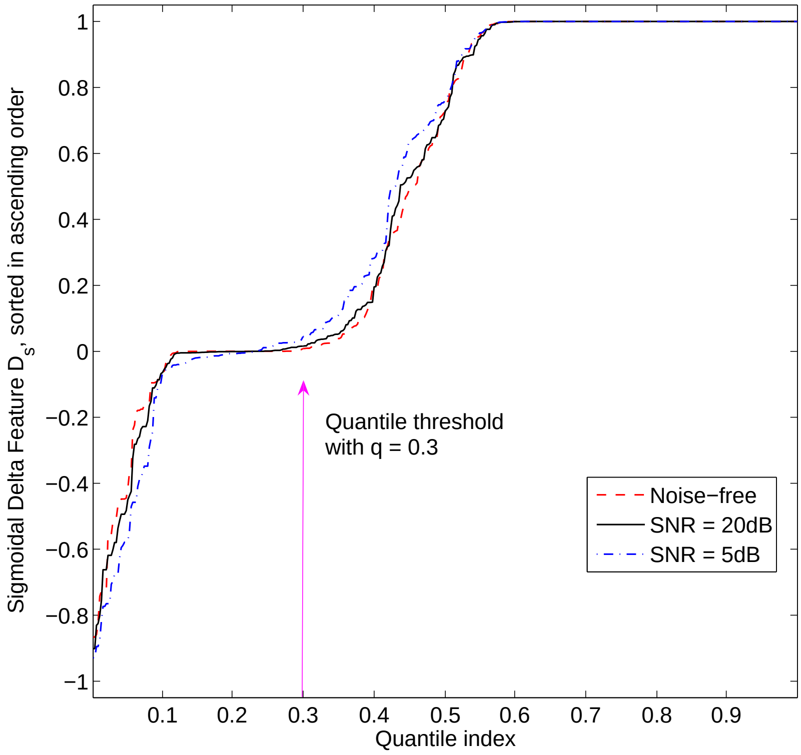

Figure 3.5: Rank order distribution of delta feature for one subband at three different conditions: noise-free, SNR=20dB and SNR=5dB observed for 427 frames. The quantile factor $q = 0.3$ is selected after visual inspection of this distribution for the strongly different noise conditions.

3.4.1 Slope generation

The frame-based values $D_s(i)$ of a speech signal are filtered by a one-pole IIR filter as:

$$D_f(i) = D_f(i-1) + D_s(i). \tag{3.8}$$

Although this filter is unstable, it is useful to distinguish between silence and unvoiced sounds which are visible as flat and downward slopes, respectively, at the output of this filter. By using finite-length buffers for processing of input segments, the steady increase towards infinity, which results from the unstable filter, can be limited. In general, the parameter D_s has positive values for voiced frames, negative values for unvoiced frames and approximates zero for silence frames. Because the filter operates as the cumulative sum of the elements of D_s, this results in the output parameter D_f as upward slope, downward slope, and almost flat regions for voiced, unvoiced, and silence classes, respectively (depicted in Fig. 3.6). In noisy environments, the filtered parameter D_f still shows the same slope characteristics as clearly even at the very low SNR of Fig. 3.7. From that, the phonetic segments can be classified by a slope detection method which is described later.

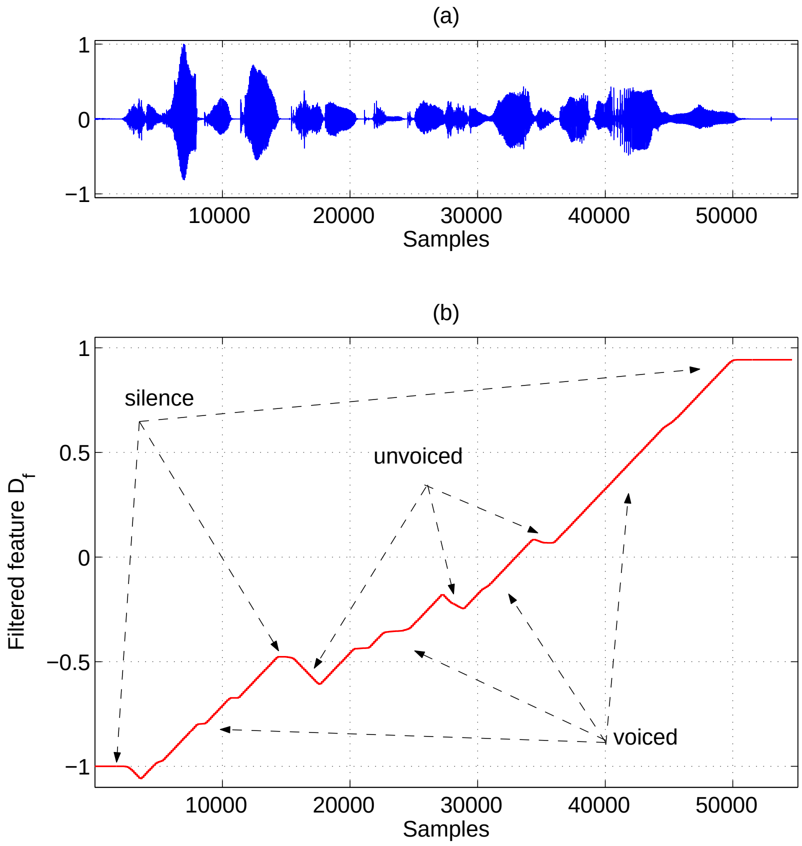

Figure 3.6: (a) Noise-free speech signal, (b) The filtered feature D_f with the assigned phonetic classes.

3.4.2 Slope detection

To detect the rising and falling slopes as well as the flat regions of the parameter D_f, a three-step method is proposed as shown in Fig. 3.8. This method detects the beginning and end points of the different phonetic segments online:

- First, if the magnitude difference of the parameter $D_f(i)$ between the current frame and the previous frame is bigger than a positive threshold $T_p = 0.5$ or smaller than a negative threshold $T_n = -0.1$, i.e. if $D_s(i)$ is bigger or smaller, then the index of the

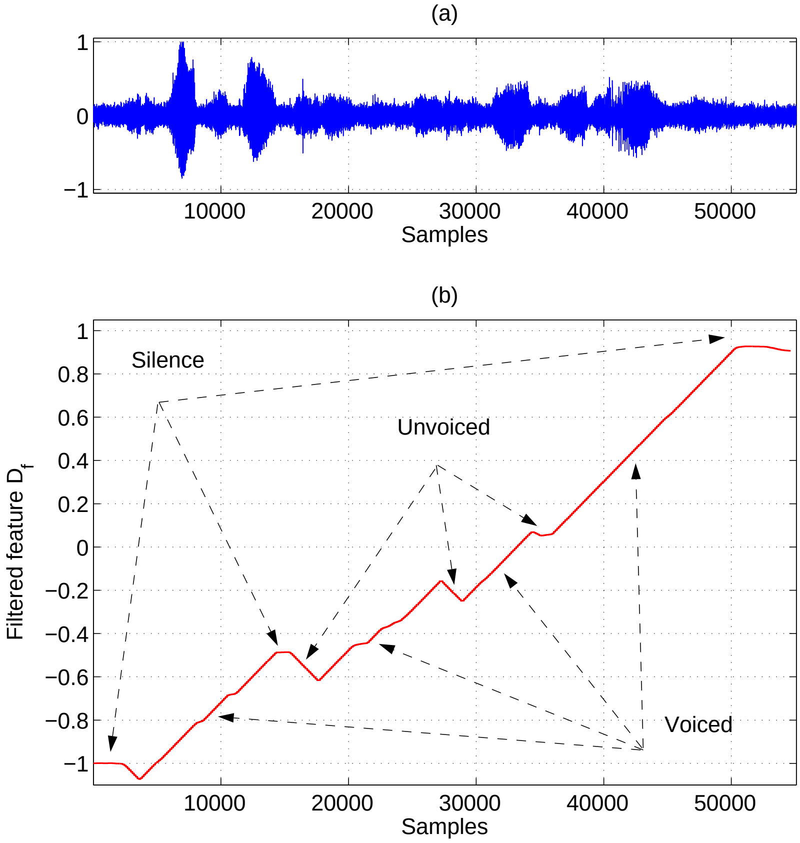

Figure 3.7: (a) Noisy speech signal at SNR=5dB, (b) The filtered feature D_f with the robustly assigned phonetic classes.

previous frame is marked as the beginning point of a non-silence phonetic segment and saved in memory.

∗ Second, the described procedure is repeated for every couple of the neighboring frames until it fails. That means the difference is smaller or larger than the selected thresholds. Then the total magnitude difference D_T of the smoothed parameter D_f between the beginning frame and the current frame is calculated and compared with another threshold $T_c = 10$ and $T_c = 1.5$ for voiced and unvoiced classes, respectively. If D_T is higher than T_c, the position of the current frame is marked as the end point and the segment is labeled with the corresponding class. Otherwise, the beginning point is replaced by the current point and the process is continued till the end of the buffer.

∗ Third, after all voiced and unvoiced segments are found, the remaining segments will be marked as silence automatically.

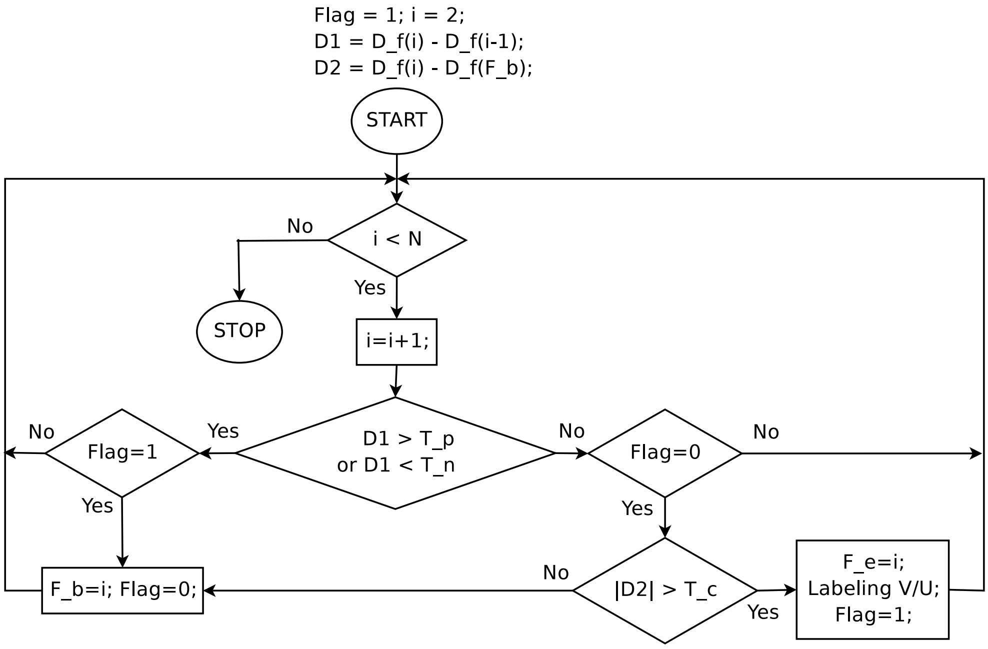

Figure 3.8: Slope detection.

The advantage of processing with the running integrator, as shown in Fig 3.7, lies in saving memory and lower delay. Only the beginning and end points need to be stored in memory. In addition, it exploits the continuity of speech sounds to classify long segments into phonetic classes without separate frame by frame detection as in most other current algorithms. This provides highly reliable performance for continuous sounds.

For both proposed adaptation methods, i.e. quantile filtering and slope tracking, occasionally occurring incorrect decisions, which are due to non-stationary noise and transient sounds, are repaired by a frame-based smoothing method which enforces sequential consistency of speech sounds such as: VVSVV → VVVVV, etc. Figure 3.9 demonstrates the evolution of the extracted delta feature which is enhanced by the sigmoidal function $D_s(i)$, and by the unstable filter, $D_f(i)$. The V/U/S labels are done based on the quantile filtering in $D_s(i)$ and the slope tracking for $D_f(i)$ in a noisy recording.

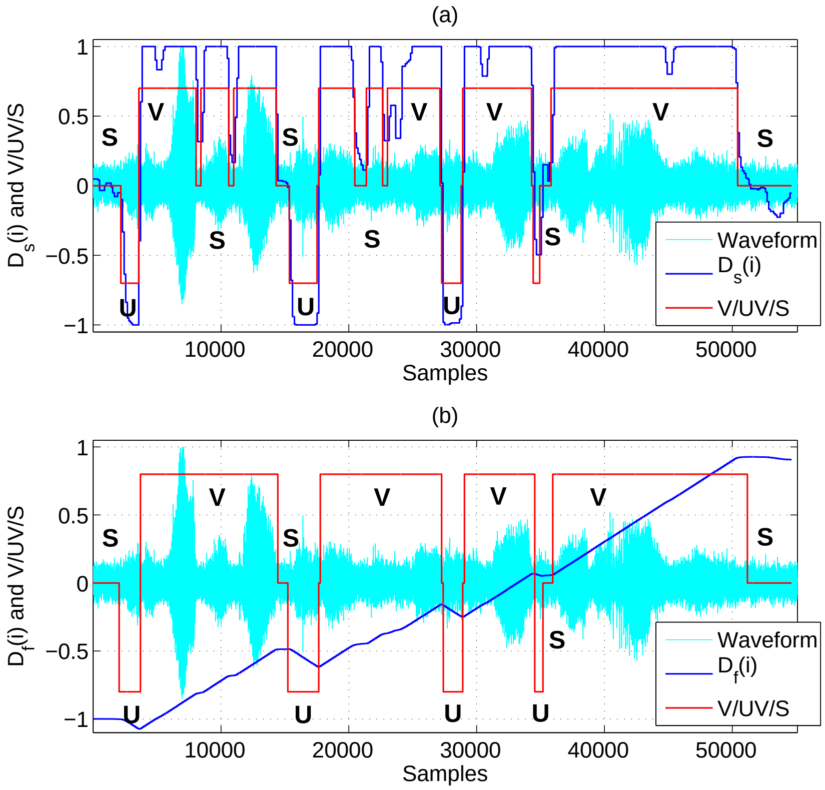

Figure 3.9: Speech signal with additive car noise at SNR = 5dB: (a) Enhanced sigmoidal delta feature $D_s(i)$ and V/U/S labels derived by quantile filtering of $D_s(i)$, (b) same but with slope tracking on $D_f(i)$.

3.5 Evaluation of classification rate

3.5.1 Experimental setup

The speech data used to build the experimental dataset is extracted from the TIMIT database [GLF+93]. Two gender-dependent sets of female (F) and male (M) speakers are selected from dialect speaking region 4 with 80 continuous utterances totally (ca. 10100 frames of 32 ms frame length and 8 ms frame rate). The selected dataset is divided into two subsets, 70% for training and 30% for testing. The speaker set contains 4 female speakers and 4 male speakers. In this research, the closure and release frames of plosives and affricates are excluded from the analysis because they are difficult to classify as voiced or unvoiced sounds. All non-speech events are also removed completely. So we are left with about 4900 frames for the training set and 2100 frames for the test set. The speech signals, which are sampled at f_s=16kHz are artificially corrupted with additive white, car, and factory noise over the SNR range of [30, 20, 10, 5]dB. This processing yields 73500 and 31500 of noisy speech frames used for the training phase and testing phase, respectively. The noise data which have obtained from the Signal Processing Information Base [StNSF95] is filtered by an anti-aliasing filter and downsampled to the sampling rate f_s=16kHz from f_s=19.98kHz. The reference labels (V/U/S) of the speech frames, which are derived automatically from the phonemic TIMIT transcriptions, are compared with the phonetic class labels at the output of the classifiers. The two proposed classifiers using the DWT with adaptive quantile filtering (AQF) and adaptive slope tracking (AST) are compared with the method proposed in [Chi00]. There glottal closure indices (GCI) is extracted for detecting voiced frames, and then the short-term log-energy is calculated to classify between unvoiced and silence frames. From our experience, the use of GCI which is based on the LP residual derived from speech frame provides quite robust pitch estimation in hars environments. This observation is also reported in [PY04]. So, it is referred as a baseline for classifying voiced and unvoiced classes.

3.5.2 Results and discussions

Because the relative frequencies of the three classes are not the same, i.e. 63.98%/ 24.20%/ 11.82% for the V/U/S classes, respectively, the average classification rate is calculated as a ratio between the sum of correctly accepted frames and the total number of tested frames over all classes. Table. 3.1-3.3 show the average classification rates calculated from the confusion matrices over all frames of three classes V/U/S for the mixed gender (Ma+Fe), female (Fe) and male (Ma) speaker datasets, including noise-free and noisy recordings. The standard deviation values are estimated by formula in appendix C.

For all three different types of noise, the AQF method provides up to 2.5% lower error rate than the AST method. Moreover, by considering very harsh conditions with SNR = 5dB, we see that the classfication rate differences between the AQF method and the AST method are very small in case of stationary white noise, and larger for complex factory noise. This is expected because the AQF method is based on a globally adaptive threshold while the AST method is not. However, the latter method with the running integrator saves memory during the tracking of phonetic classes. We observe that these wavelet-based methods have lower average error rates than the GCI-based method [Chi00] for clean speech and noisy speech with SNRs down to 5dB. Due to the highest complexity of factory noise which includes transient, colored, and non-stationary noise, the average performance over the considered SNR range derived by

SNR (dB)		Clean	30	20	10	5
Algos.	Gender					
AQF	Ma+Fe	93.31±0.55	92.16±0.59	91.09±0.62	89.25±0.68	86.62±0.74
	Fe	93.77±0.75	93.38±0.77	92.37±0.82	91.53±0.86	87.51±1.02
	Ma	93.22±0.78	92.89±0.79	91.67±0.85	90.02±0.93	86.53±1.05
AST	Ma+Fe	92.42±0.58	90.27±0.65	90.39±0.64	87.42±0.72	85.57±0.77
	Fe	93.53±0.76	92.92±0.79	92.12±0.83	90.77±0.89	86.36±1.06
	Ma	92.37±0.82	91.02±0.88	91.08±0.88	88.83±0.97	85.77±1.08
GCI	Ma+Fe	91.58±0.61	89.94±0.66	89.22±0.68	87.09±0.73	82.76±0.82
	Fe	92.88±0.79	92.61±0.81	91.37±0.87	89.61±0.94	84.19±1.13
	Ma	91.35±0.87	90.88±0.89	90.18±0.92	87.76±1.01	83.0±1.16

Table 3.1: Average classification rates (%) in case of white noise.

SNR (dB)		Clean	30	20	10	5
Algos.	Gender					
AQF	Ma+Fe	93.31±0.55	91.34±0.61	89.15±0.68	87.96±0.71	83.17±0.82
	Fe	93.77±0.75	93.00±0.79	91.84±0.84	90.07±0.92	85.44±1.09
	Ma	93.22±0.77	91.74±0.85	89.98±0.93	88.75±0.97	84.02±1.13
AST	Ma+Fe	92.42±0.58	90.63±0.64	88.28±0.70	86.69±0.74	81.71±0.84
	Fe	93.53±0.76	92.54±0.81	91.11±0.88	88.79±0.97	83.66±1.14
	Ma	92.37±0.82	91.07±0.88	89.49±0.95	87.53±1.02	82.11±1.18
GCI	Ma+Fe	91.58±0.61	90.12±0.65	88.09±0.71	87.53±0.72	77.86±0.91
	Fe	92.88±0.79	91.91±0.84	90.47±0.91	86.15±1.07	80.94±1.21
	Ma	91.35±0.87	90.22±0.92	88.63±0.98	88.74±0.98	78.77±1.26

Table 3.2: Average classification rates (%) in case of car noise.

SNR (dB)		Clean	30	20	10	5
Algos.	Gender					
AQF	Ma+Fe	93.31±0.55	90.94±0.63	87.52±0.72	83.79±0.80	81.79±0.84
	Fe	93.77±0.75	92.11±0.83	90.13±0.92	86.97±1.04	81.62±1.20
	Ma	93.22±0.78	91.22±0.87	88.47±0.99	84.91±1.10	82.83±1.16
AST	Ma+Fe	92.42±0.58	91.02±0.62	85.76±0.76	82.88±0.82	75.88±0.93
	Fe	93.53±0.76	91.18±0.88	88.07±1.00	85.33±1.09	79.91±1.24
	Ma	92.37±0.82	91.66±0.85	86.19±1.06	83.79±1.14	76.73±1.30
GCI	Ma+Fe	91.58±0.61	89.19±0.68	83.61±0.81	78.29±0.90	73.79±0.96
	Fe	92.88±0.79	90.22±0.92	86.95±1.04	82.04±1.18	76.19±1.31
	Ma	91.35±0.87	89.31±0.95	84.16±1.13	79.48±1.25	74.99±1.34

Table 3.3: Average classification rates (%) in case of factory noise.

both wavelet-based methods drops down by about 6% and 3% in comparison with the performance of the white and car noise cases, respectively.

As observed from Tables 3.1-3.3, the mixed-gender dataset results in higher error rates than the gender-dependent datasets for all three classifiers. The average difference of the average error rates over the given range of SNRs obtained by the wavelet-based methods is lower than the difference observed for the GCI-based method [Chi00], (1.33% compared with 1.77%), but higher than the one in [PK05a] (1.14%). The classification performance for the three classes obtained by the proposed algorithms for clean speech are higher than the ones obtained by the joint neural network classifiers proposed in the previous chapter. This is expected because the joint FNN classifier has been designed to classify six classes. There, the false acceptance rates of the voiced and unvoiced classes rise because of the wrong assignments of speech frames belonging to the mixed-excitation class.

Table 3.4: Confusion matrix in [%] for three classes tested with Ma+Fe data set which is contaminated by additive white noise at SNR = 5dB.

Class assignments		V	UV	S
Transcriptions	Methods			
V	AQF	**86.87±0.92**	4.60±0.57	8.53±0.76
	AST	**85.57±0.96**	5.38±0.62	9.05±0.78
	GCI	**83.59±1.01**	5.85±0.64	10.56±0.84
UV	AQF	5.33±0.99	**85.89±1.54**	8.78±1.26
	AST	5.82±1.04	**85.52±1.56**	8.66±1.25
	GCI	7.72±1.18	**82.15±1.70**	10.13±1.34
S	AQF	5.93±1.50	7.34±1.66	**86.73±2.15**
	AST	6.50±1.57	8.67±1.79	**84.83±2.28**
	GCI	8.56±1.78	11.87±2.05	**79.57±2.56**

Now we examine the performance of each phonetic class in terms of their confusion matrices. Tables 3.4-3.6 show the true acceptance, false rejection and false acceptance rates for three classes obtained for the Ma+Fe data set with white, car, and factory noise at SNR = 5dB, respectively. Depending on the characteristics of the additive noise, the detection rates of different classes vary differently. We first discuss the results obtained by AQF, the most reliable method. There is a significant decrease of the classification rate of the silence class when going from white to car noise, then factory noise (86.73%, 81.55% and 78.82%, respectively). The reason is that more

Class assignments		V	UV	S
Transcriptions	Methods			
V	AQF	**84.18±1.00**	10.07±0.82	5.75±0.64
	AST	**82.93±1.03**	10.33±0.83	6.74±0.68
	GCI	**79.20±1.11**	12.51±0.90	8.29±0.75
UV	AQF	11.14±1.40	**81.27±1.73**	7.59±1.18
	AST	12.93±1.49	**78.85±1.81**	8.22±1.22
	GCI	14.02±1.54	**75.27±1.91**	10.71±1.37
S	AQF	10.26±1.93	8.19±1.74	**81.55±2.46**
	AST	10.34±1.93	8.72±1.79	**80.94±2.49**
	GCI	13.23±2.15	10.90±1.98	**75.87±2.72**

Table 3.5: Confusion matrix in [%] for three classes tested with Ma+Fe data set which is contaminated by additive car noise at SNR = 5dB.

Table 3.6: Confusion matrix in [%] for three classes tested with Ma+Fe data set which is contaminated by additive factory noise at SNR = 5dB.

Class assignments		V	UV	S
Transcriptions	Methods			
V	AQF	**82.78±1.03**	12.91±0.91	4.31±0.55
	AST	**77.63±1.14**	14.68±0.96	7.72±0.73
	GCI	**75.81±1.17**	15.84±1.00	8.35±0.75
UV	AQF	12.18±1.45	**80.63±1.75**	7.19±1.15
	AST	17.06±1.67	**73.96±1.95**	8.98±1.27
	GCI	18.23±1.71	**71.18±2.01**	10.59±1.37
S	AQF	13.68±2.18	7.50±1.67	**78.82±2.59**
	AST	19.12±2.50	10.43±1.94	**70.45±2.89**
	GCI	20.73±2.57	11.02±1.99	**68.25±2.96**

non-stationary noise occurs in car and factory conditions. A smaller change occurs for the unvoiced class, with a smaller gap between car noise and factory noise (85.89%, 81.27% and 80.63%). The results show that the classification rates of the voiced class vary slightly (86.87%, 84.18% and 82.78%). These observations prove that, by applying the wavelet-based classifiers, the ability of detecting voiced and unvoiced classes are somewhat robust in noisy conditions. With the colored noise in case of car and factory conditions, the false acceptance rates of the voiced class rise to higher values than the one of white noise.

3.6 Evaluation in application domain

As an extended evaluation of the robustness of the proposed phonetic classifier, we assess its performance via its effect on a speaker verification (SV) system. The phonetic outputs of the speech classifier are used to build a voice activity detector (VAD) for the front-end unit of the SV system. The incorporation of VAD is one of the approaches to make the SV system robust against noise. As reported in [MCGB01], VAD has a main impact on SV performance under noisy conditions. In our research, a VAD is integrated into an SV system which was designed in [NPHKon] to increase the security level of air traffic control (ATC). When a pilot registers the first time to a control sector his voice is automatically enrolled. At any later occurrence of the same aircraft identification tag [NPHKon], the received voice message is verified against the existing speaker model. The design of the ATC-oriented SV system is based on training Gaussian mixture models (GMMs) and a universal background model (UBM) [Bea00].

3.6.1 VAD as a pre-processing stage

[4] The main challenges for the ATC-oriented SV system are the degradation of the transmitted noisy signal, caused by the fading channel and the bandwidth limitation of the transceiver equipment. A frequency range of only $300-2500$ Hz is used for speech transmission in ATC [NPHKon]. Due to the lack of an ATC speech database, we only deal with the bandwidth limitation by simulating this condition on other databases. The design of the SV system consists of four phases as shown in Figure 3.10. In phase 1, gender-dependent UBMs are trained. These trained models are then applied in phase 2 for speaker dependent modeling using gender information from gender recognition. Retraining of a speaker model is performed in phase 3, and finally, in phase 4 the verification task is carried out. A detailed description of the SV system can be found in [NPHKon].

As an important pre-processing stage, VAD is employed in the front-end unit as depicted in Fig. 3.10 to detect only segments with voice activity which are used to

[4] Thanks to M. Neffe for his SV system, and his acceptance for re-editing text material for introducing ATC, SV system, and experimental database setup.

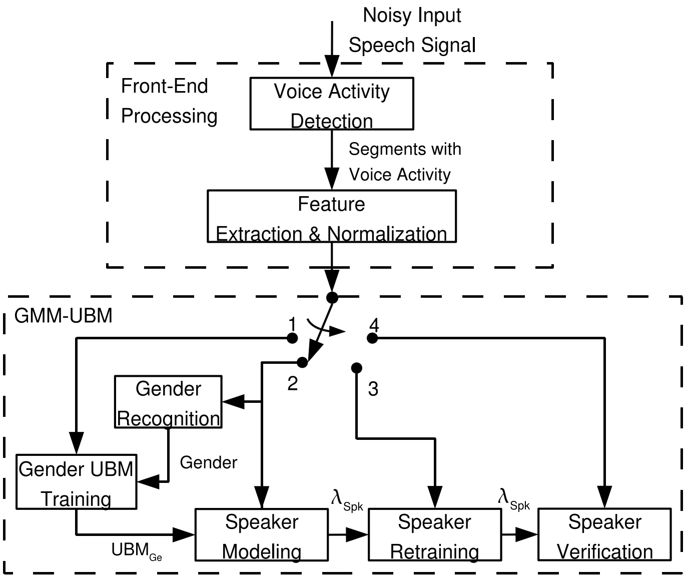

Figure 3.10: VAD as pre-processing stage in the SV system [NPHKon].

extract linear-frequency cepstral coefficients. Robust VAD is crucial in the SV system in order to extract suitable speaker dependent data. Non-speech data which is contaminated by noise of the transmission channel may drive the model training process into incorrect convergence, thereby leading to an unreliable SV system. A robust VAD is designed according to following scheme:

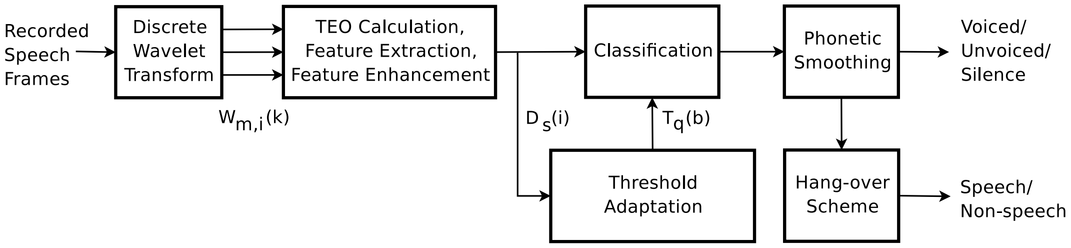

Figure 3.11: A block diagram showing the VAD based on outputs of the phonetic classifier (Figure 3.1).

To meet the specific properties of the air traffic voice channel, some adjustments on the proposed phonetic classifier have been done. Dealing with the voice signal with a bandwidth of $[0.3 - 2.5]$ kHz, we choose a decomposition scale $m = 1$ to consider the relation between a low-frequency subband of $[0.3 - 1.1]$ kHz and the remaining higher-frequency subband. From the temporal characteristics of the input utterances, we observe that the ratio between the number of speech frames and the number of non-speech frames varies for different buffers. Thus, the quantile factor $q(b)$ is apdated for each buffer to achieve a better noise threshold estimate. The method is based on a comparison of feature differences cross five consecutive rank-sorted frames and a predetermined level $\varepsilon = 10^{-3}$. The test is performed from the beginning of the buffer and is stopped when the difference is larger than this level. Then the quantile factor $q(b)$ is selected as:

$$q(b) = i', \qquad \text{if} \quad D_s(i') - D_s(i' - 4) > \varepsilon \tag{3.9}$$

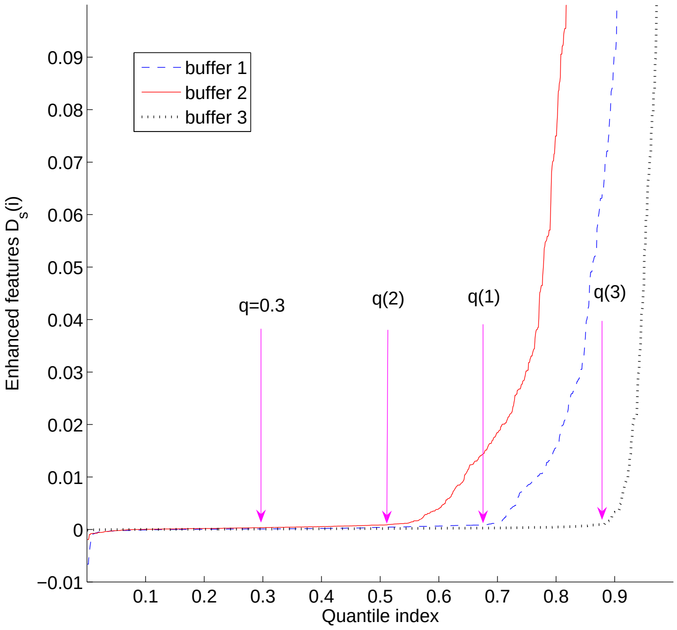

Figure 3.12: Adaptive quantile factors selected for different examples of frame buffers.

At the output of the speech classifier, all voiced and unvoiced frames are relabeled as speech frames and the remaining silence frames are considered as non-speech frames. For smoothing of fluctuations resulting from strong non-stationary noise in the VAD outputs, a hang-over scheme [PKW+06] has been applied. The output sequence VAD(i) is smoothed by using the 100ms/200ms hangover scheme to bridge short voice activity regions, preserving only candidates with a minimal duration of 100ms, and being not more than 200ms apart from each other. This excludes talk-spurts shorter than 100ms and relabels pauses smaller than 200ms. The impact of the bridging rule is considered during experiments.

The designed VAD based on wavelet sigmoidal delta feature is compared with other VAD methods based on short-term log-energy with constant quantile filtering [NPHKon], and short-term log-energy with SNR-based adaptive threshold which was designed for the advanced front-end of the distributed speech recognition ETSI ES 202

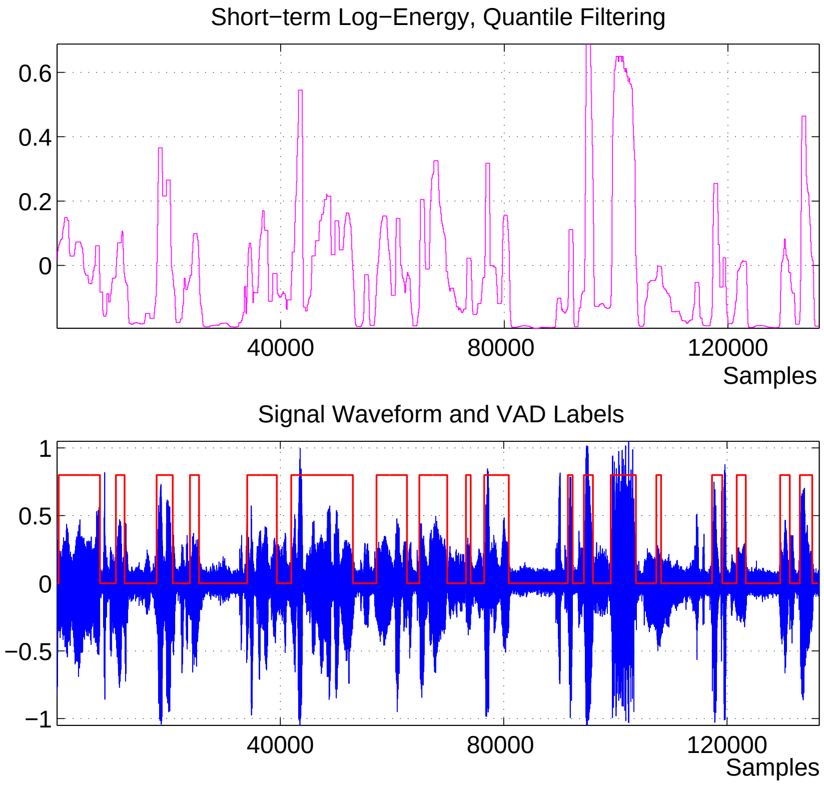

Figure 3.13: VAD based on short-term log-energy with constant quantile filtering: (top) Extracted feature, (bottom) speech signal at SNR = 10dB and VAD labels.

050 [ETS03]. Tests of three examined VAD methods on a car noise recording with SNR = 10dB are demonstrated from Figure 3.13 to Figure 3.15. These informal tests show the best results for voice activity detection with the proposed wavelet method.

3.6.2 Experiments and evaluations

For the experimental evaluation, we use the fixed SPEECHDAT-AT database, consisting of noisy, fixed network telephone recordings [BEK00], and the WSJ0 database consisting of almost noise-free broadcast recordings [GGPP93]. Dialect regions and speaker ages were selected randomly. In order to simulate the conditions of ATC, all files were band-pass filtered to a bandwidth from 300 Hz to 2500 Hz and down-sampled to a sampling frequency of 6 kHz. Furthermore, the databases were cut artificially in utterances of 5 seconds which corresponds to typical talk spurt lengths in ATC.

Gender-dependent UBMs were trained as GMMs with 38 Gaussian components using two minutes of recordings for each of 50 female/male speakers selected ran-

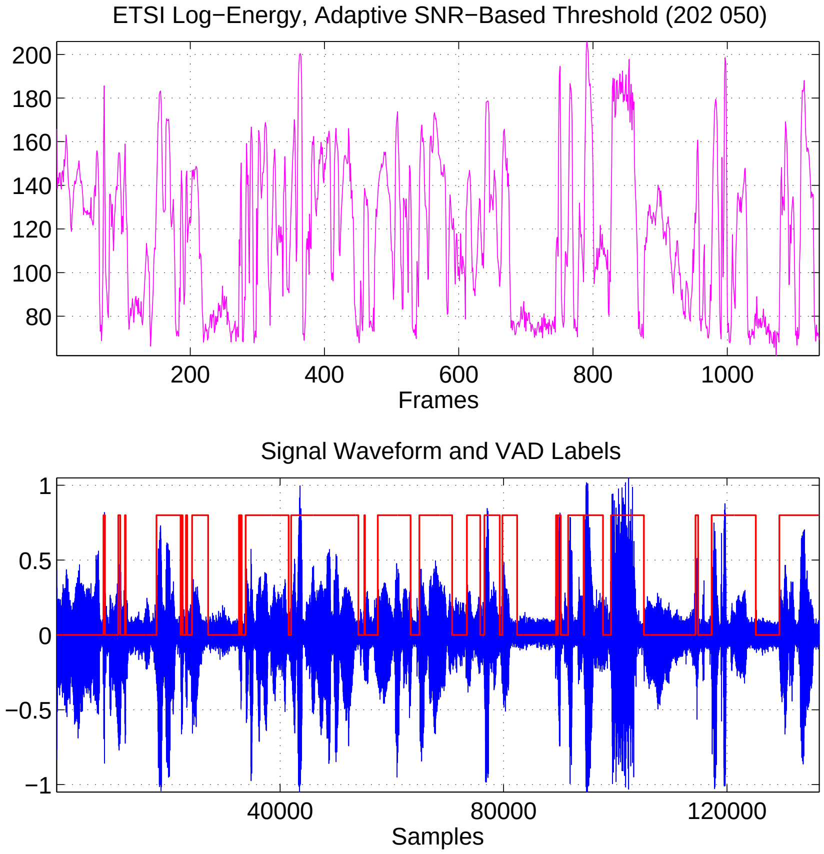

Figure 3.14: VAD based on short-term log-energy with updated SNR-based threshold: (top) Extracted feature, (bottom) speech signal at SNR = 10dB and VAD labels.

domly from each of the 100 female/male speakers extracted from the SPEECHDAT-AT database. Out of the remaining 50 female and 50 male speakers for testing, 20 speakers were marked as reference speakers, 10 male and 10 female each. For each speaker, 6 utterances were used for verification. So each reference speaker model was compared to 600 utterances, yielding a total of 12000 test cases for 20 reference speaker models all together. For the experiment on the WSJ0 database, 23 female and 28 male speakers from CD 11_2_1 was used to train the gender dependent UBMs. CD 11_1_1 with 26 female and 19 male speakers were taken for testing. Out of them, 24 speakers were marked as reference speakers, 12 male and 12 female each. For each speaker, 12 utterances were used for verification. So each reference speaker model was compared to 540 utterances, yielding a total of 12960 test cases for 24 reference speakers.

The SV performance is assessed by using the detection error trade-off (DET) curve [MDK+97] and the equal error rate (EER) [PM98]. The energy-based VAD [NPHKon] using logarithmic short-term energy with constant quantile filtering is compared with

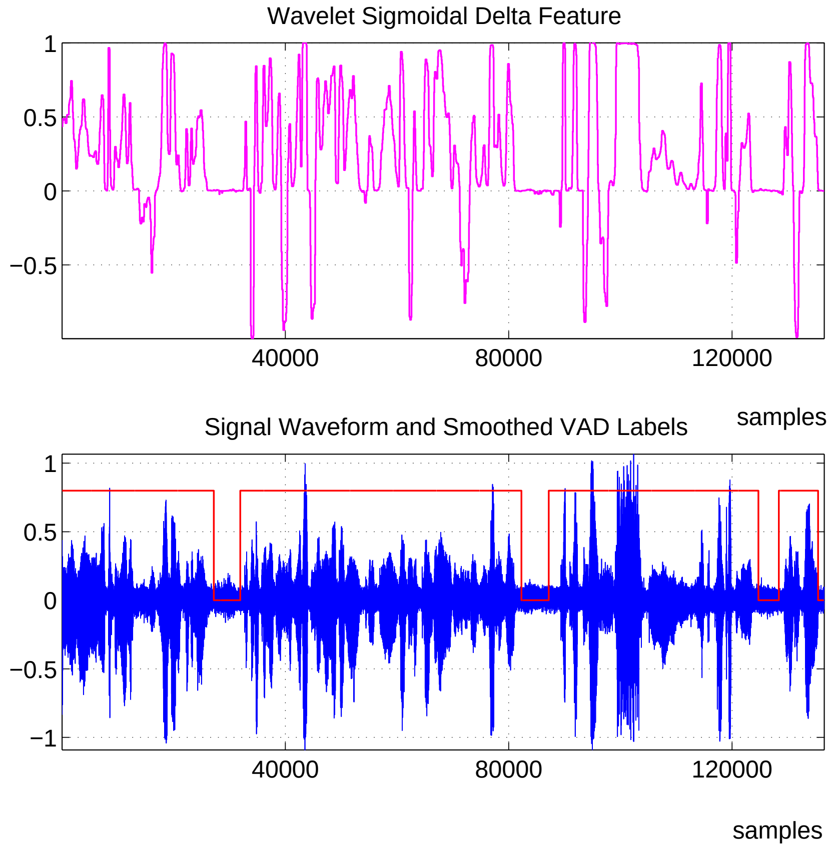

Figure 3.15: VAD based on wavelet sigmoidal delta feature with adaptive quantile filtering: (top) Extracted feature, (bottom) speech signal at SNR = 10dB and VAD labels.

the optimized wavelet-based VAD in terms of the SV performance. The hangover schemes for output smoothing are the same in both VAD methods. As reported in Table 3.6.2, the usage of both VAD methods improves SV performance significantly, compared to the case without using VAD, for the SPEECHDAT-AT database. However, for the WSJ0 database, the obtained results are almost similar. This shows a positive effect of VAD in removing noise-dominated non-speech segments which may lead to an unreliably trained SV system. For the SPEECHDAT-AT database, and comparing the wavelet-based VAD with the energy-based VAD, the EER is reduced from 11.7% to 9 % without smoothing, and from 6.52% to 4.75% with smoothing as illustrated in Fig. 3.16. Thus, by using the proposed wavelet-based VAD, we obtain a 23% and 27% relative improvement compared to the energy-based VAD in both cases. In addition, from the observed results, we discover that not only an accurate detection of speech frames but also a smoothing hangover scheme to bridge short pauses between speech frames helps to improve the SV performance.

VAD Algorithms	NoVad	EVad	WaVad
Hang-over		wo/w	wo/w
Databases			
SPEECHDAT-AT	25.12	11.7 / 6.52	9 / 4.75
WSJ0	10.15	- / 10.37	- / 10

Table 3.7: EER [%] results derived from both databases, without using VAD and using energy-based (EVad), wavelet-based VADs (WaVad) without (wo) and with (w) applying the hangover scheme.

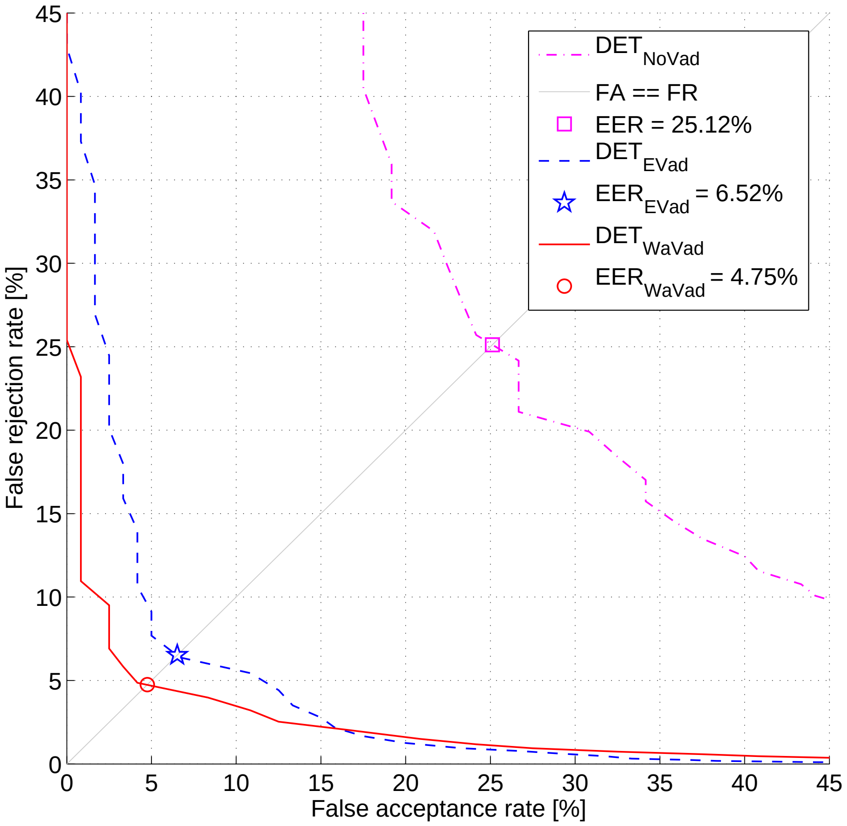

Figure 3.16: DET-curves showing the false acceptance (FA) rate versus the false rejection (FR) rate, and obtained EERs when using energy-based (EVad), wavelet-based VADs (WaVad) and without VAD (NoVad) [NPHKon].

To assess the impact of environmental mismatch between training and test conditions, a cross testing has been performed using the SPEECHDAT-AT database for training the UBMs but the WSJ0 database for testing, and vice versa. The wavelet-based VAD is employed for these experiments. In the former condition, the EER is 11.8 % which is worse than the above results because the models were trained with noisy speech and tested with clean speech. In the latter condition, the slight improvement of EER to 11 % may result from the VAD reducing noisy non-speech segments in the testing phase. In both conditions, the use of the VAD cannot solve the mismatch between the training and testing phases.

3.7 Conclusions

High-performance phonetic classification algorithms have been developed based on the discrete wavelet transform and the Teager energy operator. They exhibit a very low-complexity as the classifiers use only a single parameter. The results presented in this chapter illustrate the effectiveness of the time-scale feature extracted from the wavelet coefficients. The quantile filtering which generalizes the concept of the minimum statistics is an excellent technique for estimating the noise level accurately, especially in case of non-stationary noise. As reported in the experiments, the quantile-based threshold adaptation method provides quite robust performance but requires a buffer of frames to determine the quantile threshold, and thus results in delay. The slope tracking method uses a running integrator in three steps and saves memory and delay, while providing slightly poorer performance.

Based on the proposed phonetic classifier, a voice activity detector has been developed and built into the front-end processing stage to improve speaker verification. The VAD is specially designed for the needs of air traffic control with respect to bandwidth restriction and high noise conditions due to the ATC radio channel. The quantile filtering technique is adjusted with an adaptive quantile factor to meet those demands. Systematic tests have been performed using this system without VAD, with the energy-based VAD, and the wavelet-based VAD . The last method which is based on the enhanced wavelet feature and the adaptive noise threshold provides the best performance. The hangover scheme which smooths the VAD output also contributes to the EER reduction. Significant EER reduction is achieved under harsh environments when applying the wavelet-based VAD.

In future work, the testing on a larger database is necessary to validate the robustness of the proposed phonetic classifiers. More complicated types of noise should be considered for example babble noise consisting of interfering speaker background noise. Parameters of the wavelet-based VAD will be fine tuned to maximize EER performance. The relationship between VAD objective performance and speaker verification performance should be considered. The performance of a VAD developed from original phonetic classifier using quantile filtering with constant quantile factor will be compared with the one using the adaptive quantile factor in terms of precision and recall measures.

Chapter 4

Statistical wavelet filtering for speech enhancement

4.1 Introduction

Mobile communication has been researched and developed to meet the demands of human communication. Due to the nature of mobile technology, the communication processes are often carried out in real environments which contains a wide range of potential sources of noise and distortion that can degrade the quality of the speech signal. This makes robustness to acoustic background noise a highly motivated challenge in speech communications. Out of different approaches to enhance the transmitted speech signal such as bandwidth extension, source separation, etc., a wide range of techniques for noise reduction have been proposed to deal with varying levels and types of background noise as encountered on telephone channels, constructions, cars, babble noise, etc. The noise reduction methods not only improve the quality, acceptability and intelligibility of the speech signal for human listener, but they also crucial to achieve high performance of speech recognition.

In our research, we only study single-channel speech enhancement systems which differ from multi-channel systems where two or more channels are employed for both spatial and temporal processing. Nowadays, since most mobile devices are designed with only one microphone channel, noise estimation from a second microphone channel which would be very helpful for noise reduction algorithms is impossible. Due to the spectral overlap between speech and noise signals, however, the denoised speech signals obtained from single-channel methods exhibit more speech distortion when tuned for significant noise reduction. Nevertheless, this kind of method is relevant to keep the low cost and small size of mobile devices. Fig. 4.1 shows a block scheme of single-channel noise reduction. The signal is firstly transformed into other domains to get a better presentation of the speech signal. Based on the observed coefficients, the two following steps are implemented:

- The noise level is estimated by noise estimation block.
- Then the noise signal component is removed out of the noisy speech signal by the gain function.

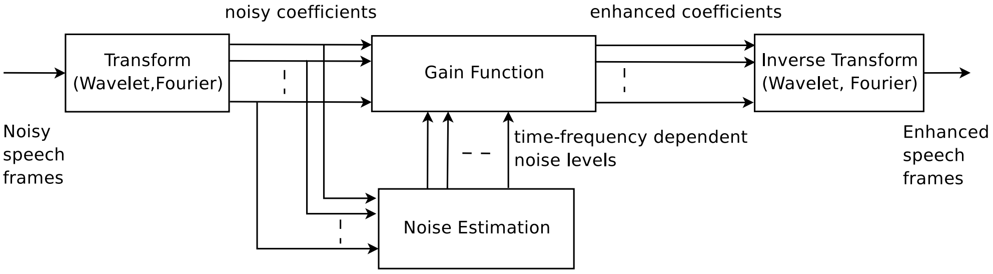

Figure 4.1: Block scheme of standard single-channel noise reduction approach.

Noise reduction can be considered as an estimation of the unknown clean speech from the observed noisy speech only. Based on different linear estimators and non-linear estimators, a gain function is designed and applied on the noisy coefficients to reduce the noise properly, thereby enhancing the speech quality. Because only the noisy speech signal is observed, an estimate of the noise signal component is required. It is one of the most difficult parts in the noise reduction algorithms, especially for non-stationary and non-white whose characteristics change the speech signal over various frequency bands. Several noise reduction methods may work fine with certain noise environments and certain applications but can fail with others. Thus, given the large diversity of noisy environments and various applications, it is clear that there cannot be one single perfect noise reduction algorithm. Single-channel noise reduction methods can be divided as follows:

- Exploiting the periodicity of voiced speech.
- Auditory model based systems.
- Optimal linear estimators.
- Statistical model based systems using optimal non-linear estimators.

The following paragraphs introduce some methods which belong to the first two approaches. After that, the widely used approaches using optimal estimators in the frequency domain and the wavelet domain are reviewed.

The first approach utilizes the basic principle that waveforms of voiced sounds are almost periodic. With this observation, an adaptive comb filter is applied on the noisy speech signal to pass the harmonics of speech but reject the frequency components between the harmonics [FSBO76]. A system closely related to comb filtering is developed in [Par76] which integrates a pitch extractor. These approaches require an accurate voiced/unvoiced classification and a good estimate of the pitch period which are problematic in harsh noisy conditions. Besides, this periodicity based method only can enhance voiced frames of the recorded speech signals. This results in very low perceptual quality and bad intelligibility in many languages where unvoiced fricatives and plosives contribute a significant amount to the word meaning.

In the second approach, application of psychoacoustic principles to speech enhancement is introduced in [CO91, TMK97] and further developed following the ideas successful audio coding based on auditory masking [Joh88]. Noise components that lie below the masking threshold can be left untouched which results in lower distortion of the enhanced speech signal [Vir99]. The incorporation of auditory masking into a subspace-based speech enhancement system is introduced in [Vet01, JC03]. Critical-band wavelet decomposition is used with a noise masking threshold in [LW03], and a perceptual wavelet packet decomposition which simulates the critical bands of the psychoacoustic model in [JM03, CW04].

4.1.1 Optimal estimation for frequency-domain denoising

Considered as a classical approach, the noise-free speech signal $\widehat{x}$ is estimated as the output of a linear frequency-response function H driven by the observed noisy speech signal y. The Wiener filter [LO79, SF96] represents the optimal solution in the minimum mean-square error (MMSE) sense for the linear estimation technique where the speech signal and the noise signal are assumed to be jointly Gaussian. Besides, the common spectral subtraction (SS) method [Bol79] is considered as a direct estimation of the short-time spectral amplitude (STSA) of the clean speech signal. The enhanced speech is obtained by subtracting the noise spectrum from the noisy speech spectrum computed with the discrete Fourier transform (DFT). The noise spectrum is estimated by several methods such as voice activity detection (VAD) [BJF94], soft-decision VAD [SS98], minimum statistics [Mar94], and quantile filtering [SFB00, PK05b]. The SS method introduces artifacts in the processed speech signal which is called "musical noise" and will be discussed late in this text. Kalman filter is also applied for speech enhancement [GKG91, KGG89].

In an effort of reducing musical noise and lowering the speech distortion, instead of applying an optimal linear estimator, optimal non-linear estimators which require knowledge of the joint probability density function of the speech and noise Fourier expansion coefficients is applied to estimate the clean speech DFT coefficients given the noisy speech DFT coefficients. The required a priori probability distribution can be estimated from training data of speech and noise or, alternatively, based on assuming a reasonable statistical modelling assumption [Eph92b]. Gaussian models are discussed in [EM84, EM85]. In [Mar02, BM03], with an assumption of super-Gaussian (Gamma, Laplace) models, MMSE estimates of the speech DFT coefficients are derived. To estimate the parameters of a speech model, different objectives such as maximum likelihood [MM80], maximum a posteriori (MAP) [WG01] and MMSE [EM84, EM85] are applied.

4.1.2 Wavelet shrinking estimator

Based on the mathematical properties of the wavelet transform explained in section 1.2 and section 1.3, David Donoho and other researchers have developed wavelet denoising methods using the thresholding/shrinking technique which is based on optimal linear and non-linear estimates. A brief summary of wavelet shrinkage denoising is reported in [Tas00]. The principle of removing noise is based on thresholding/shrinking wavelet coefficients towards zero. These techniques have been widely used in signal processing, image compression and image denoising. Recently, wavelet thresholding/shrinking is applied for speech processing in order to enhance speech quality.

Two thresholding functions are considered by Donoho et al. [Don95, DJ98] as hard thresholding and soft thresholding. These thresholding functions are obtained by applying optimal linear estimation technique. Modifications of the thresholding functions are studied in [ZD97, ZL99] with an adaptive higher order derivatives shrinking function. A so-called wavelet firm shrinkage which generalizes hard and soft thresholding is proposed in [BG95, BG96, BG97]. All of these variants try to get smoother shrinking functions to denoise effectively while preserving more useful information of the noise-free signal. The noise threshold is estimated in the RiskShrink approach with the minimax threshold [DJ98], in the VisuShrink approach with the universal threshold [DJ94], and in the SureShrink approach with the SURE (Stein's unbiased risk estimator) threshold [DJ95].

Many improvements of wavelet thresholding techniques for enhancing noisy speech signals have been studied such as semisoft thresholding with selected threshold for unvoiced regions [SB97], efficient hard and soft thresholdings [SMM02], and a smooth hard thresholding function based on μ-law [SA01, CKYK02]. In [LGG04], the combination of soft and hard thresholding is adapted to different properties of the speech signal. The more sophisticated shrinking functions with better characteristics than soft and hard thresholdings are developed for speech recognition in [KKH03]. To obtain lower speech distortion, wavelet thresholding/shrinking method are integrated with other techniques such as Teager energy operator and masked adaptive threshold [BR01]. A blind adaptive filter of speech from noise is designed in wavelet domain [VG03].

Motivated by the classical wavelet shrinkage pioneered by Donoho and his colleagues, optimal non-linear estimates have been applied in Bayesian-based wavelet shrinking methods. Contributions to the development of Bayesian approaches to wavelet shrinkage are presented in [Vid98, ASS98]. Bayes' rule is applied to achieve non-linear thresholding. Adaptive Bayesian wavelet shrinkage is proposed in [CKM97] which is applied for image denoising with promising performance [SA96, JB99]. As a development from [Vid98], the local false discovery rate is linked with Bayes factor shrinkage to form a novel shrinking method called Bayesian adaptive multiscale shrinkage [Lav06]. A very good overview of noise reduction by non-linear wavelet shrinking derived from optimal linear and non-linear estimates can be found in [Jan00]. Application of Bayesian-based wavelet shrinkage into speech enhancement is studied in [KYK01]. Thresholds are estimated by minimizing the Bayesian risk.

In this chapter, a novel speech enhancement system based on statistical wavelet shrinking is designed to eliminate musical noise as well as handle colored and non-stationary noise. The denoising process is considered in a sequence of buffers consisting of several overlapping speech frames. The wavelet coefficients of 128 channels are extracted by performing the full WPD on every speech frame. For every wavelet channel of all frames in the buffer, the universal thresholds [DJ94] are calculated. Then quantile filtering method [SFB00, PK05b] is applied to statistically estimate more accurate thresholds relating to the noise levels. This procedure is done recursively for the sorted universal thresholds along each wavelet channel. Next, to deal with non-stationary noise, adaptive weighting functions are built based on the temporal threshold variation. In addition, colored noise is handled by another static non-linear weighting function. The wavelet shrinking gain function proposed in [SA01, CKYK02] is optimized by an adaptive factor for listening comfort. The designed statistical wavelet shrinking algorithm is evaluated on the AURORA3 and NTIMIT databases. Besides objective tests measuring the segmental SNRs, subjective tests such as an overall quality evaluation based on Comparison Category Rating (CCR) and based on ITU-T standards are applied. As a new contribution to subjective testing methods, a Comparison Diagnostic Test (CDT) is developed from the Diagnostic Rhyme Test standard to improve the evaluation quality for speech enhancement system.

The chapter continues with the following section for reviewing and discussing state-of-the-art speech enhancement methods in more depth. Then, the wavelet denoising approach is treated with different wavelet gain functions for removing noise. The three next sections present in detail all steps for noise threshold estimation by scale-dependent universal threshold, quantile threshold, and time-frequency weighted threshold, respectively. In the evaluation section, all designed methods are assessed by an objective test and subjective tests on different databases. A new evaluation method is designed in this section. Finally, the conclusion section reviews the designed system and presents an outlook for future work to improve the speech enhancement method.

4.2 State-of-the-art speech enhancement

From modeling point of view, the speech signal can be distorted by additive noise from the environment or convolutive noise created by the (linear) transfer function of the communication channel. Additive noise is considered in most noisy signal models although it is not always the case. Its model is linear in the power spectral domain while the model of convolutive noise is linear in the log-spectral or cepstral domain. In this thesis, we are interested only in the additive noise model as follows:

$$y(n) = x(n) + d(n) \ . \tag{4.1}$$

where a noisy speech signal $y(n)$ results from a clean speech signal $x(n)$ which is corrupted by the additive noise signal $d(n)$. Noise signals are considered as undesired. They can be stationary, non-stationary, narrowband, and broadband noise. Even interfering speech signals, such as multi-talker babble or a single concurrent speaker, can be classified as background noise. Noise can originate from a localized position or from virtually all directions which is called diffuse noise. The additive model can also be presented in the frequency domain as:

$$Y_w(k) = X_w(k) + D_w(k). \tag{4.2}$$

where $Y_w(k)$, $X_w(k)$ and $D_w(k)$ denote the short-time spectra of the time-domain signals $y_w(n)$, $x_w(n)$ and $d_w(n)$ which are obtained by multiplying a sliding time-limited window w with the signals. k is the discrete frequency index, $k = 0, 1, ..., K-1$. Since the speech signal and the noise signal are uncorrelated, the following relation holds, too:

$$P_y(k) = P_x(k) + P_d(k). \tag{4.3}$$

where $P_y(k)$, $P_x(k)$ and $P_d(k)$ are the power spectral densities (PSD) of the time-domain signals $y(n)$, $x(n)$, and $d(n)$, respectively. For stationary signals, as $K \to \infty$, the PSD can be defined as $P_x(k) = E\{|X_w(k)|^2\}$. Based on these additive models, different suppression rules based on different optimal estimators have been proposed. The a priori and a posteriori SNR which are used to build these rules are introduced after McAulay and Malpass [MM80], Ephraim and Malah [EM84] as follows:

$$\xi_k \triangleq \frac{P_x(k)}{P_d(k)}, \tag{4.4}$$

$$\gamma_k \triangleq \frac{|Y_w(k)|^2}{P_d(k)}. \tag{4.5}$$

The differences and similarities between various speech enhancement approaches will be discussed in the following sections.

4.2.1 Spectral subtraction and its variants

The most common approach is spectral subtraction (SS) [Bol79] in the frequency domain. The denoising process is performed by subtracting an average estimate of the noise spectrum $\overline{|D_w(k)|}$ from the observed noisy speech spectrum $|Y_w(k)|$ as follows:

$$|\widehat{X}_w(k)| = \max\left(|Y_w(k)| - \overline{|D_w(k)|}, 0\right). \tag{4.6}$$

In its variant, so-called power spectral subtraction [LO79], an estimate of the short-time energy spectrum of the speech signal is calculated as:

$$|\widehat{X}_w(k)|^2 = \max\left(|Y_w(k)|^2 - E\{|D_w(k)|^2\}, 0\right). \tag{4.7}$$

From the observed data $y_w(n)$, the short-time energy spectrum $|Y_w(k)|^2$ of noisy speech signal is calculated directly. The noise level $E\{|D_w(k)|^2\}$ is estimated by averaging $|D_w(k)|^2$ over many non-speech frames where the background noise is assumed to be stationary. Negative values resulting from spectral subtraction are replaced by zero. This process results in a well known artifact so-called “musical noise” or “running spring water” which yields residual noise with a very unnatural and disturbing quality. Strong fluctuations may appear in the enhanced speech signal after denoising by the spectral subtraction method. The reason is that a succession of randomly spaced spectral peaks emerges in the frequency bands. This results in the residual noise which is composed of narrow-band components located at random frequencies that turn on and off randomly in each short-time frame [Cap94]. This “musical noise” artifact can be reduced by an improved estimation of the average noise spectrum. A generalization of the SS method is proposed in [LO79, BSM79] as follows:

$$|\widehat{X}_w(k)|^p = \max\left(|Y_w(k)|^p - \alpha E\{|D_w(k)|^p\}, 0\right). \tag{4.8}$$

And a frequency response of the enhancement system or a suppression rule is represented as a function of the a posteriori SNR γ_k as:

$$H_{SS}(k) = \frac{\widehat{X}_w(k)}{|Y_w(k)|} = \left(1 - \frac{\alpha}{\gamma_k}\right)^{\frac{1}{p}}. \tag{4.9}$$

A degree of freedom is described by the exponent p where $p = 1$ and $p = 2$ correspond to magnitude and power spectral subtraction, respectively. Noise overestimation, i.e. $\alpha \geq 1$, helps to remove noise as much as possible while increasing speech distortion. In case of underestimation, $0 < \alpha < 1$, the subtraction of noise is reduced to avoid the introduction of artifacts, and to keep the background noise at a certain level which maybe comfortable for the human ear.

4.2.2 Wiener filtering

The goal of the Wiener filter is to filter out noise that has corrupted a signal by statistical means. With this technique, a frequency weighting for an optimum filter is first estimated from the noisy speech. This filter is then applied to obtain an estimate of the noise-free speech signal. Wiener's solution requires information regarding the spectral content of the clean speech signal and the noise. Based on assumption that the speech signal and the noise signal are uncorrelated stationary random processes, the linear estimator of $x(n)$ which minimizes the mean square error is obtained by filtering $y(n)$ with the non-causal Wiener filter [LO79]:

$$H(k) = \frac{E\{|X_w(k)|^2\}}{E\{|X_w(k)|^2\} + E\{|D_w(k)|^2\}}. \tag{4.10}$$

This solution does not modify the phase of the noisy speech signal, it has zero phase. The phase associated with the estimate $\widehat{X}_w(k)$ is that of $Y_w(k)$. The approximated PSD $E\{|X_w(k)|^2\}$ of the clean speech signal can be estimated by first estimating $E\{|Y_w(k)|^2\}$, then $E\{|D_w(k)|^2\}$ is subtracted from the estimated $E\{|Y_w(k)|^2\}$ to obtain an estimate of $E\{|X_w(k)|^2\}$. A generalization of the Wiener filtering has been studied as parametric Wiener filters:

$$H_{WF}(k) = \left[\frac{E\{|X_w(k)|^2\}}{E\{|X_w(k)|^2\} + \alpha E\{|D_w(k)|^2\}}\right]^{\beta}. \tag{4.11}$$

where α and β are constants which are used to tune the characteristics of the Wiener filter. The Wiener filter solution corresponds to $\alpha = \beta = 1$. For $\alpha = 1$ and $\beta = 1/2$, we obtain a formula for power spectral subtraction. This means the Wiener gain function is just simply the square of the suppression rule for the power spectral subtraction method. Its gain function is rewritten as a function of the a posteriori SNR γ_k as follows:

$$H_{WF}(k) = \left(1 - \frac{\alpha}{\gamma_k}\right)^{\beta}. \tag{4.12}$$

4.2.3 Optimal non-linear estimator

As another approach, the modification of the SS is proposed in [MM80]. The estimation problem is formulated by assuming the noise at each frequency channel is Gaussian and the resulting estimate is derived from the ML estimate:

$$|\widehat{X}_w(k)| = \frac{1}{2}|Y_w(k)| + \frac{1}{2}\left(|Y_w(k)|^2 - E\{|D_w(k)|^2\}\right)^{\frac{1}{2}}, \tag{4.13}$$

and a gain function is calculated as:

$$H_{ML}(k) = \frac{1}{2} + \frac{1}{2}\sqrt{1 - \frac{1}{\gamma_k}}. \tag{4.14}$$

By considering the fact that speech does not always appear in all recorded speech frames $y_w(n)$, a probabilistic factor for the present of speech is proposed as a function of $|Y_w(k)|$ [MM80].

The optimal non-linear spectral amplitude estimation based on the minimum mean square error criterion and its modification [EM83, EM84, EM85] achieves a significant noise reduction while reducing the "musical noise" and maintaining good speech quality. In [EM84] a combination between the MMSE amplitude estimate and the MMSE phase estimate improves the optimality of the resulting estimate. The MMSE amplitude estimate is derived as following gain function:

$$H_{MMSE}(k) = \frac{\sqrt{\pi \nu_k}}{2\gamma_k} \exp\left(\frac{-\nu_k}{2}\right) \left[(1+\nu_k) I_0(\frac{\nu_k}{2}) + \nu_k I_1(\frac{\nu_k}{2})\right], \tag{4.15}$$

where $I_0(.)$ and $I_1(.)$ are the modified Bessel functions of order zero and one, respectively, and ν_k is defined as:

$$\nu_k = \frac{\xi_k}{1+\xi_k}\lambda_k. \tag{4.16}$$

The performance of this constrained MMSE amplitude estimate was shown to be better than the ML estimate [MM80] using the same Gaussian model. The gain function in Equation (4.15) requires high computational effort due to Bessel functions. Simplifications of this suppression rule are proposed in [WG01] with three simpler rules corresponding to the joint maximum a posteriori (MAP) estimate of the amplitude and the phase, MAP estimate of the amplitude, and MMSE estimate of the spectral power [Sri05]. The third approach provides the best approximation to Equation (4.15) with the following suppression rule:

$$H_{SP}(k) = \sqrt{\frac{\xi_k}{\xi_k+1}\frac{\nu_k+1}{\nu_k}}. \tag{4.17}$$

By applying a MMSE log-spectral amplitude estimator which provides a better distortion measure than the MSE of the spectral amplitude only [EM85], the following gain function is obtained by minimizing $E\left\{\left(\log X_w(k) - \log \widehat{X}_w(k)\right)^2\right\}$:

$$H_{LSA}(k) = \frac{\xi_k}{\xi_k+1} \exp\left(\frac{1}{2}\int_{\nu_k}^{\infty} \frac{e^{(-t)}}{t} dt\right). \tag{4.18}$$

where $X_w(k)$ and $\widehat{X}_w(k)$ are the amplitudes of the k^{th} spectral component of the noise-free speech and estimated speech signals, respectively. As shown in [EM85], this suppression rule results in lower residual noise than the one derived by minimizing the MSE in the spectral domain. Motivated by the observation that the speech DFT coefficients are better modeled by a Gamma distribution, and by using a Gaussian or

Laplacian distribution for the noise coefficients, two non-linear MMSE estimators are derived in [Mar02] which reduce “musical noise” for low a-priori SNRs. With the same assumption, it was proved in [BM03] that the use of a Gaussian distribution for the noise results in “musical noise” while it is reduced by using Laplacian distribution. A nice overview of the MMSE estimation approach assuming different super-Gaussian models for speech and noise is presented in [Mar05].

4.2.4 Noise estimation

A good noise estimate may lead to high quality of the denoised speech signal [Chi00, Vas00, Dav02]. As a general method, the noise estimate can be calculated from non-speech frames which are detected previously by VAD [BJF94, BJF95, SS97b, TO00, CK01]. The soft-decision VAD overcomes some disadvantages of the binary VAD by considering the probability of non-speech/speech presence in specific segments [SS98]. A drawback of the VAD-based spectral subtraction approach is low temporal consistency if the unreliable speech/non-speech detection is occupied. Another drawback of VAD is that the noise estimate cannot be updated during speech periods which results in poor noise estimation. To overcomes the problems of VAD, the noise estimate can be updated during the speech activity regions by the minimum statistics method which is proposed in [Mar93, Mar01]. As the minimum is sensitive to outliers, a related method which is called quantile method is presented in [SFB00, Eva01]. The noise level is estimated by taking the temporal quantile instead of the minimum at each frequency bin. One more approach which provides a more accurate noise estimate uses trained statistical models [Eph92a, SSKon] to exploit the prior knowledge about noise. HMMs, GMMs, and codebooks are such statistical models that are trained on a specific database.

4.3 Wavelet denoising approach

While the approaches described so far are based on the short-time Fourier transform, the noise reduction topic has been studied and developed recently in the wavelet domain by using the wavelet transform and the wavelet packet decomposition. The short-time Fourier transform which is used in the spectral subtraction method has a limitation due to its fixed absolute bandwidth resolution in the time-frequency plane. The discrete wavelet transform which provides a fixed relative bandwidth in the time-frequency analysis can avoid that drawback [VK95, Mal99]. This advantage of the DWT results from its short basis functions which allow to analyze high-frequency signal components while the long ones are helpful for low-frequency signal components.

4.3.1 Hard and soft thresholding functions

To be compared with other denoising methods using optimal non-linear estimator reviewed in previous sections, wavelet denoising is considered as a non-parametric statistical estimation which does not need to estimate parameters of the model prior. Due to the linearity of the wavelet transform, the additivity of the model in Equation (4.1) is preserved. The wavelet coefficients of noisy speech $Y_k(n)$ can be expressed as the sum of the coefficients of clean speech $X_k(n)$ and noise $D_k(n)$ as:

$$Y_k(n) = X_k(n) + D_k(n)\ . \tag{4.19}$$

In this noise model, we assume noise coefficients is white and Gaussian distribution with zero mean and variance σ^2. The goal of wavelet thresholding/shrinking is, with selected T, to minimize the mean squared error or formally the risk $R(T) = E\left\{\|\widehat{X}_k(n) - X_k(n)\|^2\right\}$. The heuristic of the wavelet-based denoising technique originates from this statement: every empirical wavelet coefficient contributes noise of variance σ^2, but only very few wavelet coefficients contribute signal [DJ94]. As we know, the wavelet transform decorrelates the signal and leaves uncorelated noise uncorrelated. Because the noise is spread out equally over wavelet coefficients, i.e. the wavelet decomposition leads to a sparse representation, the important coefficients can be distinguished from noisy coefficients by their higher absolute values. This suggests a denoising strategy (so-called wavelet thresholding) by replacing the noisy DWT coefficients which are lower than a certain threshold T by zero and doing the inverse DWT to get a denoised signal. In case of very low SNRs, this technique can not hold because it is difficult to distinguish between signal coefficients and noisy coefficients.

With the hard-thresholding gain function $G^H(T)$ [DJ94], all wavelet coefficients which are lower than the threshold T are removed while the others are left untouched:

$$\widehat{X}_k^H(n) = G^H(T,Y) = \begin{cases} Y_k(n) & ,\text{if } |Y_k(n)| > T \\ 0 & ,\text{if } |Y_{k(n)}| \leq T \end{cases}\ . \tag{4.20}$$

Because of the strong discontinuity of this input-output characteristics as depicted in Figure 4.2, some applications do not use hard-thresholding. The pure noise coefficients may pass a threshold which results in artifacts introduced in the outputs such

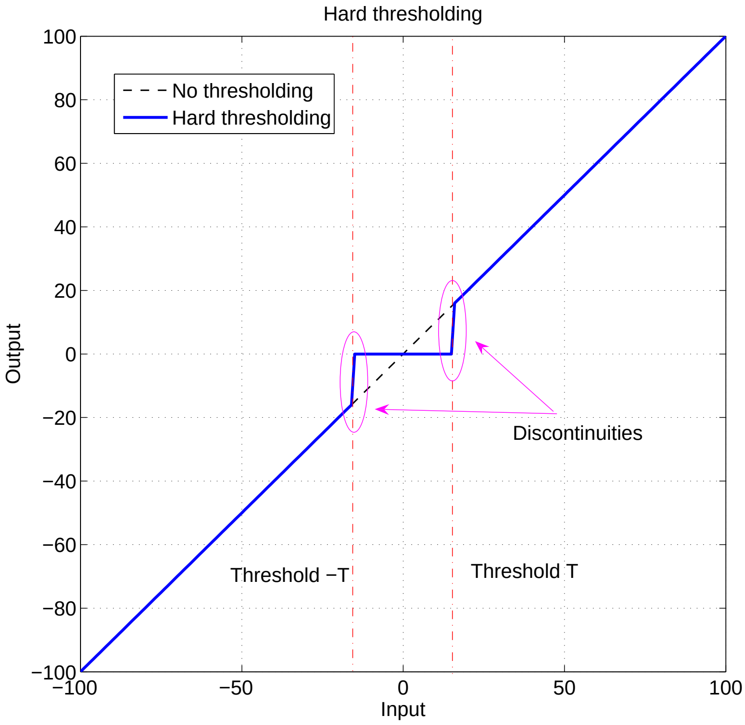

Figure 4.2: Input-output characteristic of hard-thresholding and its discontinuities.

as annoying 'blips' and increased residual signal distortion. With the soft-thresholding proposed in [DJ94, Don95], these artifacts can be reduced by shrinking the wavelet coefficients above the threshold by an amount equal to the absolute threshold value T:

$$\widehat{X}_k^S(n) = G^S(T,Y) = \begin{cases} \operatorname{sgn}\left(Y_k(n)\right)\left(|Y_k(n)| - T\right) & \text{, if } |Y_k(n)| > T \\ 0 & \text{, if } |Y_k(n)| \leq T \end{cases} \tag{4.21}$$

Consequently, the input-output characteristics becomes continuous, see Figure 4.3. It reduces the discontinuity of hard-thresholding but is still sub-optimal because its high-order derivatives are all zero and because of the strict setting to zero of the coefficients whose absolute values are below the threshold. This leads to the destruction of wavelet coefficients of unvoiced speech due the high similarity in terms of signal characteristics between unvoiced consonants and noise. By deriving bias, variance, and risk function for the hard thresholding and soft thresholding, Andrew Bruce in [BG95] analyzed the basic difference between hard and soft thresholding. Hard thresholding tends to have bigger variance because of the discontinuity of the gain function while soft thresholding tends to have bigger bias due to its shifting of all coefficients which are bigger than threshold T towards zero by an amount of T. Contribution of threshold

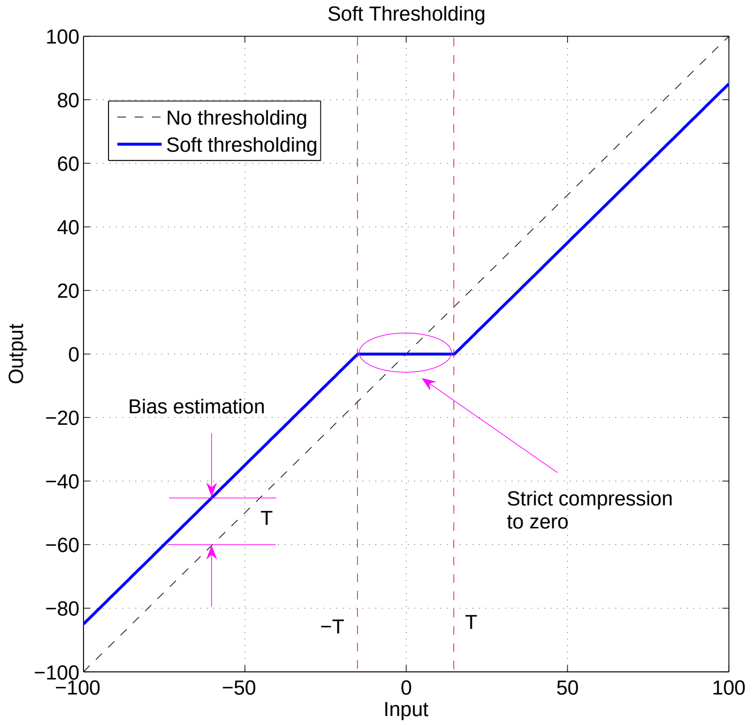

Figure 4.3: Input-output characteristics of soft-thresholding and strict setting to zero.

selection to bias and variance is discussed in next section.

4.3.2 Shrinking functions

The shrinking terminology is to be distinguished from thresholding. Thresholding means all wavelet coefficients whose values are below a selected threshold are strictly set to zero, while shrinking preserves more information of coefficients by setting them to a fraction of their original values, only a small part of them is replaced strictly by zero. As one of the methods for achieving small variance and small bias at once, a semisoft shrinkage function was proposed in [BG95]. It employs two thresholds T_1 and T_2, and represents a compromise between soft and hard thresholding and is defined as the following function:

$$\widehat{X}_k^{SS}(n) = G^{SS}(T_1, T_2, Y) = \begin{cases} 0 & ,\text{if } |Y_k(n)| \leq T_1 \\ \operatorname{sgn}(Y_k(n)) \dfrac{T_2(|Y_k(n)| - T_1)}{T_2 - T_1} & ,\text{if } T_1 < |Y_k(n)| \leq T_2 \\ Y_k(n) & ,\text{if } |Y_k(n)| > T_2 \end{cases} \tag{4.22}$$

The semisoft shrinkage function includes both hard thresholding ($T_1 = T_2$) and soft thresholding ($T_2 = \infty$) as special cases. This kind of generalized shrinkage can prevent the sensitivity to small fluctuation of the wavelet coefficients which is very critical in hard thresholding. Due to its high discontinuity, the hard thresholding output changes dramatically when the values of wavelet coefficients vary slightly around the threshold.

There are several variants of wavelet shrinkage which try to smoother the gain function. A differentiable shrinkage function presented in Equation (4.23) was studied in [ZD97, ZL99]. These shrinking functions have higher order derivatives which allow to develop adaptive schemes of denoising based on gradients and will have better numerical properties in general.

$$\widehat{X}_k^{DS}(n) = G^{DS}(T,\rho,Y) = \begin{cases} Y_k(n) + T - \dfrac{T}{2\rho+1} & \text{, if } Y_k(n) \leq -T \\ \dfrac{1}{(2\rho+1)T^{2\rho}} Y_k^{2\rho+1}(n) & \text{, if } |Y_k(n)| \leq T \\ Y_k(n) - T + \dfrac{T}{2\rho+1} & \text{, if } Y_k(n) > T \end{cases} \tag{4.23}$$

Motivated by a novel shrinking function which is continuous around the threshold and adapts to the characteristics of the input signal, a customized shrinking function was designed in [YV04] as follows:

$$\begin{aligned} \widehat{X}_k^{CS}(n) &= G^{CS}(T,\tau,\epsilon,Y) \\ &= \begin{cases} Y_k(n) - \operatorname{sgn}(Y_k(n))(1-\tau)T & \text{, if } |Y_k(n)| \geq T \\ 0 & \text{, if } |Y_k(n)| \leq \epsilon \\ \tau \left(\dfrac{|Y_k(n)|-\epsilon}{T-\epsilon}\right)^2 \left[(\tau-3)\left(\dfrac{|Y_k(n)|-\epsilon}{T-\epsilon}\right) + 4 - \tau\right] & \text{, otherwise} \end{cases} \end{aligned} \tag{4.24}$$

where $0 < \epsilon < T$ is the cut-off values below which the wavelet coefficients are set to zero, and $0 < \tau < 1$ is a parameter controlling the function shape. With $\tau = 0$ we obtain the soft thresholding function and with $\tau = 1$ we get a smoothed hard thresholding function which is discussed in the next section.

4.3.3 Optimal shrinkage: Smoothed hard thresholding

An enhanced shrinkage, i.e. smoothed hard thresholding based on the μ-law, is proposed in [SA01, CKYK02] for speech enhancement. It preserves the larger coefficients and has a smooth transition from noisy coefficients to signal coefficients

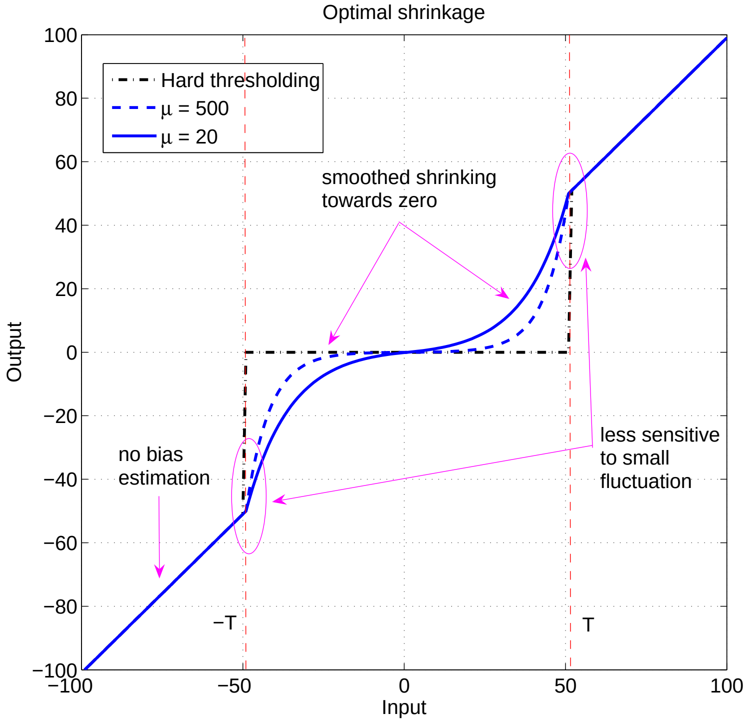

Figure 4.4: Enhanced wavelet shrinking by smoothed hard-thresholding.

$$\tilde{X}_k^{SH}(n) = G^{SH}(T, \mu, Y)$$

$$= \begin{cases} Y_k(n) & , \text{if } |Y_k(n)| > T \\ \dfrac{T \ \text{sgn}(Y_k(n))}{\mu} \left((1+\mu)^{\dfrac{|Y_k(n)|}{T}} - 1 \right) & , \text{if } |Y_k(n)| \leq T \end{cases} \tag{4.25}$$

This suppression rule is further modified by [Kot04] for denoising in robust speech recognition with an adaptive parameter μ_k which is defined as:

$$\mu_k = \theta \frac{\max_n \{|X_k[n]|\}}{T} , \tag{4.26}$$

where θ is a constant factor selected to minimize the MSE between the noise-free speech signal $x[n]$ and its estimate $\widehat{x}[n]$ [Kot04]. In section 4.6.3, an adaptive factor θ will be proposed to improve perceptual speech quality. The μ-law shrinkage as well as the semisoft shrinkage [BG96] present a compromise between soft and hard thresholding. As discussed, the hard thresholding function has larger variance but smaller bias, while the soft thresholding shows higher bias but smaller variance. In other words, hard thresholding tends to keep closeness to the signal, while soft thresholding achieves smoothness of the signal [BG96]. A big advantage of the μ-law shrinkage over others is that it does not strictly set to zeros ll or parts of the wavelet coefficients, whose absolute values are below the threshold, as done by hard and soft thresholding [DJ94]. Besides, the drawback of the semisoft shrinkage [BG96] (higher computational complexity with two thresholds) is avoided here.

Fig. 4.4 shows the input-output characteristic of the μ-law based optimal shrinkage function. The lower the adaptive parameter, the more wavelet coefficients are retained and vice verse. The smoothed hard-thresholding turns into conventional hard thresholding when the adaptive parameter goes to infinity. This shrinking preserves more wavelet coefficients which fall below the threshold. This behavior of saving coefficients is demonstrated in Figure 4.5 and Figure 4.6, with wavelet coefficients derived from a 3^{rd} scale DWT of an unvoiced fricative segment at SNR $=$ 15dB, and their versions processed by hard thresholding and smoothed hard thresholding.

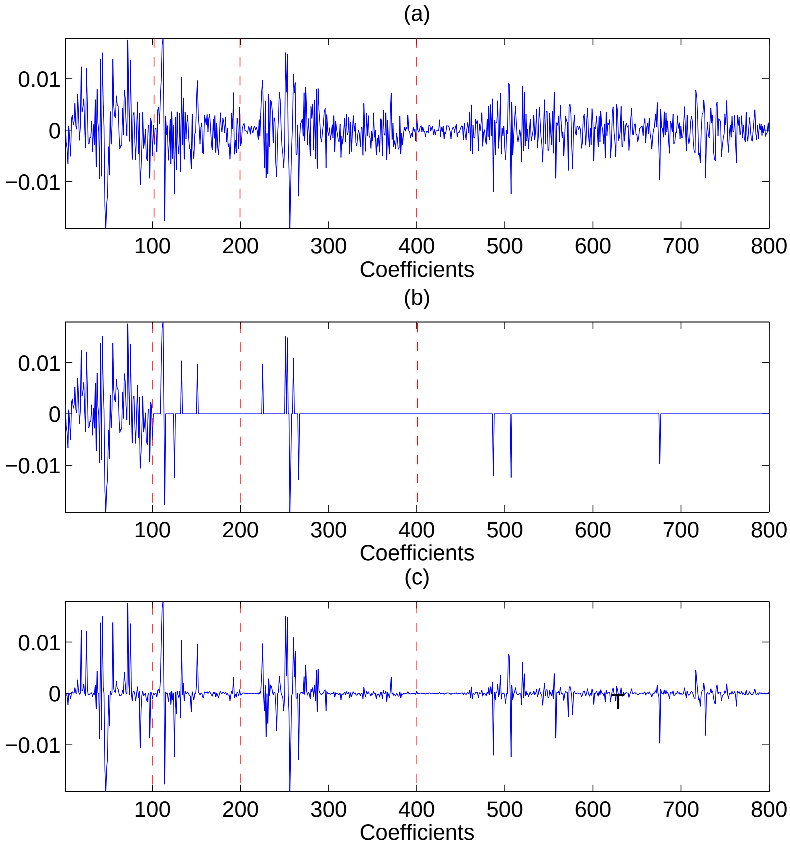

Figure 4.5: (a) Wavelet coefficients derived from an unvoiced segment at SNR $=$ 15dB, (b) Denoising by hard-thresholding, (c) Denoising smoothed hard-thresholding.

Another advantage of the smoothed hard thresholding is that it maintains the structure of voiced speech better than hard thresholding for high noise level. As shown in

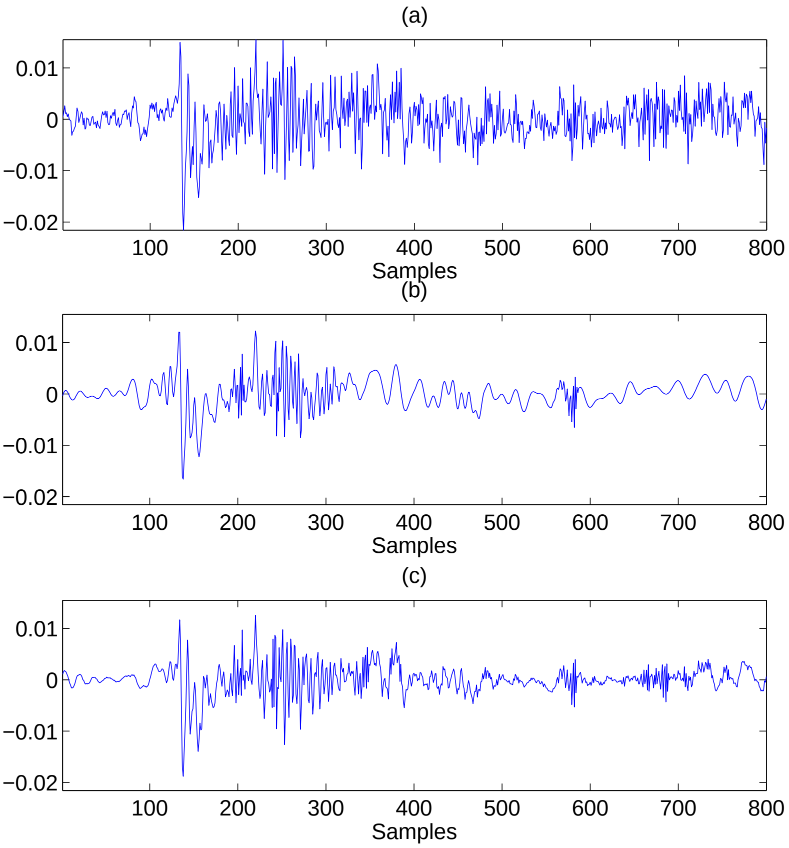

Figure 4.6: (a) A fricative segment with SNR = 15dB in the time domain, (b) Its enhancement by hard-thresholding, (c) Its enhancement by smoothed hard-thresholding.

Figure 4.7, the almost periodic structure of the reconstructed waveform of the phoneme $/e/$ which is corrupted by car noise is maintained better after denoising by smoothed hard-thresholding than the one using hard-thresholding. The introduced artifacts in Figure 4.7 (c) may come from the unsuitable selection of the wavelet basis.

4.4 Estimate of thresholds

4.4.1 Bias and variance

We consider the contributions of bias and variance into risk function and focus on the threshold that minimizes this objective function. The content is this section is adapted from original material of chapter 3 in [Jan00]. Thresholding can be considered a trade-off between data fitting and smoothing. A small threshold achieves a close fitting to input speech with high probability of retained noise. On the other hand, a

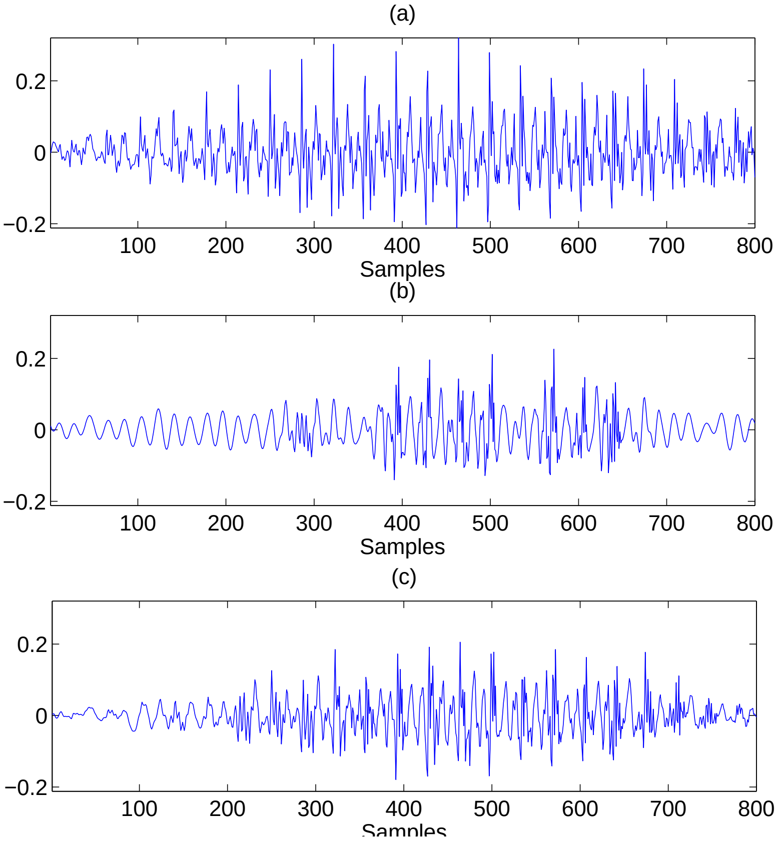

Figure 4.7: (a) A vowel segment with SNR = 15dB, (b) Its enhancement by hard-thresholding, (c) Its enhancement by smoothed hard-thresholding.

large threshold shrinks much more wavelet coefficients to zero so as to smooth the input signal while degrading the speech characteristics by introducing distortions, especially for noise-like, i.e. unvoiced sounds. The expected risk function is presented as a function consisting of bias and variance terms:

$$\begin{aligned} E\left\{R(T)\right\} &= E\left\{\|E\{\widehat{X}_k(n)\} - X_k(n)\|^2\right\} + E\left\{\|\widehat{X}_k(n) - E\{\widehat{X}_k(n)\}\|^2\right\}, \\ Risk &= Bias + Variance. \end{aligned} \tag{4.27}$$

Obviously replacing the smallest coefficients by zero lowers the variance but increases the bias, and vice versa. By increasing the threshold, we get smaller variance which produces a smoother signal, while decreasing the threshold corresponds to a lower bias which represents a closer fitting. The minimum risk threshold is the best compromise between variance and bias. The contribution of each individual coefficients to the risk function affects the threshold estimation [DJ94]. Assuming the noise source is Gaussian, an important conclusion is derived: the small coefficients with less information are best thresholded by high thresholds, whereas the big coefficient containing useful information should be served with small thresholds [Jan00].

4.4.2 Universal threshold procedure

The universal threshold derived by Donoho [Don95] is based on the principle of minimization of the risk between the expected signal and the noise-reduced output signal. He proved that, under assumptions such as orthogonality of DWT and i.i.d. noise with variance σ^2, the universal threshold is proportional to the standard deviation and the log of the length N of the wavelet coefficient sequence:

$$T = \sigma\sqrt{2\log N} \tag{4.28}$$

This threshold which is derived for the asymptotic behavior of the minimum risk is not actually the optimal threshold vale. However, it is applicable for any signal with length N and sufficiently smooth [Jan00].

4.4.3 White and non-white noise

The wavelet decomposition provides not only a sparse representation but also a multiscale representation. The later property is an important wavelet characteristic which helps to choose the threshold in case of colored noise where the noise power depends on the scale. We see that the noise level is equal over all scales only for the white noise case. Thus, the noise threshold can be estimated from the detail coefficients at the 1^{st} analysis scale because this detail contains most of the small coefficients. However, in case of colored noise, the noise level must be assessed at each scale, and the variance at that scale is re-calculated. Scale-dependent threshold T_k was proposed by [Joh99]. It is more adaptive to speech characteristic.

$$T_k = \sigma_k\sqrt{2\log N_k} \tag{4.29}$$

where N_k is the length of one wavelet packet containing the sequence of wavelet coefficients $Y_k(n)$ in the k^{th} frequency channel. σ_k is the standard deviation of the noise in $Y_k(n)$. To avoid outliers, the noise level is estimated by the median absolute deviation (MAD) of the sequence of wavelet packet coefficients which is a robust estimate of the standard deviation:

$$T_k = \frac{1}{\gamma_{MAD}}\text{Median}(|Y_k(n)|)\sqrt{2\log N_k}\ , \tag{4.30}$$

where $\gamma_{MAD} = 0.6745$ is the conversion factor between the standard deviation and MAD in case of white Gaussian noise [Don95]. In our research, the universal threshold and the smoothed hard shrinking function are used to estimate the threshold and removing noise in case of stationary and non-white noise. Because the speech enhancement system processes the speech signal frame by frame, the frame index i is introduced into the formula of the shrinking function as:

$$\begin{aligned}\tilde{X}^{SH}_{k,i}(n) &= G^{SH}(T,\mu,Y)\\ &= \begin{cases} Y_{k,i}(n) & ,\text{if } |Y_{k,i}(n)| > T_{k,i} \\ \dfrac{T_{k,i}\ \mathrm{sgn}(Y_{k,i}(n))}{\mu}\left((1+\mu_{k,i})^{\dfrac{|Y_{k,i}(n)|}{T_{k,i}}} - 1\right) & ,\text{if } |Y_{k,i}(n)| \le T_{k,i} \end{cases}\end{aligned} \tag{4.31}$$

where the adaptive parameter $\mu_{k,i}$ is defined as:

$$\mu_{k,i} = \theta \frac{\max\limits_{n}\{|Y_{k,i}[n]|\}}{T_{k,i}} \ , \tag{4.32}$$

The universal threshold calculation is applied for every k^{th} frequency channel at the i^{th} frame as:

$$T_{k,i} = \frac{1}{\gamma_{MAD}}\mathrm{Median}(|Y_{k,i}(n)|)\sqrt{2\log N_{k,i}} \ , \tag{4.33}$$

From the analysis of the universal threshold approach, we conclude that this is a very local estimate. It does not take into account the correlations between different coefficients across scales. This consideration is necessary for denoising speech signals contaminated by non-white noise which has different characteristics at different frequency channels. This drawback is surmounted by a nonlinear frequency weighting function proposed in subsection 4.6.1. Besides, the universal thresholds are calculated from the wavelet packets of every frequency channel at a certain frame. This means the temporal characteristics of speech and noise are not accounted for the estimation of the threshold, especially for non-stationary noise. In order to handle the addressed problem, the calculated universal thresholds are further modified by quantile filtering over a buffer containing N_f speech frames. The thresholds are updated for every buffer by a recursive scheme in section 4.5.1 and temporally weighted by an adaptive function in section 4.6.2.

4.5 Statistical quantile filtering

It is well known that speech information does not always appear in all frequency channels simultaneously, even in speech intervals. The energy in each frequency channel is on the noise level over a significant part of the time as reported in [SFB00]. Thus, the noise level can be estimated by taking the q^{th} quantile observed over the duration of the utterance in every channel. From this observation, we develop a quantile-based algorithm to estimate the thresholds related to the noise level in each wavelet packet channel.

4.5.1 Quantile filtering and recursive buffer

To meet the memory saving and delay requirements, overlapping buffers instead of the whole utterance (which might continue indefinitely) are used. Because of the properties of quantile method, the buffer length and its effect on the perceptual quality should be considered. Some settings of these parameters are described in Table 4.1. The buffers (L_b = 960ms length and 480ms overlap) which store N_f = 47 speech frames (L_f = 40ms length and 20ms overlap) are used because of the good results from informal listening tests and their suitability to real-time applications. After sorting the threshold values stored in a buffer for every wavelet packet channel, we observe that the threshold values derived from non-speech frames occupy up to 60% of the buffer as depicted in Fig. 4.8. This lower quantile part should be memorized and transferred into the next buffer to maintain continuity with the estimated noise level of the previous buffer. This is done by applying a recursive scheme as described in Fig. 4.9 which is constructed from the overlapping buffers. The use of recursive buffering mechanism helps to track non-stationary noise because the thresholds related to noise levels are updated for every buffer automatically.

$L_b(ms)$	$N_f(frames)$	$L_f(ms)$
480	23	40
960	47	40

Table 4.1: The setting of frame and buffer sizes

4.5.2 Noise threshold estimation

Noise threshold estimation is implemented in three steps as follows:

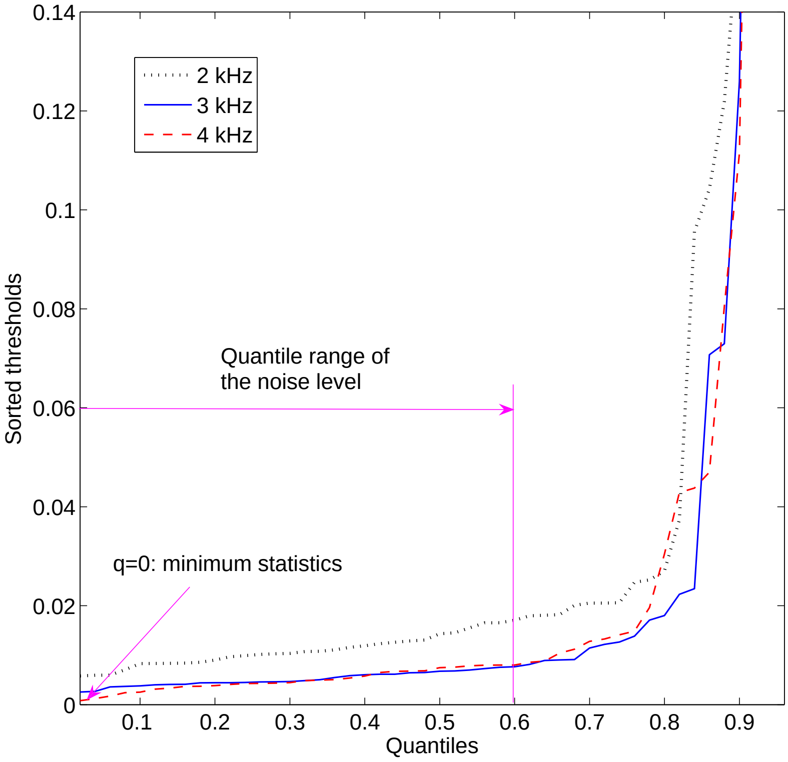

Figure 4.8: Quantiles of sorted threshold values over a buffer of 960ms at three selected wavelet subbands.

- First, the threshold $T_{k,i}$ are calculated for all wavelet packet channels of all frames from the current buffer (e.g., the 2^{nd} buffer in Fig. 4.9) using Equation 4.33, and stored in the corresponding threshold buffer.

- Second, the new thresholds $T_{k,i}$ of the frames $i = 24, \ldots, 47$ are selected and merged with the sorted thresholds $T_{k,i}$ which are selected from quantile range $q = 0.1, \ldots, 0.6$ of the previous sorted threshold buffer to form a recursive buffer. The selection of this quantile range guarantees the 'fading memory' property for the recursive buffering scheme.

- Then, for each WPD channel, the thresholds of all speech frames in the recursive buffer are sorted in ascending order which leads to $T_{k,i'}$, where $i' = 1, \ldots, N_f$ are the frame indices after sorting with $N_f = 47$. This sorted recursive buffer will be used in the next loop.

- Finally, the threshold related to the noise level, the so-called quantile threshold Γ_k, for all frames in the sorted recursive buffer at the k^{th} channel, is determined as the

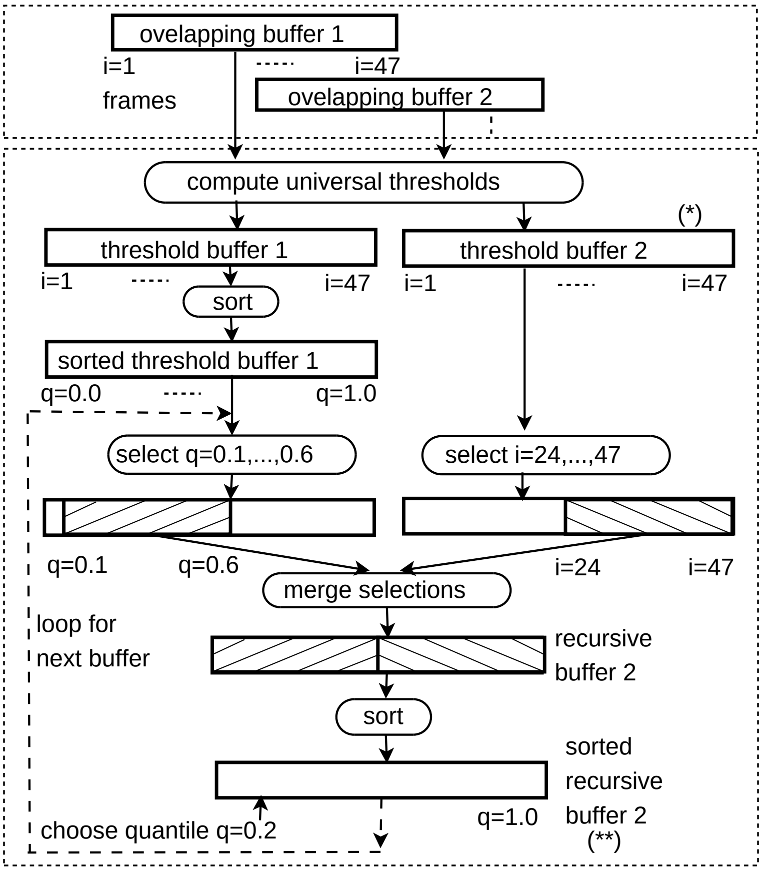

Figure 4.9: Recursive scheme.

q^{th} quantile:

$$\Gamma_k = T_{k,i'} \quad |_{i'=\lfloor qN_f \rfloor} \tag{4.34}$$

We have performed informal listening tests over the range of possible values $q = 0.0,\ 0.1, \ldots,\ 0.6$. With $q = 0$, we get the minimum statistics estimation of the noise threshold as in [Mar01]. There is, of course, an obvious question about the optimal selection of q. Some research in [SFB00] tried to compare the use of mean, modal, and median of quantiles in spectral subtraction based speech recognition. However, quantile selection in the wavelet threshold domain is still an open question for speech enhancement. The quantile $q = 0.2$ is a good choice which yields the best performance in informal listening test.

4.6 Time-frequency weighted quantile threshold

As mentioned in section 4.4.3, in order to handle the non-stationary and non-white noise effectively, the quantile threshold Γ_k derived for every k^{th} channel at each i^{th} frame in a certain buffer is weighted by a time-frequency dependent non-linear adaptive function as follows :

$$\tilde{\Gamma}_{k,i} = \lambda_{k,i}\eta_k\Gamma_k, \tag{4.35}$$

where $\lambda_{k,i}$, η_k are nonlinear parameters in the time and frequency domains, respectively. $\tilde{\Gamma}_{k,i}$ is the weighted estimate of the quantile threshold.

4.6.1 Non-linear frequency weighting

Correlated or colored noise has high variation of noise energy over the various wavelet channels. As reported in [Jan00], by the analysis of the correlation matrix of the noisy wavelet coefficients, if the noise is not white, then the colored noise is maintained after the wavelet transform. To suppress colored noise effectively, the quantile threshold Γ_k is weighted by the static non-linear function η_k as follows:

$$\eta_k = (a_1\Gamma_k)^{-b_1} + d_1 \ , \tag{4.36}$$

where $a_1 = 10, b_1 = 0.4, d_1 = 0.2$ are constants and selected manually to obtain good perceptual quality from our informal listening tests.

By employing η_k, a better estimate of the noise level thresholds over all wavelet channels is achieved, thereby the quality of the processed speech is enhanced. As shown in Fig. 4.10, the weighting function η_k amplifies strongly the small quantile thresholds and less those in the larger range. With this behavior, the weighting function meets the requirements discussed in section 4.4.1. In general, the small coefficients which lead to small universal thresholds, and thus small quantile thresholds (in the range of $[0.0, \ldots, 0.1]$ depicted in Fig. 4.10) should be best thresholded with high thresholds. Whereas the big coefficients which lead to large universal thresholds, and thus large quantile thresholds (which are higher than 0.1) should be served by less amplified thresholds. In other words, by applying frequency weighting in Equ. 4.36, the big coefficients containing speech information mostly at low-frequency channels result in low thresholds, and thus are slightly impacted. This leads to a low distortion of the enhanced speech signal while removing background noise effectively. In [PK05b], there was another proposed function which provides quite similar behavior. The only difference is that the very low threshold range of $[0.0, \ldots, 0.05]$ should not be thresholded by too high a threshold in order to make the remaining background noise sound comfortable. However this function is more expensive than the function in Equation (4.36). Besides, its usage is effective only when dealing with white noise which is not always realistic. For other types of background noise, the quality of speech sounds enhanced

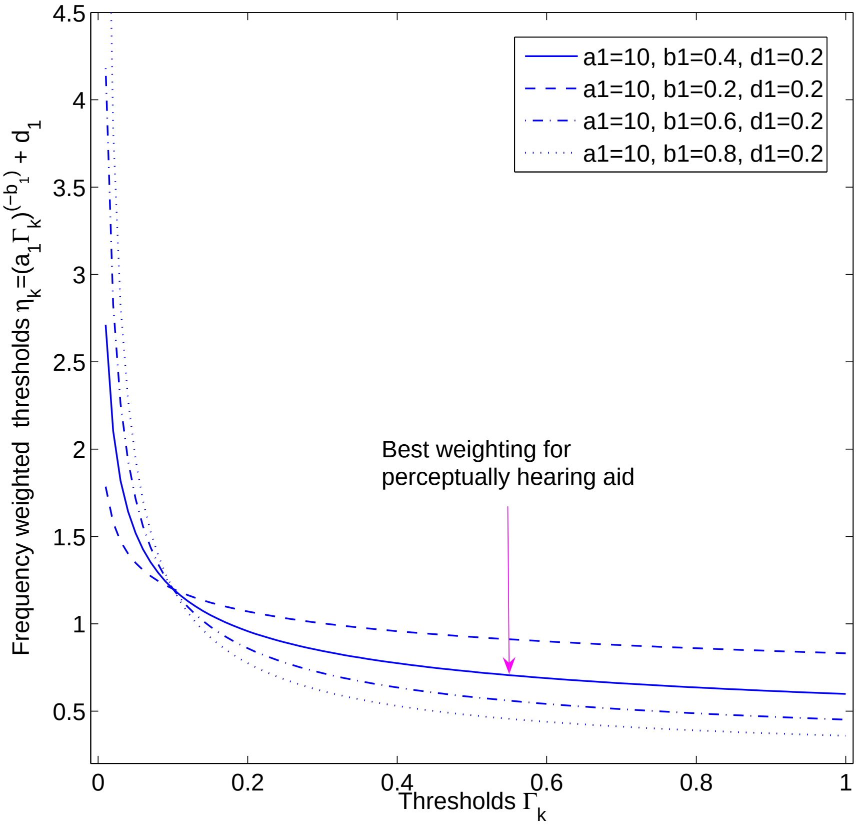

Figure 4.10: Frequency non-linear weighting function.

by these two functions is very similar. Design of this function is reported in appendix E.1.

4.6.2 Adaptive temporal weighting

Non-stationary noise can be handled by adaptive weighting method in the time domain. In principle, the method is based on the observed temporal variation of the universal threshold values of all frames over a buffer. The time-varying threshold dependent curve (TDC) $\lambda_{k,i}^{TDC}$ is built to track where speech and noise appear along the buffer. The frames with smaller estimated thresholds $T_{k,i}$, which are obtained from quantile filtering, might correspond to noise and will undergo stronger thresholding. The frames with large $T_{k,i}$ values always contain more speech information and are treated in the reverse way to preserve speech quality, see Fig. 4.11. This is expressed as:

$$\lambda_{k,i}^{TDC} = (a_2 T_{k,i}^*)^{-b_2} + d_2 \ , \qquad (4.37)$$

where $a_2 = 1, b_2 = 0.14, d_2 = 0.2$ are selected experimentally from informal listening tests, $T^*_{k,i}$ is the universal threshold in the current threshold buffer (step (*) in Fig. 4.9). By introducing this parameter $\lambda^{TDC}_{k,i}$, the weighted threshold $\tilde{\Gamma}_{k,i}$ in Equation 4.35 becomes frame-dependent threshold.

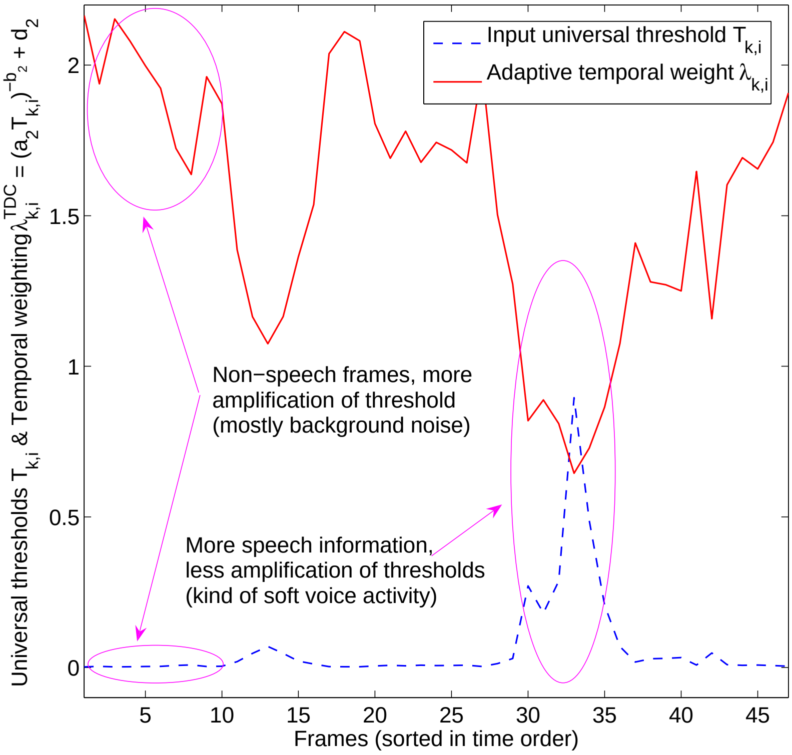

Figure 4.11: Input universal threshold $T_{k,i}$ and adaptive temporal weighting $\lambda^{TDC}_{k,i}$.

After the nonlinear weighting by the time-frequency parameter, the weighted thresholds $\tilde{\Gamma}_{k,i}$ are smoothed by median filtering to reduce the fluctuations between estimated noise thresholds of neighboring frames. Figures 4.12 - 4.14 shows noise thresholds estimation in the time-frequency plane by applying frequency weighting, time and frequency weighting, and smoothed time and frequency weighting on the calculated universal thresholds.

4.6.3 Adaptive factor for smoothed hard shrinking

From our initial listening tests, we felt that the processed speech sound is not as good as natural speech if we keep the factor θ constant. A high value of θ eliminates the noise effectively with a rather quiet residual in the non-speech frames. But it also creates

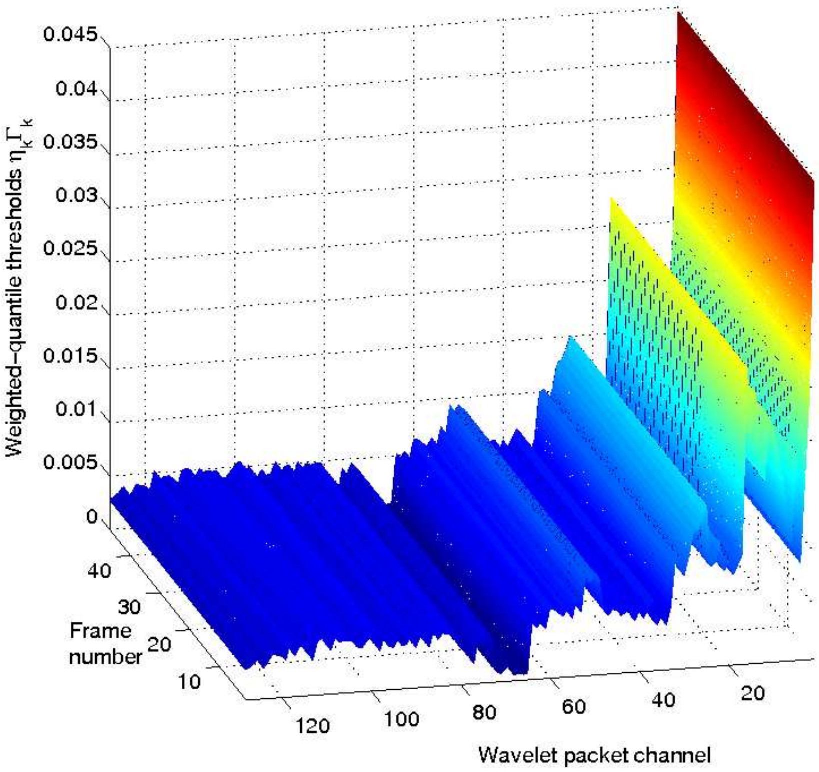

Figure 4.12: Frequency weighting over all channels in a buffer.

some discontinuities in the speech frames that result in speech distortion artifacts. To overcome this phenomenon, the factor θ is adapted itself by the normalized smoothed thresholds $\tilde{\Gamma}_k$:

$$\theta_{k,i} = \exp\left(\alpha \frac{\tilde{\Gamma}_{k,i}}{\max_i\{\tilde{\Gamma}_{k,i}\}}\right) \tag{4.38}$$

where $\alpha = 5.8$ is a slope constant. Due to this adaptive factor $\theta_{k,i}$, at each wavelet packet channel, the part of the curve below the threshold in Fig. 4.4 is flatter for the speech frames so as to preserve more coefficients and steeper for the non-speech frames so as to compress noise. This improves the perceptual quality for the speech parts significantly and keeps the background noise at a very small level for the non-speech parts.

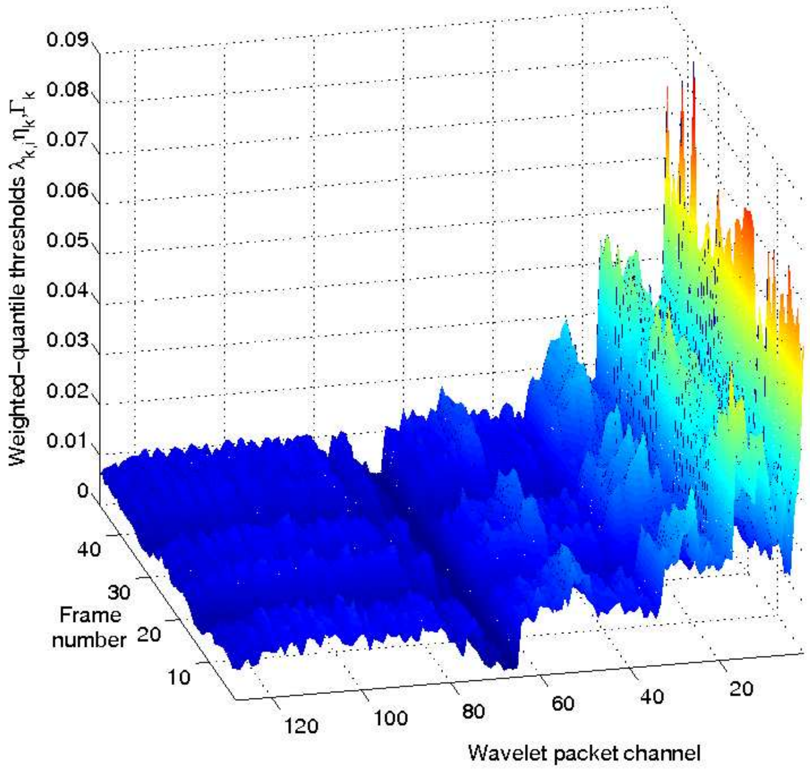

Figure 4.13: Time-frequency weighting over all channels for all frames in a buffer.

4.7 Speech enhancement evaluation

Based on quantile filtering and smoothed hard shrinking of wavelet packet coefficients, the statistical wavelet filtering (SWF) algorithm using time-frequency weighting has been implemented. The proposed algorithm is evaluated on the AURORA3 and NTIMIT databases, and compared with other STFT-based algorithms such as spectral subtraction (SS) [Bol79], nonlinear spectral subtraction (NSS) [EM83], and the NSS algorithm with the noise estimator based on the minimum statistic principle [Mar01] (NSS-MM) (which has been implemented by [Ser03]). All four algorithms are evaluated by objective tests and subjective tests explained in the next sections. Samples of denoised files are available on our website [Pha05].

4.7.1 Experimental setup

The algorithms are tested on a set of 40 utterances each from the AURORA3 database [aur01] and the NTIMIT database [FDGM+93]. The AURORA3 database [aur01] is a subset of the SpeechDat-Car database for the German language. All recordings were

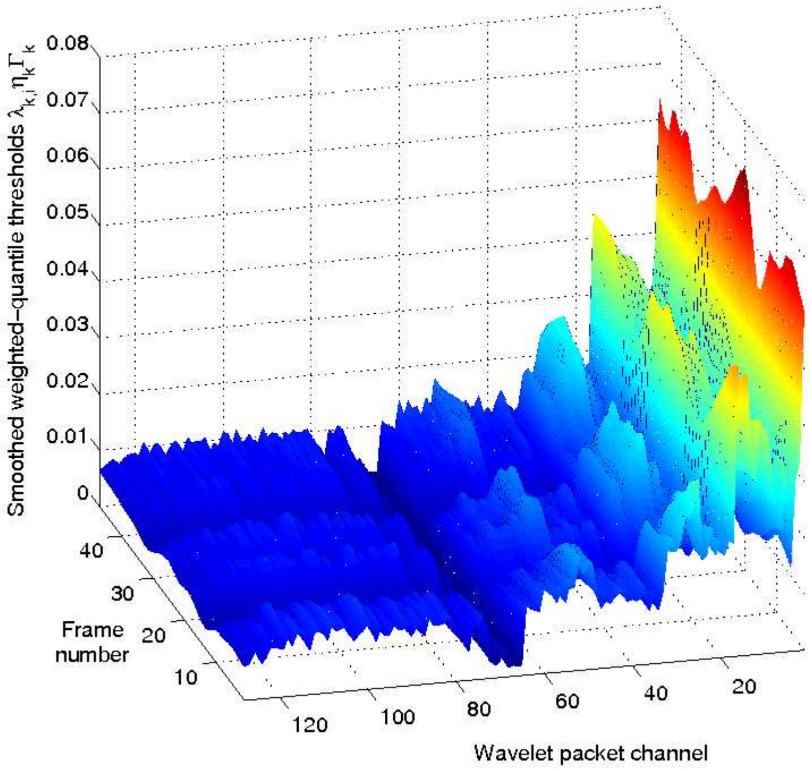

Figure 4.14: Smoothed time-frequency weighting over all channels for all frames in a buffer.

sampled at $F_s = 8kHz$. It contains isolated and connected German digits recorded in car noise conditions. The NTIMIT database [FDGM+93] was built by transmitting all original TIMIT recordings [GLF+93] in the English language through a telephone handset and over various channels in the NYNEX telephone network and redigitizing them. All recordings were sampled at $F_s = 16kHz$. For every database, 40 utterances are selected randomly and each is denoised by four different algorithms. While an objective test is quite straight forward, subjective tests use more complicated procedures and require much more effort. Most of listening evaluation standards were developed for speech synthesis and speech coding applications. In our research, we perform overall quality evaluation by using Comparison Category Rating and the P.85 ITU-T standard. Besides, a segmental evaluation method, the so-called Comparison Diagnostic Test, is developed in the basis of the Diagnostic Rhyme Test standard. All subjective tests have been done by 10 native German male listeners.

4.7.2 Objective test

The segmental signal-to-noise ratio (SegSNR) is estimated for a total of 80 enhanced utterances. As shown in Fig. 4.15, the statistical wavelet filtering algorithm SWF provides higher output SegSNR than the NSS-MM for noisy speech an input SegSNR $\geq$ 5dB and a little bit lower output SegSNR in case of input SegSNR $\leq$ 0dB. The SS and NSS algorithms achieve higher output SegSNR. However, the SegSNR does not faithfully express the speech quality. Thus, we have also performed three more subjective tests on processed data sets.

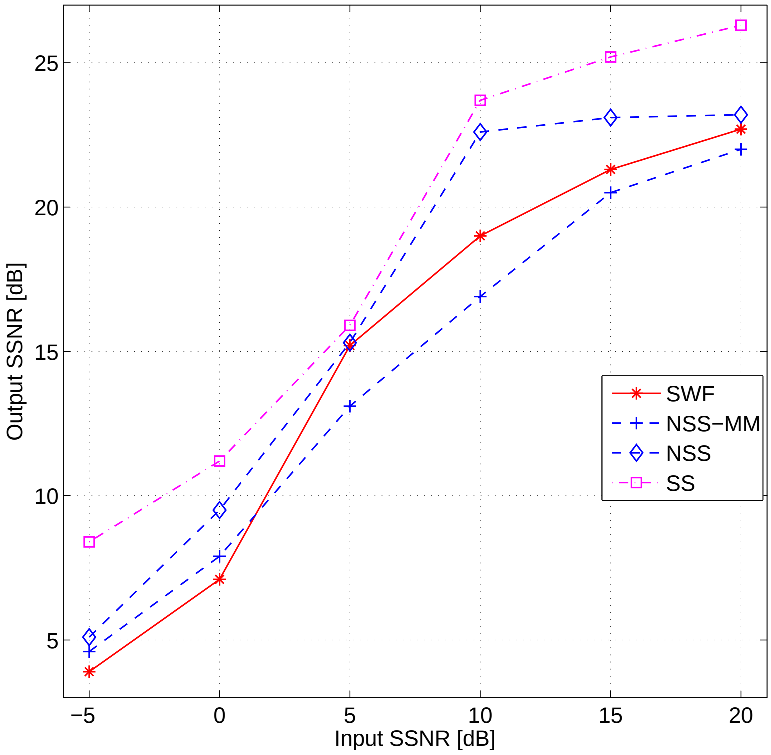

Figure 4.15: Input-ouput SegSNR for different algorithms.

4.7.3 Segmental evaluation: Comparison Diagnostic Test

The Comparison Diagnostic Test is developed on the basis of the Diagnostic Rhyme Test standard introduced by Fairbanks in 1985 [GMW97]. The test material is focused on consonants only because they are more problematic than vowels. From 40 processed files of the AURORA3 data set, 8 utterances spoken by 4 female and 4 male speakers are selected randomly. There, each of 2 files is chosen from one of four different recording conditions (low speed, high speed, stop with operating engine, and running in town).

One is recorded with close-talking microphone, and another one is recorded with far-distance microphone. These 8 files are combined with 2 files chosen from 40 processed files of NTIMIT database to build a data set of 10 files for CDT evaluation. Then 10 utterances consisting of one word or several words which store unvoiced consonants, voiced consonants, and plosives are artificially cut out of the whole sounds for listening purpose.

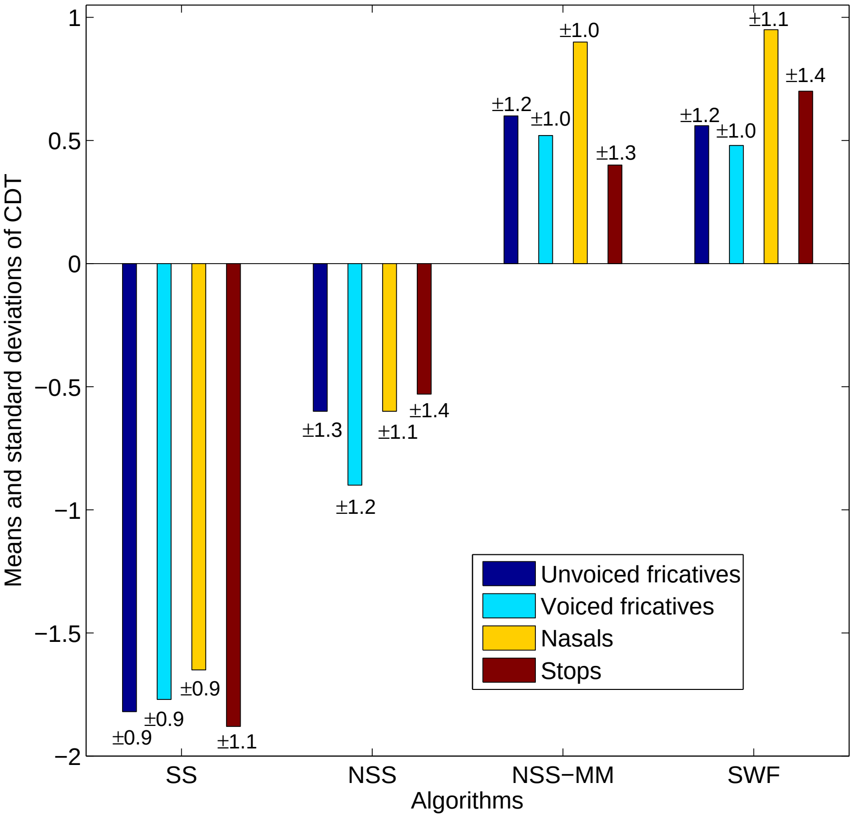

Figure 4.16: Comparison Diagnostic Test results.

The listeners hear the unprocessed file and processed utterances, and concentrate on single phonetic classes that are marked in the corresponding texts. Then they vote by Comparison Category Rating (ITU-T P.800) [GMW97] with scales from 3 (much better) to -3 (much worse) as described in Table D.2. As shown in Figure 4.16a, the SWF and NSS-MM methods maintain consonants better than the two former, ie., SS and NSS methods. The SWF algorithm is the best one in preserving plosives as expected from its nonlinear time adaptive weighting. A detailed description of the proposed CDT is presented in appendix D.1.

4.7.4 Overall evaluation: Comparison Category Rating

In this test, another data set of 10 utterances containing 8 car noise and 2 telephone noise utterances are selected randomly by the procedure of section 4.7.3. The listeners hear the noisy and the enhanced utterances, then judge which of the utterances are better overall and how much (in score values of CCR) it is improved in comparison with the original file. From Tab. 4.2, we see that the SWF algorithm achieves higher performance in case of car and telephone noise with 10dB SegSNR. In the case of 5dB SegSNR, by asking explicitly listeners the best algorithm that they prefer, more than 2/3 of the listeners prefer the output of our system over the original noisy speech.

Databases	Aurora3		NTIMIT
SegSNRs	5dB	10dB	10dB
Algorithms			
SS	-2.23 ± 1.10	-2.45 ± 0.61	-2.25 ± 1.88
NSS	-1.28 ± 1.26	-0.55 ± 1.51	-1.35 ± 1.03
NSS-MM	0.45 ± 1.75	1.00 ± 1.44	0.75 ± 1.78
SWF	0.83 ± 1.33	1.30 ± 1.24	1.30 ± 1.06

Table 4.2: Means and standard deviations of overall evaluation based on CCR.

4.7.5 Overall evaluation: ITU-T Standard P.85

Again, a third data set of 10 utterances is constructed for this listening test ITU-T standard P.85 [GMW97]. By using absolute scales per category, the listeners assess the speech quality of noisy speech and denoised speech. There are 6 categorical ratings such as acceptance, overall impression, listening effort, comprehension, articulation, voice pleasantness. The definitions and score ranges of all categories are given in appendix D.2. The lower the mean values of a category are, the higher the speech quality. The scores evaluated on the Aurora3 and NTIMIT databases are shown in Figure 4.17 and Figure 4.18, respectively. The two latter algorithms get similar or lower mean values compared with the ones of the noisy speech. In almost all categorical ratings with various SegSNRs, the performance of the proposed SWF always exceeds the one of the NSS-MM. In the categories “overall impression” and “listening effort” the SWF algorithm yields a clear improvement over the original noisy speech.

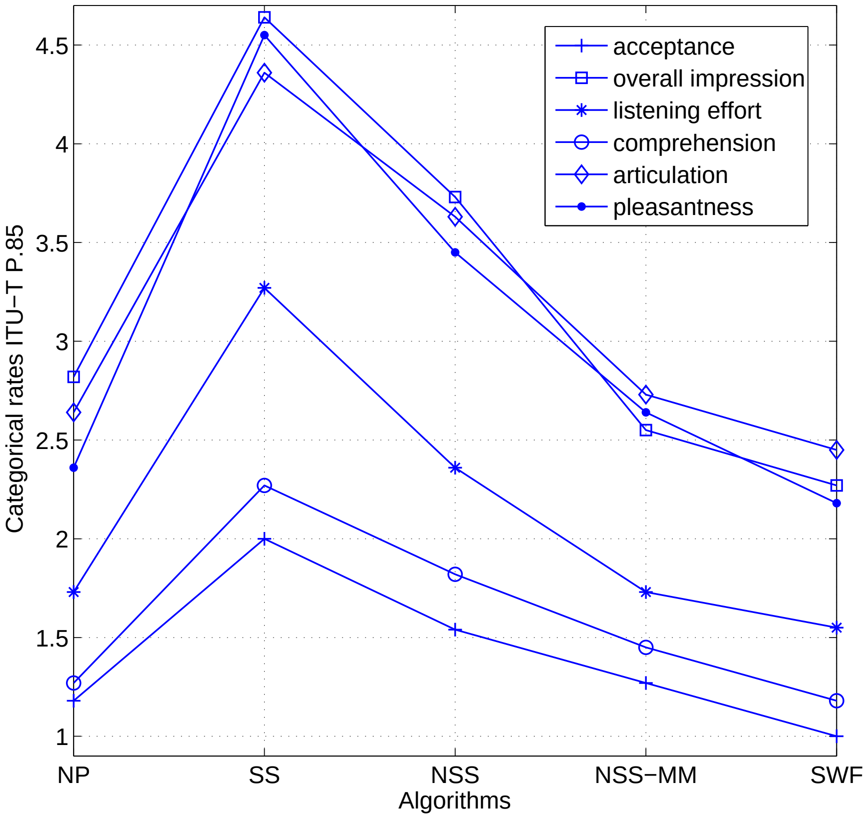

Figure 4.17: Overall test according to ITU-T (P.85), scores obtained on the Aurora3 database processed by different algorithms and without processing (NP)

4.8 Conclusion

A speech enhancement system based on a statistical wavelet shrinking is designed to eliminate musical noise as well as handle colored and non-stationary noise. The scale-dependent threshold relating to the noise level is estimated by employing statistical quantile filtering for every frequency channel and each time-domain buffer. In addition to the use of a scale-dependent threshold, a nonlinear weighting function in frequency helps to handle non-white noise. To achieve a better estimate of the noise threshold in case of non-stationary noise, a temporally adaptive weighting function has been designed and used together with quantile filtering to handle non-stationary noise. The good results of subjective tests confirm the possibility of a high-level of noise suppression while preserving speech intelligibility and naturalness. Application of frequency weighting which leads to the less impact on the voiced sounds and the proposed adaptive factor of the smoothed hard shrinking function provides an explanation for the preserved naturalness as observed in the evaluation experiments. Finally, the specific

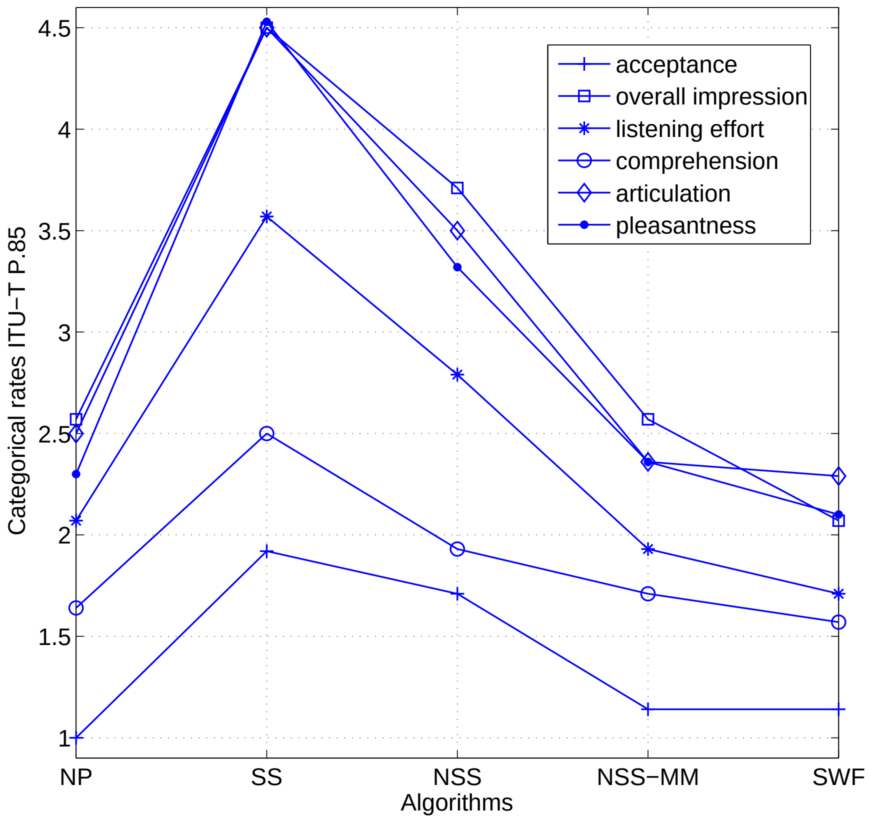

Figure 4.18: Overall test according to ITU-T (P.85), scores obtained on the NTIMIT database processed by different algorithms and without processing (NP).

time-frequency analysis of the DWT using a short basis function to analyze high-frequency components and vice versa contributes to the good maintenance of quality of the unvoiced sounds. The newly developed Comparison Diagnostic Test provides a more accurate evaluation of the quality of unvoiced sounds in speech enhancement applications.

As we see from our studies, the design of the wavelet thresholding/shrinking or gain function is very important in this wavelet-based speech enhancement approach. A survey on the application of different wavelet shrinking functions, especially the generalized form [SB97, LGG04], to enhance noisy speech signal should be done in the future. The use of optimal non-linear estimators based on the Bayesian approach [Vid98, Lav06] opens an interesting research area for deriving optimal non-linear gain functions in the wavelet domain. Beyond the universal threshold, the minimax and SURE thresholds can also be applied. In order to improve the estimate of the noise level in case of non-stationary noise, quantile filtering could be improved by using an adaptive quantile factor as proposed in section 3.6.1. The minimum statistics principle could be implemented at every wavelet channel and compared with quantile filtering in terms of speech enhancement performance. The selection of the optimal WPD tree as well as the employment of the perceptual WPD will be studied to reduce the processing time of the proposed system while maintaining the utmost speech quality. Besides, the replacement of WPD by STFT in the proposed denoising framework needs to be studied as a performance comparison between application of Wavelet transform and Fourier transform in speech enhancement as well as speech recognition applications.

Chapter 5

Noise reduction for robust speech recognition

5.1 Introduction

As one part of human-machine communication, the objective of automatic speech recognition system can be interpreted as a transcriber of linguistic contents of speech signal into words or sentences. The most interesting topic in this field over the last few years is the robustness of ASR system in noisy environments. The current laboratory ASR system is able to recognize continuous speech with average error rate from 5% to 10%. If speaker adaptation is allowed, the error rate drops below 5% after several minutes [You96]. However, the recognition performance decreases dramatically in intensively noisy environments which distort the speech signal. Besides variability in the speech signal caused by inter-speaker, intra-speaker, contextual and linguistic variations [Hil04], ambient noise of operating environment is a very important factor that can enormously reduce the recognition rate of speech recognizer in real world applications. In addition, a mismatch between training and testing conditions leads to the low recognition performance. This physical constraint, the lack of robustness to environmental variabilities, motivates a technical challenge needed to be overcome. An ASR system is environmentally robust if it is able to cope with a wide range of noise sources in different environments such as cars, construction sites, cafeterias, offices, stations, factories etc., where very low signal-to-noise ratios are encountered. The ambient noise impacts not only speech signal but also speech production which is known as Lombard effect [LT71, Jun93, JFF99]. In strong noise environments, a human speaker always increases the vocal effort, e.g., by changing the articulatory shape, which results in variations in the produced speech signal.

Since last decade, plenty methods have been proposed and studied to improve the robustness of speech recognizer in harsh environments. Roughly, the methods can be divided into three approaches: acoustical model adaptation, robust feature extraction, and noise reduction.

- **Acoustical model adaptation** is used to compensate the mismatch between training and testing conditions, thereby increasing the robustness of ASR system. A common method is parallel model combination [GY96] which uses two separated models for clean speech and noisy speech. However, the trained model which is optimized for one target environment may not handle environments with other noise sources. Also, this approach increases complexity of the ASR system.

- **Robust feature extraction** is crucial in a speech recognition system for extracting the best parametric representation of necessary linguistic information from speech waveform. The objective is to find a lower-dimensional representation of information from the speech signal that enhances the phonetic differences. The representative features may be extracted via three methods such as Fourier transform, linear prediction, and wavelet decomposition. In the first method, the mel-frequency cepstral coefficients (MFCCs) [DM80] are used widely in the current ASR systems. The second method consists of finding coefficients for a linear prediction (LP) filter [Ita75] which is equivalent to autoregressive modeling of the signal, and the cepstral coefficients are derived from the LP filter coefficients (linear prediction ceptral coefficients, LPCC) [HN97]. As observed by [DM80], the MFCC features obtained from Fourier spectrum preserve more acoustic information than LPCC, derived from the linear prediction filter transfer function, and support better compression of small spectral variations in high-frequency bands. Although speech recognition based on MFCC features achieves a high recognition rate in the condition of clean speech signals, performance is poor in the harsh environments, especially at very low SNR levels. The drawback may result from the fixed resolution in time-frequency plane of the STFT. Several parameterization methods based on DWT and WPD have been proposed to address this problem in the literature [Kot04, SPH98, GT00, GG01]. It is reported that the usage of WPD based features improves performance of speaker identification and ASR as compared to MFCC features [SPH98, Kot04], but constitutes an increased computational load.

- **Noise reduction** is considered as an effective approach for robust ASR system. As shown in Fig. 5.1, the quality of recorded speech signal is enhanced in a pre-processing stage to ensure that proper information will be extracted by feature extraction stage, thereby increasing recognition performance of the ASR systems. In comparison with the two former approaches, the third solution using a pre-processing stage for removing noise leaves an existing ASR system untouched.

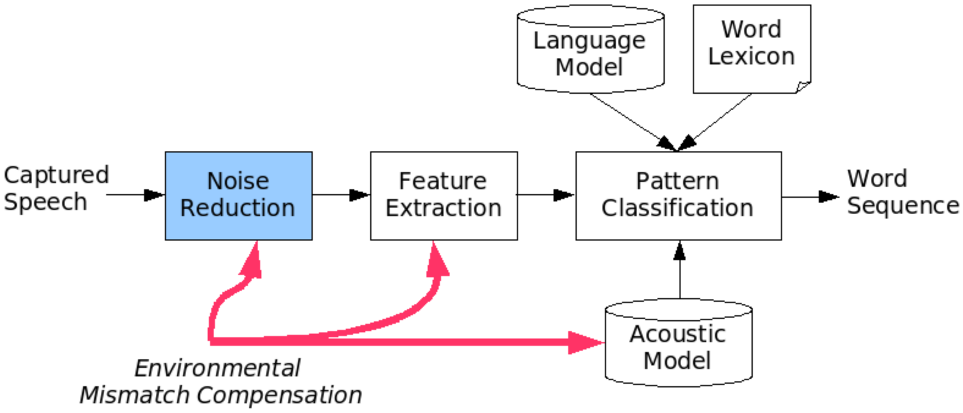

Figure 5.1: Noise reduction approach for robust ASR.

In the framework of our study, we concentrate on the third approach by optimizing the proposed statistical wavelet filtering method for robust speech recognition. As reported in chapter 4, for noise reduction of speech signals targeted at increased perceptual comfort for the human listener, the proposed algorithm was found to preserve overall quality (intelligibility, naturalness, etc.) while keeping a robust attenuation of background noise. This motivates the question whether the increase in perceptual quality of a speech signal leads to an increased recognition rate of the ASR system or the other way round. In the second investigation, several enhancements of the statistical wavelet filtering are introduced as effective solutions for improving ASR performance. The frequency weighting shape is changed to adapt with demands of ASR. Moreover, noise thresholds are estimated in critical subbands by employing perceptual wavelet packet decomposition. In a following study, by integrating the proposed noise reduction process into training phase, the retrained models are expected to provide higher recognition rates. To assess how effective the noise reduction process is under mismatch conditions between training and testing phase of the ASR system, the experiments with three different conditions as well-match, medium mismatch and high mismatch are implemented and compared. The AURORA3 [aur01] and SNOW [sno] databases which consist of complex noise from harsh acoustic environments are used to evaluate the efficiency of the optimized noise reduction algorithms.

The chapter is organized with the description of noise characteristics of practical environments and some selected noise reduction algorithms for robust ASR in the following section. Structures of standard and advanced front-ends are reported in the next section. After that, different solutions to enhancing the statistical wavelet filtering for improving ASR performance are described. Finally, the experimental results are presented and discussed. Conclusions and proposals for further development end up the chapter.

5.2 Why noise reduction for robust ASR?

In particular when ASR systems are used in environments where it is not possible to use close-talking microphones, like for hands-free applications in cars, or if strong sources of background noise, like factory noise, impact the speech signal noise suppression is required to achieve robust recognition performance. The noise sources in a car may come from inside and outside as studied in [Hil04]. The outside noise is low-frequency noise due to mechanical sources such as engine, and tires, etc. The inside noise results from audio equipments, passengers, acoustic warning signals, windscreen wiper, etc. This kind of low-frequency noise has quasi-stationary characteristic. Speech recognition in such car noise environments is studied in the AURORA3 project [aur01]. In a real-world environment, also strongly non-stationary noise sources, such as transient noise of machines in the factory floor, or strongly correlated noise sources, as harmonic noise from engines, have to be expected. The application of ASR in such a harsh environment is considered in the European project SNOW (Services for NOmadic Workers) [sno], where the task is to provide robust ASR for workers in a factory floor environment, namely in airplane maintenance. The non-white and non-stationary noise sources encountered in these environment are a big challenge for noise reduction techniques in robust ASR.

It is known that robust feature extraction and noise reduction are still far away from human performance. This motivates a research with new techniques for improving performance of the ASR system in the wide range of noise sources. Many noise reduction algorithms which were designed for speech enhancement task are optimized and integrated into the ASR system as the pre-processing module. The optimal nonlinear spectral subtraction [EM83] which achieves a significant noise reduction while maintains high speech quality are used in co-operating with the minimum statistics approach [Mar93, Mar01] for noise estimation. This method can track the non-stationary noise by updating the noise estimation during the speech activity regions. By applying an optimized spectral subtraction algorithm to remove noise, the recognition rate is improved as reported in [GMM04]. In the ETSI advanced front-end [ETS03] Wiener filtering is used to improve the quality of extracted features.

Recently, with the flexible resolutions in time-frequency analysis of DWT [VK95, AH01] and ability of maintaining unvoiced sounds of optimal wavelet shrinkage [Don95, CKYK02], the wavelet-based noise reduction methods have been applied to improve significantly recognition performance in harsh environments [Ohs93, Kot04]. The proposed noise reduction method based on statistical wavelet filtering [PK05b] which preserves high quality of denoised speech signal is employed for robust ASR in [RPK06]. With the capability of handling non-stationary and non-white noise by quantile filtering, non-linearly adaptive weighting in time-frequency domain, and optimal wavelet shrinkage, the SWF method is expected to increase recognition rate further by experimentally parameter optimization and perceptual estimate of noise threshold.

5.3 Robust front-ends

Since MFCC features are currently the most popular features used for speech recognition, the ETSI standard front-end (SFE) feature extraction [ETS00] is used for the experiment. The recent algorithm which demonstrates the best overall performance as compared to the ETSI standard will be explained briefly in its block scheme. It is selected as the new standard for the advanced front-end (AFE) for distributed speech recognition [ETS03]. The AURORA working group from the ETSI works on the development and standardization of algorithms to parameterize a representation of the speech (at 8kHz, 11kHz and 16kHz sampling rate) that is suitable for distributed speech recognition. The AFE sub-group is in charge of defining the front-end and speech processing methods. The AURORA working group has published two following DSR standards:

- ETSI ES 201 108 (2000-02) provides the algorithm for front-end feature extraction to create Mel-cepstrum parameters and compress these features for lower data transmission rate (4800 bps).

- ETSI ES 202 050 (2003-11) proposes an advanced DSR front-end to improve recognition rate in noise environments. The noise reduction is carried out by Wiener filtering

(8 kHz) and spectral subtraction (11 kHz and 16 kHz) with support of an energy-based VAD. The Mel-cepstrum features of the enhanced speech are extracted after noise reduction.

The speech enhancement methods used in the ETSI standards are well-known methods with limited efficacy in case of non-stationary and colored noise. Since the contribution to these standards is possible, the promising noise reduction method based on wavelet packet decomposition, statistical wavelet filtering, is employed as a replacement for the Wiener filtering to improve recognition rate.

5.3.1 Standard front-end parameterization

The standard front-end structure is based on the mel-cepstral feature extraction. Assuming short-term stationarity, the speech signal is framed with the frame length of 25ms, and the frame shift interval is 10ms. Then the framed input speech signal is filter with a pre-emphasis filter to amplify high-frequency components for compensating the attenuation caused by the radiation from the lips. A Hamming window is applied to the output of each pre-emphasized signal frame, and a fast Fourier transform (FFT) is applied to the windowed signal. Then the Fourier spectrum is smoothed by integrating over Fourier coefficients in Mel-scale frequency bands (channels). There are 24 channels obtained by applying triangular, half-overlapping windows. The output of Mel-filtering is log-compressed. Finally, 13 cepstral coefficients are derived by applying a discrete cosine transform (DCT) to the output of log-compression. The whole process is depicted by the scheme in Fig. 5.2:

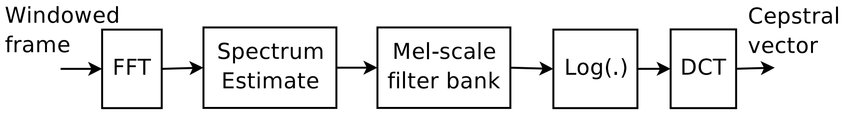

Figure 5.2: Feature extraction scheme of standard front-end

5.3.2 Advanced front-end parameterization

In the advanced front-end project, the cepstral features are calculated from the input speech signal with three main blocks as shown in Fig. 5.3. The noise reduction block is a combination of a two-stage Wiener filter [AC99] and time domain noise reduction [NSJ+01] which are described in next section. An energy-based voice activity detector is used during the Wiener filter design. After that, the SNR-dependent waveform processing block based on the Teager energy operator is applied to the denoised speech signal. The cepstrum calculation block is implemented similar as the standard MFCC extraction in Fig. 5.2 with only a few small differences.

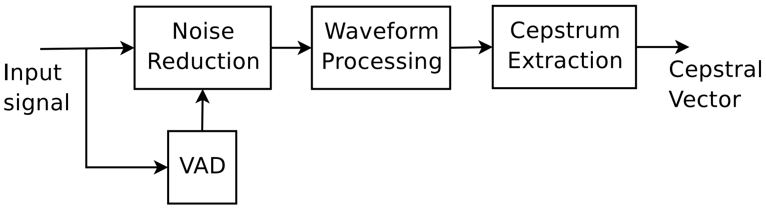

Figure 5.3: Block scheme of advanced front-end

5.4 Two-stage Mel-warped Wiener filtering

This noise reduction process is implemented in two stages of Wiener filtering [AC99] as shown in Fig. 5.4. The two stages are similar but the gain factorization block, and the position of each stage make a difference between two stages as well as their roles in the noise reduction process.

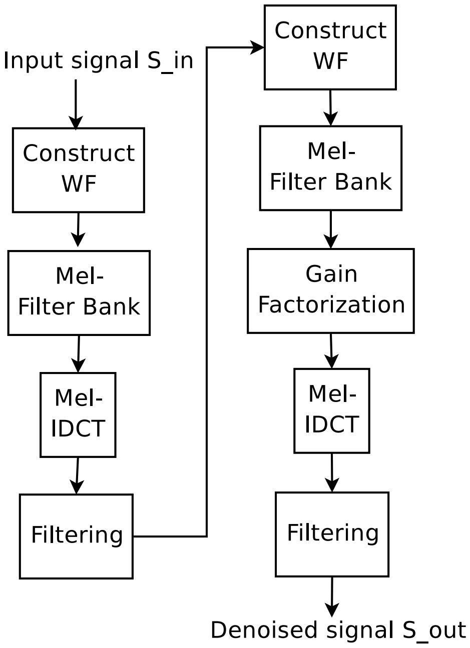

Figure 5.4: Two-stage Mel-warped Wiener filtering noise reduction scheme

The input signal is framed by the 25 ms frame length and 10 ms frame shift. After Fourier transform, by averaging each two consecutive frequency bins of the 128-bin

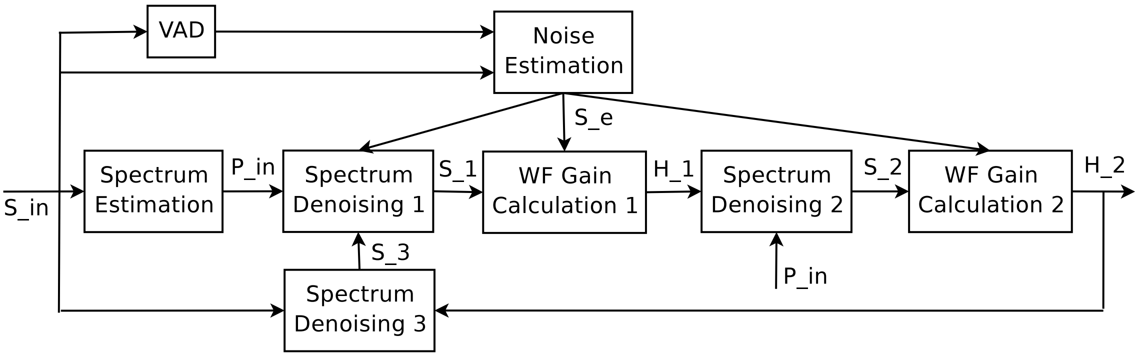

Figure 5.5: Noise spectrum estimation for the construction of Wiener filter

spectrum, the Fourier spectrum is reduced to 64 frequency bins. Next a power spectral density mean (P_{in}) is used to compute the mean over two consecutive power spectrum bins, which reduces the variance of spectral estimation. A voice activity detector based on energy measurement is used for noise estimation in Wiener filter design. A frame is decided as speech if the difference between the current frame log energy and the long-term estimate of non-speech log energy exceeds a defined threshold. The Wiener filter coefficients in frequency domain are calculated by using both the current frame spectrum and the noise spectrum estimation as shown in Fig. 5.5. First the noise spectrum $S_n(k,i)$ is estimated by using VAD, where k is spectrum bin and i is frame index. Then the transfer function of the first Wiener filter is calculated as:

$$H_1(k,i) = \frac{S_1(k,i)}{S_1(k,i) + S_n(k,i)}, \tag{5.1}$$

where the denoised spectrum $S_1(k,i)$ is obtained by:

$$S_1(k,i) = \beta S_3(k,i-1) + (1-\beta)\max\{P_{in}(k,i) - S_n(k,i), 0\}, \tag{5.2}$$

the denoised spectrum $S_3(k,i-1)$ is computed from the previous frame:

$$S_3(k,i-1) = H_2(k,i-1)S_{in}(k,i-1). \tag{5.3}$$

The second denoised spectrum $S_2(k,i)$ is computed by applying the first designed Wiener filter to the PSD mean input signal:

$$S_2(k,i) = H(k,i)P_{in}(k,i), \tag{5.4}$$

and it is combined with the noise spectrum $S_n(k,i)$ to estimate the second Wiener filter frequency response as follows:

$$H_2(k,i) = \frac{\lambda(k,i)}{1+\lambda(k,i)}, \text{ with } \lambda(k,i) = \max\left\{\frac{S_2(k,i)}{S_n(k,i)}, \lambda_T\right\} \tag{5.5}$$

In the next step, the Wiener filter spectrum is smoothed by integrating the spectral coefficients within triangular frequency bins arranged in the non-linear Mel-scale using 24 triangular frequency bins. The denoised speech signal is the output of the convolution of the noisy input speech signal with the Wiener filter impulse response which is obtained from Mel-warped inverse cosine transform (Mel IDCT). The main benefit of the two-stage approach is the flexibility of the Wiener filter design. While the SNR of the input signal in the first stage may be low, the input signal SNR in the second stage is quite high after first denoising. This helps that the gain factorization is performed more accurately.

5.5 Enhancement of statistical wavelet filtering for ASR

As explained in Chapter 4, with the ability of handling non-stationary and colored noise, and preserving phonetic information of the speech signal, the proposed statistical wavelet filtering has a high potential for the application in robust speech recognition. This leads to the question whether the increased perceptual quality of denoised speech signal goes with an increased recognition rate of the ASR system. By applying different noise reduction methods, such as the statistical wavelet filtering, nonlinear spectral subtraction, or Wiener filtering, with parameters optimized for increasing perceptual quality we observed that ASR recognition performance is quite low, even below the one without using noise reduction pre-processing as reported in section 5.6. This leads to the preliminary conclusion, that, unlike the speech enhancement systems which are designed for human hearing purposes, where dynamic comprehension, naturalness and intelligibility are essential criteria, noise reduction algorithms designed for ASR need to be optimized somehow between perceptual constraints and mathematical constraints to compensate the effects of environmental noise, thereby provide robust recognition performance. In this section, the statistical wavelet filtering algorithm is optimized for improving recognition performance. The two main modifications proposed in the next sections are the modification of the shape of the frequency weighting function and the estimation of noise thresholds in critical subbands by perceptual wavelet packet decomposition.

5.5.1 Frequency weighting shape for ASR

With the frequency weighting function η_k given by Equ. 4.36 in the previous chapter, the small quantile threshold values in the range $[0, \ldots, 0.1]$ are amplified strongly while the higher values are only slightly amplified or damped. Since the quantile thresholds are estimated from the universal thresholds calculated from all wavelet packets in each speech frame, their small values are properly obtained by noisy wavelet coefficients which have lower amplitudes than those of useful wavelet coefficients. As discussed in sections 4.4.1 and 4.6.1, this means that the small noisy coefficients will be shrunk due to the amplified thresholds while the large coefficients carrying speech information are only slightly impacted. Consequently, the process leads to low level of remained background noise while maintaining most of the speech spectrum with high quality as shown in Figs. 5.6 (b) and 5.7 (b). In the further, lets call this weighting function η_k^{HA} which is designed for hearing aid (HA).

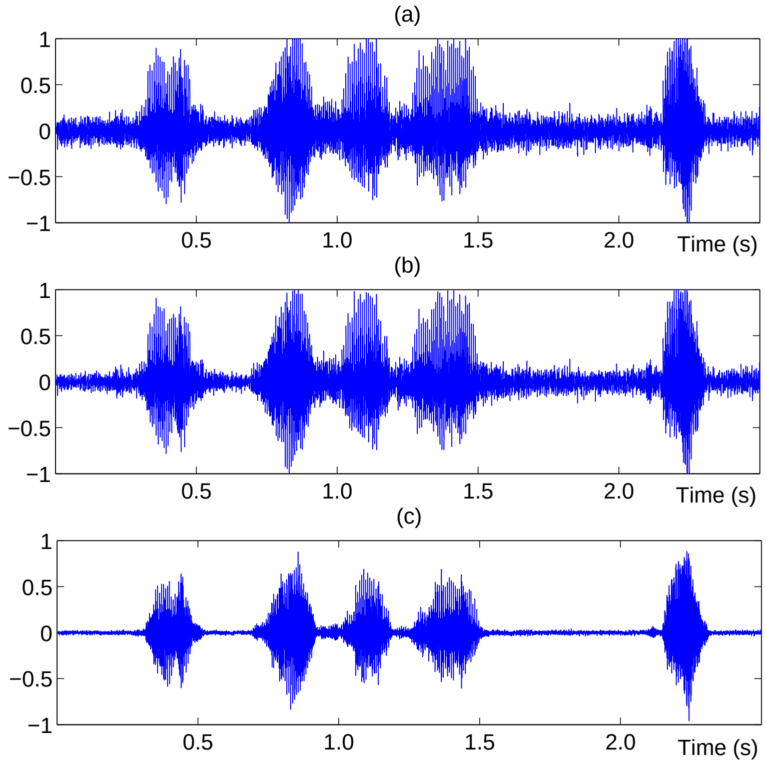

Figure 5.6: Waveform of (a) recording distorted by car noise, and denoised sounds using frequency weightings optimized for (b) hearing aid (HA) η_k^{HA}, and for (c) speech recognition (SR) η_k^{SR}.

In harsh environments where the SNR of captured signal is very low (maybe under 5dB), the less impact on large coefficients results in a high level of background noise

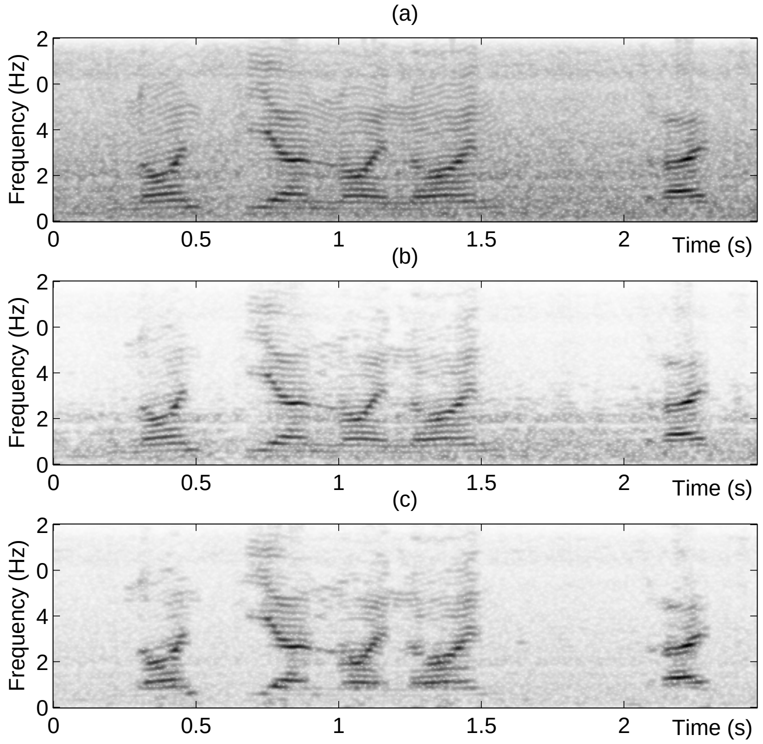

Figure 5.7: Spectrogram of (a) recording distorted by car noise, and denoised sounds using frequency weighting optimized for (b) hearing aid (HA) η_k^{HA}, and for (c) speech recognition (SR) η_k^{SR}.

during the speech regions after denoising. As reported in the listening evaluation section for subjective tests, this can be tolerated for listening aid applications because the less the speech spectrum is modified the more the naturalness of speech sound is maintained. However, the high background noise is a big problematic of speech recognition application in adverse acoustic environments. The overlapping of strong noise on speech regions may provide unsuitable information for feature extraction that results in reduced recognition rate. Moreover, the harmonic noise from engines with a high amplitude spectrum will not be removed using this type of weighting function. This is undesired because the recognizer may confuse noise with speech sound, increasing the insertion error rate. For this reason, another weighting function η_k^{SR} is designed to meet requirements of speech recognition (SR) application:

$$\eta_k^{\mathrm{SR}} = (a_4 \Gamma_k)^{b_4} + d_4 \ . \tag{5.6}$$

where a_4, b_4, d_4 are constants and selected experimentally in order to achieve the high word recognition rate (WRR). From our experimental observation, mainly the factor

a_4 and the power b_4 affect the WRR. Out of a number of parameter sets tried, the set $a_4 = 70, b_4 = 0.5, d_4 = 0.4$ was found to yield the highest WRR for the designed SWF method using full wavelet packet decomposition.

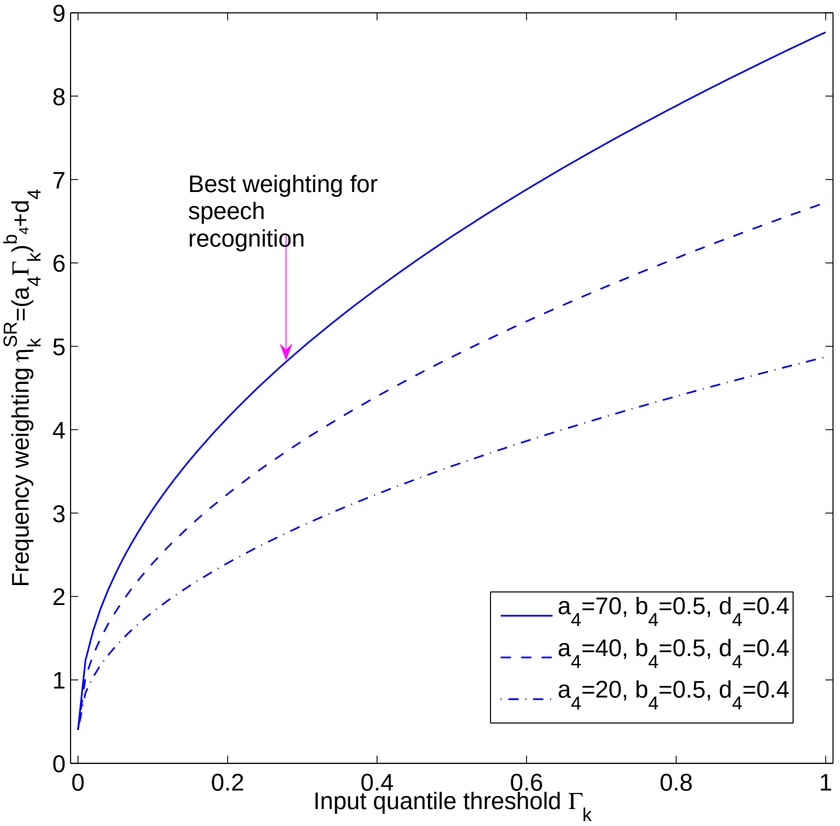

Figure 5.8: Designed weighting function for speech recognition application.

The new weighting functions in Fig. 5.8 produces stronger weighting on the large quantile thresholds Γ_k stemming from wavelet packets containing large coefficients of speech and noise. Obviously, this new weighting results in a denoised speech which is less natural, but still clear and intelligible, with a very low remaining noise level, as shown in Figs. 5.6 (c) and 5.7 (c). The influence of the frequency weighting η_k^{SR} to recognition performance will be evaluated in section 5.6. As another demonstration for the benefit of employing quantile filtering and the frequency weighting η_k^{SR}, figures 5.9 and 5.10 show the effective suppression of an alarm siren component by the method.

5.5.2 Perceptual threshold estimation

In order to improve the accuracy of noise estimation, we propose a perceptual threshold estimation method based on psychoacoustic model of human hearing. The noise

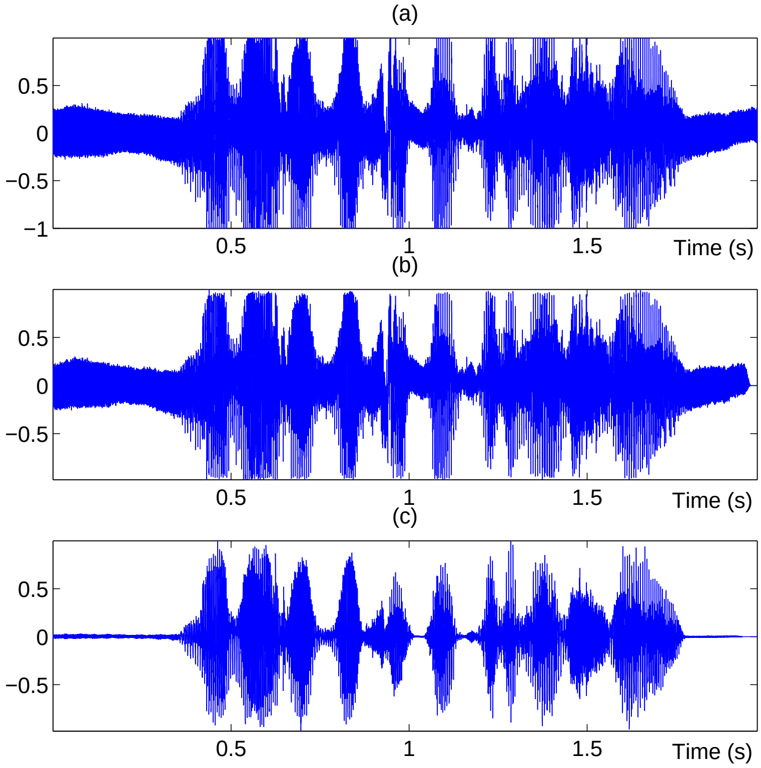

Figure 5.9: Waveform of (a) recording distorted by siren noise, and denoised sounds using frequency weightings optimized for (b) hearing aid (HA) η_k^{HA}, and for (c) speech recognition (SR) η_k^{SR}.

reduction scheme designed in Chapter 4 is modified by adding a so-called threshold mapping module. For each processed frame, instead of using directly the calculated universal threshold values of all channels obtained by full WPD, the threshold mapping function integrates these 128 universal threshold values into 17 threshold values corresponding to 17 critical subbands of the psychoacoustic model. Then statistical filtering and adaptive weighting is applied to estimate the noise thresholds for these subbands. Finally, an inverse mapping is implemented to provide the shrinking gain function for the estimated noise thresholds of all 128 wavelet channels. A block scheme of noise reduction based on perceptual threshold estimation is presented in Fig. 5.11.

In the literature, there are many examples for speech processing using the WPD with subbands designed to match the auditory critical subbands. By applying perceptual WPD (PWPD) in speech applications such as speech enhancement [FW03, CW04], speech coding [CD99, GAE04], speech recognition [FD01], and speaker identification [SPH98, SGFK05], an increased system performance is reported as compared to using conventional WPD. Within our new proposal, we still implement the full WPD. How-

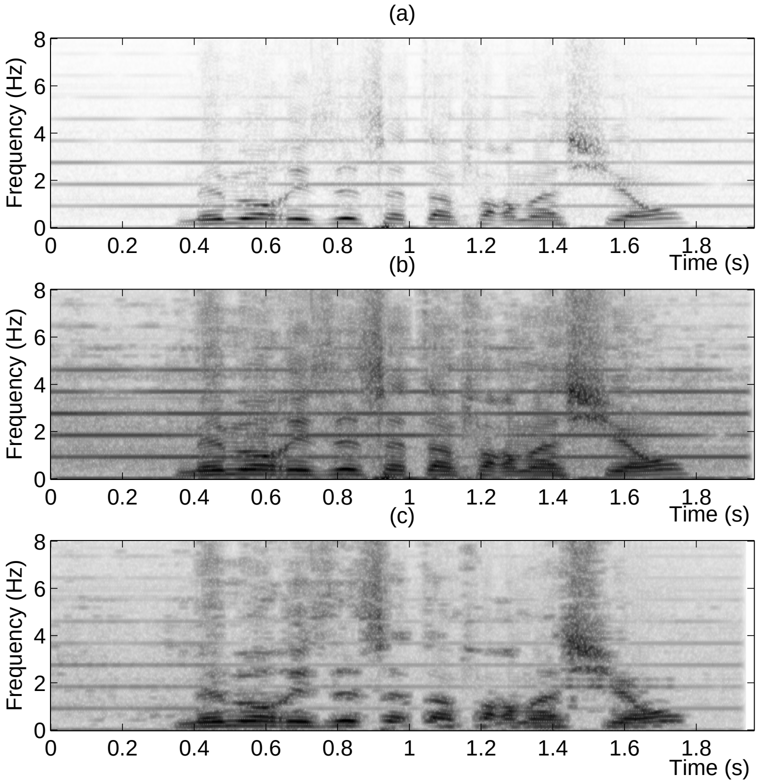

Figure 5.10: Spectrogram of (a) recording distorted by siren noise, and denoised sounds using frequency weighting optimized for (b) hearing aid (HA) η_k^{HA}, and for (c) speech recognition (SR) η_k^{SR}.

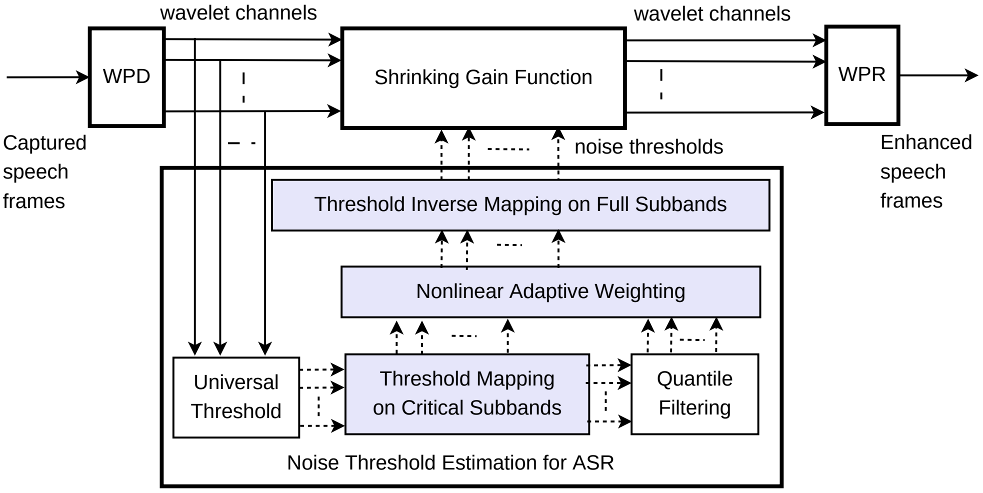

Figure 5.11: Scheme of statistical wavelet filtering using perceptual threshold estimation.

ever, the estimation of noise thresholds is carried out on the critical subbands. Then, the estimated thresholds are used to shrink noisy coefficients of all wavelet channels derived by the full WPD. By this process, the complexity of the system is reduced due to processing on limited number of critical subbands only, while noise is removed efficiently from all wavelet channels. Moreover, the estimate of noise on mel-frequency channels helps to improve the quality of recognition features which are also extracted from mel-frequency channels.

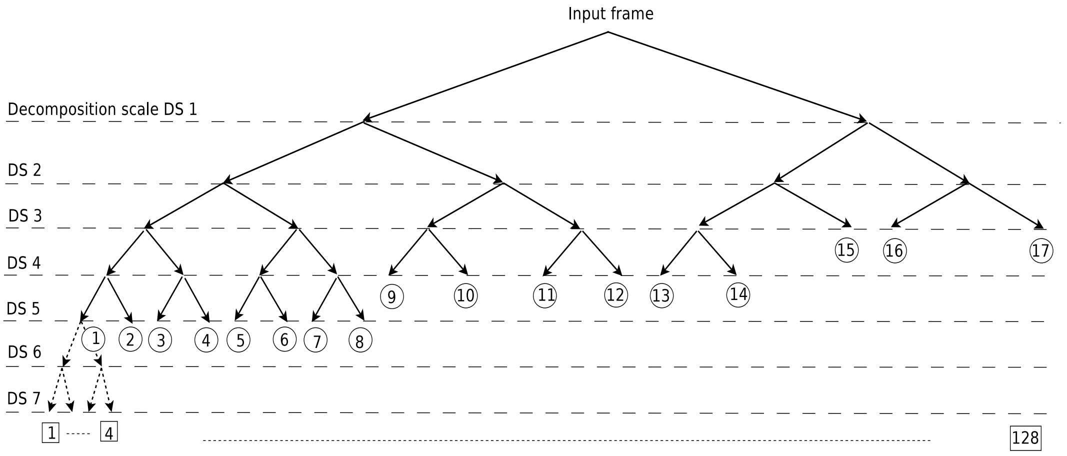

Figure 5.12: Tree structure of the perceptual wavelet packet decomposition.

Bandwidths of the PWPD subbands are designed to match approximately the critical subbands of the psychoacoustic model. As studied in [RJ93], frequency components of sounds can be integrated into critical bands in which the subjective response becomes significantly different. The relations between linear frequency f[Hz] and critical band rate z[Bark], the corresponding critical bandwidth (CBW)[Hz] of the center frequencies f_c[Hz] are described by Zwicker in [ZT80] as follows:

$$z(f) = 13\arctan(7.6x10^{-4}f) + 3.5\arctan(1.33x10^{-4}f)^2 \quad [\text{Bark}] \tag{5.7}$$

$$\text{CBW}(f_c) = 25 + 75(1 + 1.4x10^{-6}f_c^2)^{0.69} \quad [\text{Hz}] \tag{5.8}$$

In our research framework using the SNOW and AURORA3 databases sampled at 8 kHz, there are approximately 17 critical subbands obtained for the bandwidth of 4 kHz [RJ93]. According to the specifications of center frequencies, for CBW obtained from Equation 5.8 a tree structure of the PWPD is constructed [CW04] to approximate the auditory critical subbands on the Bark scale as depicted in Fig. 5.12. The perceptual thresholds $P_{m,i}$ of each critical subband m, at i^{th} frame are estimated by calculating the mean of the universal thresholds $T_{k,i}$ from corresponding wavelet channels k of the

Table 5.1: Mapping between critical subbands and wavelet channels derived by the full WPD.

Critical band no. m	Full WPD Channels $[C1_m..C2_m]$	Bandwidth (kHz)
1	[1..4]	0 - 0.125
2	[5..8]	0.125 - 0.25
3	[9..12]	0.25 - 0.375
4	[13..16]	0.375- 0.5
5	[17..20]	0.5 - 0.625
6	[21..24]	0.625 - 0.75
7	[25..28]	0.75 - 0.875
8	[29..32]	0.875 - 1
9	[33..40]	1 - 1.25
10	[41..48]	1.25 - 1.5
11	[49..56]	1.5 - 1.75
12	[57..64]	1.75 - 2
13	[65..72]	2. - 2.25
14	[73..80]	2.25 - 2.5
15	[81..96]	2.5 - 3
16	[97..112]	3 - 3.5
17	[113..128]	3.5 - 4

i^{th} frame as defined by a following equation:

$$P_{m,i} = \frac{1}{C2_m - C1_m + 1} \sum_{k=C1_m}^{C2_m} T_{k,i} \tag{5.9}$$

where $[C1_m..C2_m]$ are orders of wavelet channels derived by the full WPD. The mapping of these channels into the critical subbands are described in Table 5.1. These perceptual thresholds are then employed in statistical quantile filtering and adaptive weighting to estimate final thresholds of noise levels. With this perceptual statistical wavelet filtering method, the parameters of time and frequency weighting functions are further tuned to improve recognition rates for different front-end setups. A brief description of the tuning experiment is presented in appendix E.2.

5.6 Experiments and evaluations

5.6.1 Speech databases and recognizers

To assess the suitability of the presented noise suppression algorithms for the use in ASR systems, a number of tests were carried out on the SpeechDat-Car corpus [aur01] and on the SNOW [sno] database. The German Aurora 3 SpeechDat-Car corpus includes samples of series of digits recorded in a car environment under various driving conditions such as: car stopped with motor running, town traffic, driving at low speech on rough road, driving with high speed on good road. All speech sounds are recorded with a close talking microphone and a hands-free microphone at a sampling rate $F_s = 8\,\text{kHz}$. The SNOW database consists of a number of 40 phonetically balanced sentences and 60 commands typical for the SNOW application recorded by 10 different speakers in native French in the first data set (SNOW1), and by 12 native and non-native speakers in English for the second data set (SNOW2). The databases are recorded originally at sampling rates of $F_s = 16\,\text{kHz}$ (SNOW1) and $F_s = 22.05\,\text{kHz}$ (SNOW2), and than sub-sampled to 8 kHz. The speech data was collected in the halls of a factory. The noise is on average high level, including both stationary and non-stationary noises, and some background music. The background noise always comprises the full reverb from the halls. As observed, the noise environment for the SpeechDat-Car corpus is not as adverse as the factory floor environment in the SNOW corpus [sno], however, the corpus is widely used for assessing ASR systems and thus allows for direct comparison of noise suppression algorithms, and particularly the training/test set with "high-mismatch" (i. e., with rather clean samples used as training data and very noisy samples as test data) should allow for a rough assessment of how the proposed noise suppression algorithm would behave in a more adverse environment.

Two different speech recognizers were employed during experiments. In the first recognizer, two front-ends were employed as standard front-end MFCC specified by [ETS00] and advanced front-end defined by [ETS03] in combination with the HTK recognizer [YEG+05] for the German AURORA3 SpeechDat-Car corpus. The recognizer structure in the AURORA3 framework is described in detail by [PH00, Pea00]. Simple left to right acoustic models without skips over states are utilized. There are 16 states per word, and a mixture of 3 Gaussians per state. In the Aurora framework, two pause models known as "sil" and "sp" are defined to model the pauses before and after the utterance and the pauses between words. For tests on the SNOW database, the Loquendo hybrid HMM-NN speech recognizer [GAM99] with general purpose acoustic models trained on a large telephonic corpora, including several speakers with a statistical distribution of age, sex, and geographic areas is applied. The Loquendo recognizer employs an acoustic model for noise and end-point detection, which is similar to VAD.

5.6.2 Test results with the SWF

By applying full WPD, the proposed denoising algorithms based on statistical wavelet filtering using different frequency weighting functions built for hearing aid η_k^{HA}, Equ. 4.36 in chapter 4 (SW1), and designed for speech recognition η_k^{SR}, Equ. 5.6 in section 5.5.1 (SW2) are applied in a pre-processing stage to the front-end of the speech recognizers. In addition, the use of non-linear spectral subtraction (NSS) [EM84] and two-stage Wiener filtering (WF) [AC99] of AFE as well as without using noise reduction (w.o.NR) are evaluated as the references. Firstly, the proposed denoising algorithms are tested with standard front-end MFCC (SFE) specified by [ETS00], and used as replacement of the WF in advanced front-end (AFE) [ETS03] of the HTK recognizer [YEG+05] for the German Aurora 3 SpeechDat-Car corpus. Secondly, the statistical wavelet filtering SW2 and the elaborate spectral subtraction of Loquendo (SSL) [GMM04] modified from Ephraim-Malah rule [EM84, EM85] are tested with Loquendo speech recognizer. The performance in terms of word recognition rate (WRR) and accuracy (ACC) derived from the experiment are reported in Table 5.2. The test has been implemented in two modes as without (w.o.) retraining and with retraining of the acoustic models. The "high-mismatch" training/test set of the German SpeechDat-Car corpus and the SNOW1 data set are used for the test.

Table 5.2: Recognition performance as WRR/ACC derived by applying different noise reduction algorithms and without denoising (w.o.NR).

Recognizers	German SpeechDat-Car/HTK				SNOW1/LOQ
Mode	w.o. retraining		with retraining		w.o. retraining
Front-ends	SFE	AFE	SFE	AFE	-
Measures	WRR/ACC	WRR/ACC	WRR/ACC	WRR/ACC	WRR/ACC
Algorithms					
w.o.NR	66.70/63.23	85.34/84.27	66.70/63.23	85.34/84.27	-/-
SW1	61.98/54.49	79.37/51.43	68.83/64.01	83.44/65.96	-/-
SW2	65.63/60.41	83.21/77.30	75.30/73.20	85.20/83.90	78.40/67.10
NSS	60.68/59.39	80.11/74.61	70.44/69.75	83.46/82.39	-/-
WF	-/-	**89.78/89.45**	-/-	79.42/79.28	-/-
SSL	-/-	-/-	-/-	-/-	83.70/75.20

As addressed in the introduction of the chapter and in section 5.5.1, the interesting question whether the increase in perceptual quality of a speech signal leads to an increase in recognition rate of the ASR system or not is firstly examined. We see that the use of the SW2 algorithm instead of the SW1 increases WRR from 61.98% to 65.63% (ACC from 54.49% to 60.41%) for the SFE, and from 79.37% to 83.21% (ACC from 51.43% to 77.30%) for AFE. The achieved performance is also higher than the one derived by using the NSS algorithm optimized for hearing aid.

In case of using the recognizer without retraining, however, the use of the SWF-based noise suppression algorithms does not increase the recognition rate as compared to the use of the original front-ends (no denoising in SFE and using WF in AFE). For the "high-mismatch" training/test set of the German SpeechDat-Car corpus, the WRR is reduced from 66.7 % to 65.63 % (accuracy from 63.23 % to 60.41 %) for the standard MFCC front-end, and from 89.8 % to 83.21 % (accuracy from 89.45 % to 77.30 %) for the AFE. We attribute this mainly to the different training and test conditions, and probably to a negative interference between the proposed noise suppression algorithm and the denoising algorithm in the AFE.

The second test thus comprises the retraining of the HTK recognizer using the proposed noise suppression algorithm SW2 as a pre-processing stage for the standard MFCC front-end, and as a replacement for the Wiener filter denoising in the AFE. Here, the word recognition rate of 66.7 % using the standard MFCC front-end for training and testing is increased to 75.3 % with the proposed algorithm (accuracy from 63.2 % to 73.2 %), however, the word recognition rate for the AFE of 89.8 % is reduced to 85.2 % (accuracy from 89.5 % to 83.9 %).

The third test was performed on a corpus set up in the scope of the SNOW project, comprising 435 utterances (a total of 1135 words, utterances are commands for controlling a graphical browser display) recorded by 4 female and 4 male speakers under work conditions in an airplane maintenance facility. For this recognition test the Loquendo ASR system [Loq06] using SSL denoising algorithm [GMM04] was utilized, using a grammar where all the vocabulary words can be looped without any constraints. The proposed wavelet noise suppression algorithm was again used as a pre-processing stage, in addition to the denoising in the ASR front-end. Like in the experiment with HTK, the word recognition rate is reduced, too, from originally 83.7 % to 78.4 %.

5.6.3 Test results with perceptual SWF

Application of the perceptual noise threshold estimation is examined on the second data set SNOW2 consisting of recordings from 12 native and non-native speakers in English, and on the SpeechDat-Car corpus in three different conditions: high-mismatch (*hm*), medium mismatch (*mm*) and well match (*wm*). From the above evaluation, the frequency weighting function optimized for speech recognition is selected and integrated into the perceptual statistical wavelet filtering (PSWF) algorithm. The parameters of time and frequency weighting functions are fine tuned experimentally to achieve the highest recognition rate and accuracy as possible for different front-ends. A brief description of the tuning experiment is presented in appendix E.2. The PSWF algorithm is compared with the WF algorithm used as baseline in terms of recognition performance in Table 5.3.

Table 5.3: Recognition performance as WRR/ACC using PSWF.

Recognizers	German SpeechDat-Car/HTK			
Algorithms	PSWF		Baseline	
Mode	with retraining		-	
Front-ends	SFE	AFE	SFE	AFE
Measures	WRR/ACC	WRR/ACC	WRR/ACC	WRR/ACC
Conditions				
hm	77.71/76.73	**89.45/86.63**	66.70/63.23	**89.78/89.45**
mm	81.92/78.99	88.65/85.29	78.48/76.43	89.53/89.02
wm	92.91/91.20	95.07/93.25	90.48/87.92	95.55/94.65

In this fourth experiment, by replacing the WF by the PSWF in advanced front-end AFE, the obtained word recognition rate is almost similar to the one of baseline for all different conditions, e.g. 89.45% to 89.78%, 88.65% to 89.53% and 95.07% to 95.55% for *hm*, *mm* and *wm* conditions, respectively. This shows an improvement compared to the use of statistical wavelet filtering SW2 and SW1 in terms of WRR (89.45% to 85.20% and 83.44%) and ACC (86.63% to 83.90% and 65.96%). From Table 5.3, we observe that usage of noise reduction as a pre-processing stage only makes sense if there is a high mismatch between training and testing environments. In the *hm* condition, the WRR is improved up to 11.11% (from 66.70% to 77.71%) but the WRR improvement is lower, 3.44% and 2.43%, for the *mm* and *wm* conditions, respectively.

This observation not only confirms the crucial role of noise reduction for environment mismatch compensation, but also opints at the necessity of background model adaptation in case of a mismatch between the training and testing data sets.

With more test results reported in Tables E.2, E.3, and E.4 in appendix E.3, we observe that the need of model retraining depends on the mismatch conditions no matter the front-ends are used. The most noticeable improvements from without to with retraining are found for the *hm* condition. This makes sense since there is a big gap of environmental mismatch between training and testing sets. Thus, operating condition can be considered for decision on retraining model.

By evaluating ASR performance in terms of number of deleted words (D) and number of inserted words (I), very useful observations are made: In the AFE setup, the usage of both WF and VAD results in very low I but high D due to the impact of VAD with wrong voice activity decisions. However, with the proposed PSWF, much lower D is obtained for all tried experiments. The application of temporal weighting function which is considered as a kind of voice activity detection introduces a soft decision with the smoothed transitions at the word edges. Of course more noise frames will survive as a trade-off of this mechanism. However, a high number of false rejection is more harmful to recognition performance than high number of false acceptance.

In the final experiment, the PSWF algorithm is tested with the Loquendo speech recognizer on the SNOW2 database. The PSWF provides slightly higher word accuracy 94.78% to 94.69% which is derived from baseline of Loquendo recognizer without using end-point detector. In case of using end-point detector that tries to identify the starting point and ending point of user utterance, the Loquendo baseline ACC drops down to 68.20%. The risk is that a portion of utterance is removed when the end-point detector fails. However, the ACC is increased to 88.98% by using the PSWF. It is interesting that the PSWF algorithm really helps EPD to reduce the deletion of speech portions.

5.7 Conclusion

In [PK05b] we have shown that an elaborate SWF-based speech enhancement algorithm allows for consistent attenuation of background noise while preserving speech naturalness and intelligibility. In this chapter the algorithm has been adapted to suit the requirements of noise suppression for the use with ASR systems. The modification to the algorithm allow for a more aggressive suppression of low-frequency background noise compared to the previous setting [PK05b]. A further optimization is done by applying the perceptual wavelet packet decomposition tree and the estimate of noise threshold in critical subbands. The experiments with ASR systems show, on the one hand, that, for the application in adverse noise environment, no improvement in recognition rate can be achieved when the proposed algorithm is used as a pre-processing module in addition to ASR internal noise reduction methods without re-training. On the other hand, if the ASR system is trained with the noise suppression algorithm, a significant improvement is achieved using the ETSI 201 108 standard front-end, and the proposed algorithm almost achieves the performance of the noise reduction in the ETSI 202 050 advanced front-end.

Thus, this promising SWF-based algorithm can be applied for both speech enhancement and ASR, and should be further investigated and optimized for ASR in future research. In spite of the more aggressive noise suppression approach targeted at the application with ASR, the proposed algorithm still provides better performance regarding naturalness of the speech signal in a comparison with other denoising algorithms in informal listening tests. In particular, the combination with voice activity detection, as used in the AFE, should be beneficial to increase accuracy by reducing the number of insertions. To conclude, one should think how to fully integrate this noise suppression algorithm into the front-end processing in order to extract more accurate and robust recognition features.

Chapter 6

Conclusions and perspectives

The dissertation has investigated several applications of wavelet analysis and wavelet denoising in robust speech processing and advanced applications. Now, we summarize the achieved results and outline perspectives for future research.

Through most chapters of the thesis, robustness to acoustic background noise for speech processing and applications is the focus. In modern speech technology, the need for extracting efficient acoustic features to improve phonetic classification, the need for enhancing the quality of recorded speech signals distorted by background noise from real-word environments, and the need of reducing noise for achieving robust word recognition in harsh environments are well motivated. The Wavelet transform with its smart ability of decomposing signals into well-localized time-scale features and optimal wavelet shrinkage as a powerful noise removing tool is applied for the addressed goals.

In our research we designed novel phonetic classifiers by focusing on reliable feature extraction and by applying advanced machine learning approaches. Based on the analysis of acoustic properties of the phonetic sound classes, the DWT is exploited to extract useful time-scale features which characterize these classes. We classify sequences of speech frames into six phonetic classes, i.e., voiced, unvoiced, silence, mixed-excitation, voiced closure, and transient classes. An efficient joint classifier was built by combining a multi-threshold decision classifier with two trained feed-forward neural networks. This joint classifier provides optimal non-linear thresholds learned by FNN and low complexity due to the usage of the linear multi-threshold model. As another approach to pattern recognition, we trained Bayesian networks for the task. The obtained results show that the proposed time-scale features provide classification rates comparable to the baseline MFCC features. Additionally, the use of only 7 time-scale features compared to 13 MFCC features leads to a simpler classifier. Discriminative parameter/structure training improves the classification rate in most cases and gives slightly better classification performance than the one derived by the joint FNN classifiers. The study of gender-dependent/independent classifiers opens an approach towards the realization of gender-independent phonetic classification.

With the promising results obtained from above research, we developed a robust speech classifier to detect voiced/unvoiced/silence segments and speech/non-speech segments from noisy speech signals. After applying Teager's energy operator on wavelet coefficients, a single time-scale feature is extracted and further enhanced by the hyperbolic tangent sigmoidal function and median filtering to make it robust against noise. A quantile filter and slope tracker are designed as two advanced methods for adapting the decision threshold. We developed a quantile filtering method which is based on the minimum statistics method in order to estimate the adaptive threshold accurately, especially in case of non-stationary noise. We also developed the slope tracking method in three steps to classify phonetic classes while meeting delay and memory requirements of real-time applications. As a further evaluation, we designed a robust voice activity detector with adaptive quantile factor and integrated it into the speaker verification system as a pre-processing stage. The robustness of the invented method is again confirmed by the improved verification rates in a simulated environment of air traffic communication.

Based on the adaptive estimate of the quantile filtering technique, we design a speech enhancement system using statistical wavelet filtering (SWF) which is able to eliminate musical noise as well as handle non-white and non-stationary noise. The proposed algorithm operates in the wavelet packet domain and employs the smoothed hard shrinking gain function to suppress noise. Noise thresholds are estimated adaptively by quantile filtering at every frequency channel over a recursive buffer. Together with the use of a scale-dependent threshold, a nonlinear weighting function in the frequency domain helps to handle non-white noise. In addition to the adaptive quantile threshold estimate, temporal adaptive weighting functions have been designed to provide a better noise threshold estimate from non-stationary noise. With the proposed adaptive factor, the smoothed hard shrinking gain function can suppress noise as much as possible while maintaining perceptual quality of the enhanced speech signal. The superior performance obtained by subjective tests shows that the invented algorithm is comparable to state-of-the-art single channel speech enhancement methods operating in the Fourier domain. To cover more aspects of enhanced speech quality evaluation, we proposed the Comparison Diagnostic Test to examine the quality of unvoiced sounds which are most problematic in the speech enhancement field.

To increase the robustness of automatic speech recognition (ASR) in adverse environments, the proposed statistical wavelet filtering is employed as pre-processing stage before the front-end feature extraction unit of a speech recognizer. The goal is to ensure that proper features for word recognition will be extracted from the enhanced speech signal, thereby increasing the recognition performance of the ASR systems. We have shown in previous research that the elaborate SWF-based speech enhancement algorithm allows for consistent attenuation of background noise while preserving speech naturalness and intelligibility. This brings us to a tricky question whether the increase in perceptual quality of a speech signal leads to an increase in recognition rate of the ASR system or not. From our test results, we see that the SWF needs to be further optimized with respect to recognition performance. Three adaptations have been studied to achieve robust word recognition performance: changing the shape of the frequency weighting function, experimentally tuning the parameters, and estimating the noise thresholds for critical subbands derived from perceptual WPD. The experiments with ASR systems show, on the one hand, that no improvement in recognition rate can be achieved when the proposed algorithms are used as a pre-processing module in the ASR systems without re-training. On the other hand, if the ASR system is re-trained on the training data set enhanced by the proposed noise suppression algorithms, a significant improvement is achieved using the ETSI 201 108 standard front-end, and the proposed algorithm almost achieves the performance of the noise reduction in the ETSI 202 050 advanced front-end. In the last study, we show the need of employing noise reduction as pre-processing stage to compensate the mismatch between training and testing phases of the ASR system. With this solution, the recognition rates are increased significantly in case of highly mismatched environments.

The promising results obtained from the studied systems raise some open issues that are interesting to be investigated in the furture. A better set of time-scale features should be extracted to improve classification performance for the mixed-excitation, voiced closure, and transient classes. The usage of hidden Markov models with Viterbi decoding is a good candidate for replacing the phonetic interpolation rule in detecting plosives. The optimization of the robust WT-based VAD needs to be constrained to the maximization of EER performance. It is very interesting to study the relationship between VAD objective performance and application performance, for example speaker verification in our research. Dealing with the novel speech enhancement algorithm, a comparison of different wavelet shrinking functions to enhance noisy speech signals could be done. The use of optimal non-linear estimators based on the Bayesian approach opens an interesting research area for deriving optimal non-linear gain functions in the wavelet domain. We want to combine the robust VAD with the proposed noise reduction algorithm to increase the accuracy of the speech recognizer by reducing number of insertions. To conclude, one should address how to fully integrate this noise suppression algorithm into the front-end processing in order to extract more accurate and robust recognition features.

From the significant achievements of the presented study, we do believe that the approach of applying wavelet analysis as well as wavelet shrinkage is an excellent candidate for the next generation of modern speech technology.

Appendix A

Wavelet properties

The continuous wavelet transform (CWT) of any signal $x(t)$ is defined as:

$$CWT_x(a,b) = \int_{-\infty}^{+\infty} x(t)\psi_{a,b}(t)dt = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} \psi(\frac{t-b}{a})dt, \tag{A.1}$$

One of the most important properties of the wavelets is their admissibility condition [Mal99]:

$$C_\psi = \int \frac{|\Psi(\omega)|^2}{\omega} d\omega < +\infty \tag{A.2}$$

where $\Psi(\omega)$ is the Fourier transform of $\psi(t)$. The condition allows a decomposition and a restruction of a signal $x(t)$ without loss of information [Val04]. To make sure this integral is finite, $\Psi(\omega)$ must be zero at $\omega = 0$. This also means the average of the wavelet equals zero:

$$\int \psi(t)dt = 0 \tag{A.3}$$

It is essential to select wavelet basis with desired properties such as *vanishing moments* and *compact support* due to their influence to the sparsity of the wavelet representation. We interpret here the concept of number of vanishing moments. As presented in [She96], the CWT in A.1 can be expanded into the Taylor series till order P at $t = 0$ as:

$$CWT_x(a,0) = \frac{1}{\sqrt{a}}\left[\sum_{p=0}^{P} x^{(p)}(0) \int_{-\infty}^{+\infty} \frac{t^p}{p!}\psi\left(\frac{t}{a}\right) dt + \mathcal{O}(p+1)\right], \tag{A.4}$$

where the shift factor $b = 0$ for simplicity, $x^(p)(t)$ is the p^{th} derivaties of $x(t)$. Now we rewrite the expansion in terms of M_p moments as:

$$CWT_x(a,0) = \frac{1}{\sqrt{a}}\left[ax(0)M_0 + a^2\frac{x^{(1)}(0)}{1!}M_1 + ... + a^{P+1}\frac{x^{(P)}(0)}{P!}M_P + \mathcal{O}(s^{(P+2)}\right], \tag{A.5}$$

where the moment M_p is defined as

$$M_p = \int_{-\infty}^{+\infty} t^p \psi(t) dt \tag{A.6}$$

Because the average of the wavelet equals zero, we have the 0^{th} moment $M_0 = 0$, and the first term goes to zero. If we have a wavelet with P vanishing moments, the the polynomial of order P will be cancelled out of the wavelet coefficients. These polynomials are moved to scaling space. Actually, the moments do not have to be zero, but small values are good enough [Val04] for compression or noise removal tasks.

Appendix B

Multi-threshold decision classifier

The multi-threshold decision (MTD) model which was developed in [PK04] considers four input features to decide the classes of every speech frame. The hard thresholds which are employed by the model are chosen via experimental pattern classification. It works by comparing these thresholds with the extracted input features. Based on these results, the classification decision is carried out. The procedure of determining good thresholds can be described as follows:

- Dataset: About 40.000 training segments from 110 sentences of female and male speakers, dialect speaking region 1 (DR1), have been selected from TIMIT database [GLF+93].

- Wavelet basis selection: The good wavelet basis is found as $bior5.5$ (among 39 wavelet bases of Wavelet toolbox in [MMOP02]) whose WPR values are consistent over all kinds of frames when checking WPR over nearby 300 frames of phonemes $/A/, /V/, /S/$ in the database of sustained phonemes. The initial thresholds $WPR1$, $WPR2$ are taken from [Agb96], and $SAE1$ is obtained from observing 10 silence frames.

- Next, 396 frames of female speakers are classified by these initial thresholds. Based on the TIMIT transcription file, these thresholds are modified and some more thresholds are added as $WPR3$, $PVD1$, $PVD2$, $SAE2$, $SAE3$. These procedures are repeated for other recordings to improve classification accuracy.

- Then, the selected thresholds are applied to male speakers and modified to trade off performance between female and male speakers. Finally, the good thresholds are obtained. This procedure has the advantage of revealling automatically gender-dependencies of the classifier. Fig. B.1 shows scatter plot for 3-dimensional features extracted from 2000 randomly samples.

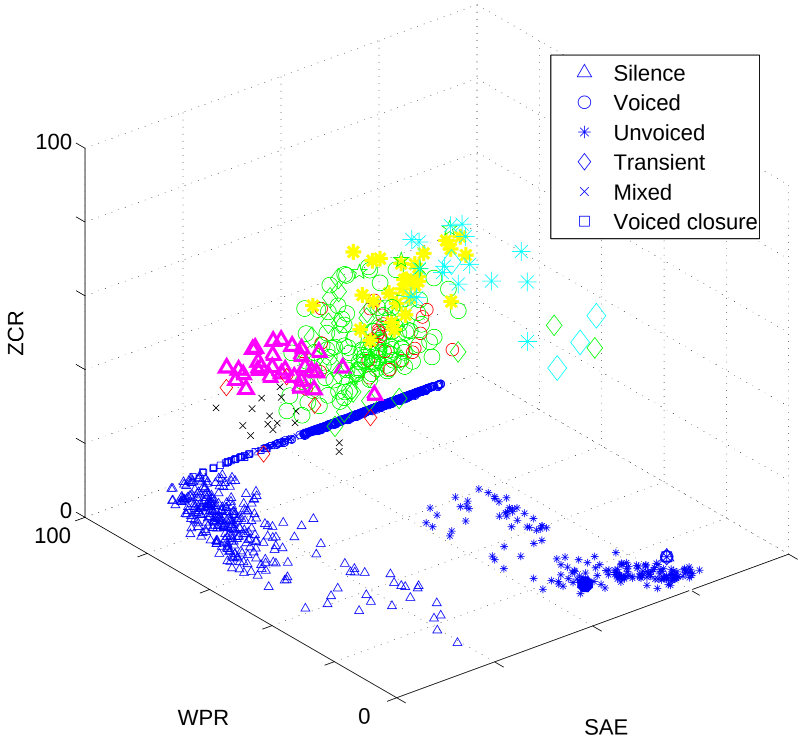

Figure B.1: Scatter plot for feature space derived from 2000 frames (all samples denoted with different size and color but similar marker belong to the same phonetic class).

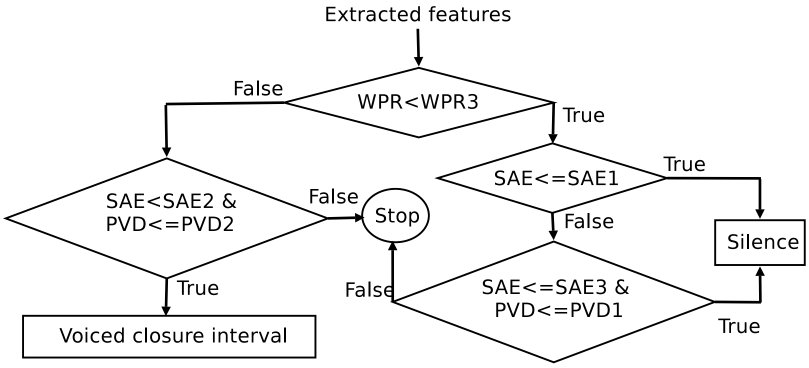

Figure B.2: The flow chart of the multi-threshold linear classifier.

Fig. B.2 shows operating flow chart of the linear classifier with the turned hard thresholds as $WPR3 = 99\%$, $PVD1 = 0.15$, $PVD2 = 0.35$, $SAE1 = 0.001/160$, $SAE2 = 0.025/160$ and $SAE3 = 0.016/160$. Subsequently, presmoothing is used to eliminate some wrong decisions.

Appendix C

Evaluating hypotheses

It has shown in [Mit97] that error rate E is a random variable, and it obeys the Binomial distribution. Thus, the standard deviation of the random variable E is estimated as:

$$\sigma_E = \sqrt{Np(1-p)} \tag{C.1}$$

where N, p are the parameters of the Binomial distribution. In the concept of evaluation, p is a misclassifying probability, and N is the number of instance in the test set $\mathcal{S}$. Actually we don't know p but we can approximate $p = r/N$ where r is the number of instance from $\mathcal{S}$ misclassified by the classifier. So, the standard deviation for the error rate E can be approximated as:

$$\sigma_E \approx \sqrt{\frac{(1-FR)FR}{N}} \tag{C.2}$$

where $FR = r/N$ is defined as a false rejection rate (%).

Appendix D

Subjective tests for speech enhancement

D.1 Comparison Diagnostic Test

Based on the Diagnostic Rhyme Test standard introduced by Fairbanks in 1985 [GMW97], we propose a comparison diagnostic test (CDT) on determined phonetic classes which are constructed from some groups of phonemes as follows:

Phonetic classes	Unvoiced fricatives	Voiced fricatives
Assigned phonemes	/f/,/th/,/s/,/sh/	/v/,/dh/,/z/,/zh/
Phonetic classes	Nasals	Stops
Assigned phonemes	/m/,/n/,/ng/	/b/,/d/,/g/,/p/,/t/,/k/

Table D.1: Assigned phonetic classes for evaluating with CDT.

As shown in Table D.1, the test materials are focused on consonants only because they are more problematic than vowels. From 40 processed sounds of AURORA3 data set, 8 utterances spoken by 4 female and 4 male speakers are selected randomly. There, each of 2 sounds is chosen from one of four different recording conditions (low speed, high speed, stop with operating engine, and running in town). One sounds is recorded with close-talking microphone, and another one is recorded by far-distance microphone. These 8 sounds are combined with 2 sounds chosen from 40 processed sounds of NTIMIT database to build a data set of 10 sounds for CDT evaluation. Then 10 utterances consisting of one word or several words which store unvoiced fricatives, voiced fricatives, nasals and plosives are artificially cut out of the whole sounds for listening purpose. The listeners hear the unprocessed and processed utterances, and concentrate on single phonetic classes that the corresponding texts are marked. Then they vote by Comparison Category Rating (ITU-T P.800) described in Table D.2.

Quality	Much better	Better	Slightly better	About the same
Scales	3	2	1	0
Quality	Slightly worse	Worse	Much worse	
Scales	-1	-2	-3	

Table D.2: Comparison Category Rating (ITU-T P.800)

D.2 Overall quality test ITU-T P.85

This isthe a proposal by ITU-T (1993) for comparative evaluation of overall quality. It consists of eight categorial estimation scales [GMW97]. However, for evaluating performance of speech enhancement algorithms, we selected only six suitable categories as follows:

- *a. Acceptance* (Do you think that this voice could be used for a telephone conversation?) 1: yes, 2: no.

- *b. Overall impression* (How do you rate the quality of the sound of what you have just heard?) 1:excellent, 2:good, 3:fair, 4:poor, 5: bad.

- *c. Listening effort* (How would you describe the effort you were required to make in order to understand the message?) 1:complete relaxation possible, no effort required, 2:attention necessary, no appreciable effort required, 3:moderate effort required, 4:effort required, 5:no meaning understood with any feasible effort.

- *d. Comprehension problems* (Did you find certain words hard to understand?) 1:never, 2:rarely, 3:occasionally, 4:often, 5:all of the time.

- *e. Articulation* (Were the sounds distinguishable?) 1:yes, very clear, 2:yes, clear enough, 3:fairly clear, 4:no, not very clear, 5:no, not at all.

- *f. Voice pleasantness* (How would you describe the voice?) 1:very pleasant, 2:pleasant, 3:fair, 4:unpleasant, 5:very unpleasant.

To do this subjective test, we construct a data set of 10 utterances which consists 8 and 2 recordings selected randomly from Aurora3 and NTIMIT databases, respectively.

For each test category, every listener is requested to listen the un-processed utterance (NP) and its four denoised utterances derived by five corresponding algorithms (SS, NSS, NSS-MM, SWF). Then he/she rates the scale for noisy utterance and denoised utterances. Listeners are allowed to listen recordings as many times as they wish. Test results are reported in Table D.3.

Algorithms		NP	SS	NSS	NSS-MM	SWF
Measures		M ± S	M ± S	M ± S	M ± S	M ± S
Terms	Utt.(dB)					
a.	A3 (5)	1.45 ± 0.27	2.00 ± 0.00	1.81 ± 0.16	1.55 ± 0.27	1.27 ± 0.42
	A3 (10)	1.18 ± 0.16	2.00 ± 0.00	1.54 ± 0.27	1.27 ± 0.22	1.00 ± 0.00
	NT (10)	1.00 ± 0.00	1.92 ± 0.07	1.71 ± 0.22	1.14 ± 0.13	1.14 ± 0.13
b.	A3 (5)	3.45 ± 0.87	4.64 ± 0.45	3.82 ± 0.56	2.73 ± 0.62	2.64 ± 0.65
	A3 (10)	2.82 ± 0.56	4.64 ± 0.25	3.73 ± 0.62	2.55 ± 0.87	2.27 ± 0.42
	NT (10)	2.57 ± 0.26	4.50 ± 0.27	3.71 ± 0.22	2.57 ± 0.26	2.07 ± 0.23
c.	A3 (5)	2.36 ± 0.65	3.27 ± 1.02	2.91 ± 0.69	2.09 ± 0.69	1.82 ± 0.56
	A3 (10)	1.73 ± 0.22	3.27 ± 0.62	2.36 ± 0.45	1.73 ± 0.62	1.55 ± 0.47
	NT (10)	2.07 ± 0.23	3.57 ± 0.73	2.79 ± 0.80	1.93 ± 0.53	1.71 ± 0.37
d.	A3 (5)	1.45 ± 0.47	2.27 ± 1.02	2.00 ± 1.20	1.45 ± 0.47	1.45 ± 0.47
	A3 (10)	1.27 ± 0.22	2.27 ± 0.62	1.82 ± 0.16	1.45 ± 0.47	1.18 ± 0.16
	NT (10)	1.64 ± 0.25	2.50 ± 0.73	1.93 ± 0.53	1.71 ± 0.22	1.57 ± 0.26
e.	A3 (5)	1.82 ± 0.76	3.55 ± 1.07	2.91 ± 0.49	2.18 ± 0.36	2.09 ± 0.49
	A3 (10)	2.64 ± 0.45	4.36 ± 0.45	3.63 ± 0.65	2.73 ± 0.62	2.45 ± 0.27
	NT (10)	2.50 ± 0.27	4.50 ± 0.42	3.50 ± 0.58	2.36 ± 0.86	2.29 ± 0.37
f.	A3 (5)	2.64 ± 0.45	4.36 ± 0.45	3.64 ± 0.65	2.73 ± 0.62	2.45 ± 0.27
	A3 (10)	2.36 ± 0.65	4.55 ± 0.27	3.45 ± 0.67	2.64 ± 0.85	2.18 ± 0.16
	NT (10)	2.30 ± 0.27	4.53 ± 0.42	3.32 ± 0.57	2.36 ± 0.86	2.1 ± 0.37

Table D.3: Detail performance in terms of mean (M) and standard deviation (S) of overall quality test ITU-T P.85 on utterances (Utt.) at different SegSNRs (dB) selected from Aurora3 (A3) and NTIMIT (NT) databases.

Appendix E

Robust speech recognition

E.1 Frequency weighting for hearing aid

As firstly proposed in [PK05b] to handle the correlated or colored noise, a frequency weighting η_k is applied on noise thresholds over all wavelet channels to enhance quality of denoised speech sound:

$$\eta_k = (a_0 \Gamma_k)^{-(a_0 \Gamma_k)^{b_0}} + d_0 \ , \tag{E.1}$$

where $a_0 = 10, b_0 = 0.55, d_0 = 0.6$ are constants and selected manually from informal listening tests.

Figure E.1 shows the weighted thresholds by applying the weighting function η_k in Equ. E.1. In general, the large coefficients which leads to large universal threshold, and thus large quantile thresholds (which are larger than 0.1) should be served by less amplified thresholds. Whereas, the small coefficients which leads to small universal threshold, and thus small quantile thresholds (in a range of $[0.0, \ldots, 0.1]$ as depicted in Fig. E.1) should be best thresholded with high thresholds. Though, from our pre-formal listening, the very low threshold (in a range $[0.0, \ldots, 0.05]$) should not be thresholded by too high threshold in order to keep remained background noise at a certain low level which creates comfortable sound. That's why we end up with the formula in Eq. E.1. This function results in low distortion of enhanced speech signal while keeping background noise remained at a certain level.

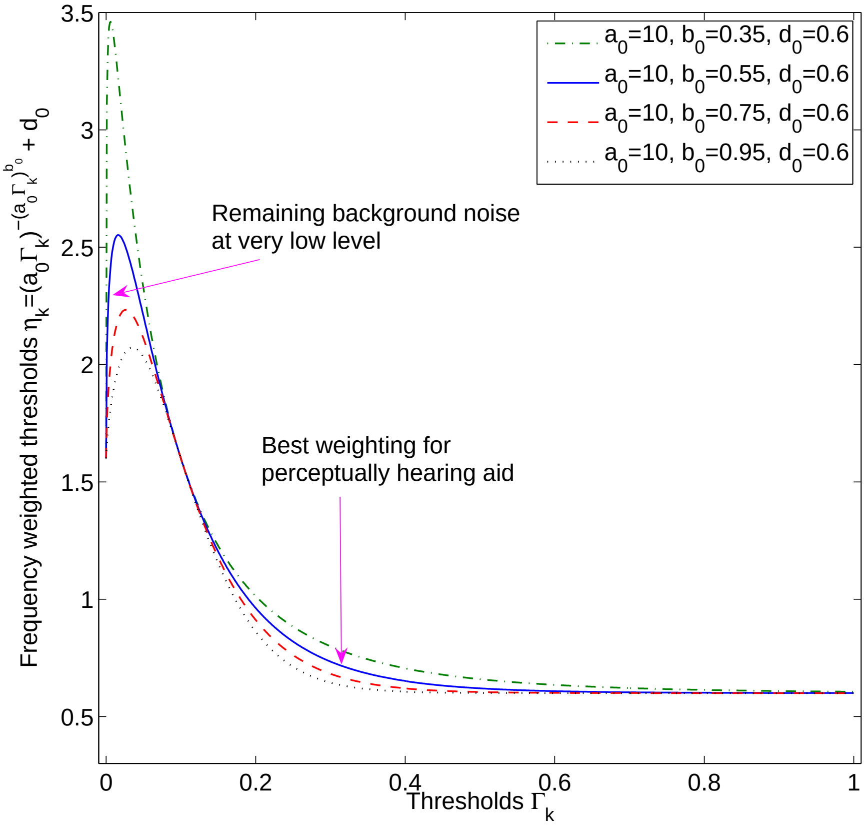

Figure E.1: Frequency non-linear weighting function.

E.2 Parametric turning of the time-frequency weighting for ASR

With the perceptual SWF method, we again turn parameters of time and frequency weighting functions in order to obtain high recognition performance for different used front-ends in two different modes: without and with retraining acoustic model. The quantile threshold Γ_k derived for every channel k^{th} atThe network configuration corresponding to experiment number 2 could be a best candidate. each frame i^{th} in a certain buffer is further estimated by:

$$\tilde{\Gamma}_{k,i} = \lambda_{k,i}\eta_k^{SR}\beta_k\Gamma_k, \tag{E.2}$$

where $\lambda_{k,i}$, η_k^{SR} are adaptive weighting functions in time and frequency domains, respectively, β_k is introduced as an overestimate factor for quantile threshold Γ_k, $\tilde{\Gamma}_{k,i}$ is the weighted estimate of the quantile threshold. The temporal weighting is formulated as:

$$\lambda_{k,i} = (a_2 T_{k,i})^{-b_2} + d_2 \ , \tag{E.3}$$

and the frequency weighting that its shape was changed for ASR is reused:

$$\eta_k^{SR} = (a_4 \Gamma_k)^{b_4} + d_4 \ . \tag{E.4}$$

In general, all three terms - time weighting, frequency weighting and overestimation factor - involves to recognition performance as mentioned previously in chapter 5. We start the turning experiment with the previous configuration used for the SWF ($a_2 = 10$, $b_2 = 1$, $d_2 = 2$, $a_4 = 70$, $b_4 = 0.5$, $d_4 = 0.4$, and $\beta_k = 1$) which provides performance reported in Table 5.2. From our first trials, we found that besides the overestimation factor β_k, the frequency non-linear weighting mostly affects the performance of ASR in terms of number of delected words (D), number of substituted words (S), and number of inserted words (I). Especially, the factors a_4 shows a parabolic tendency of these measures. This means when keeping all other factors constant and varying only a_4, a local minimum is found from obtained values of each measure as shown later. This behavior does not occur with moderate changes in other factors. Figures E.2 to E.5 presents tendencies of D, S and I when varying factor a_4 while keeping other factors constant as $a_2 = 10$, $b_2 = 1$, $d_2 = 2$, $b_4 = 0.5, d_4 = 0$, and $\beta_k = 1$. Based on this observation, the value of factor a_4 is selected to obtain minimum number of D+S+I. After that, the overestimation factor is further adjusted to obtain better recognition performance. The sets of final values of factors which achieve highest WRR and ACC for different used front-ends in different modes are reported in Table E.1. A detailed procedure is reported in [GP06].

Table E.1: Best configurations of factors for different front-ends in without (w.o.) and with (w.) retraining modes.

Factors		a_2	b_2	d_2	a_4	b_4	d_4	β_k
Front-ends	Modes							
SFE	w.o.	10	1	2	150	0.5	0	0.5
SFE	w.	10	0.8	2	95	0.5	0	0.5
AFE	w.o.	10	1	2	150	0.5	0	0.2
AFE	w.	10	1	2.5	64	0.5	0	0.5

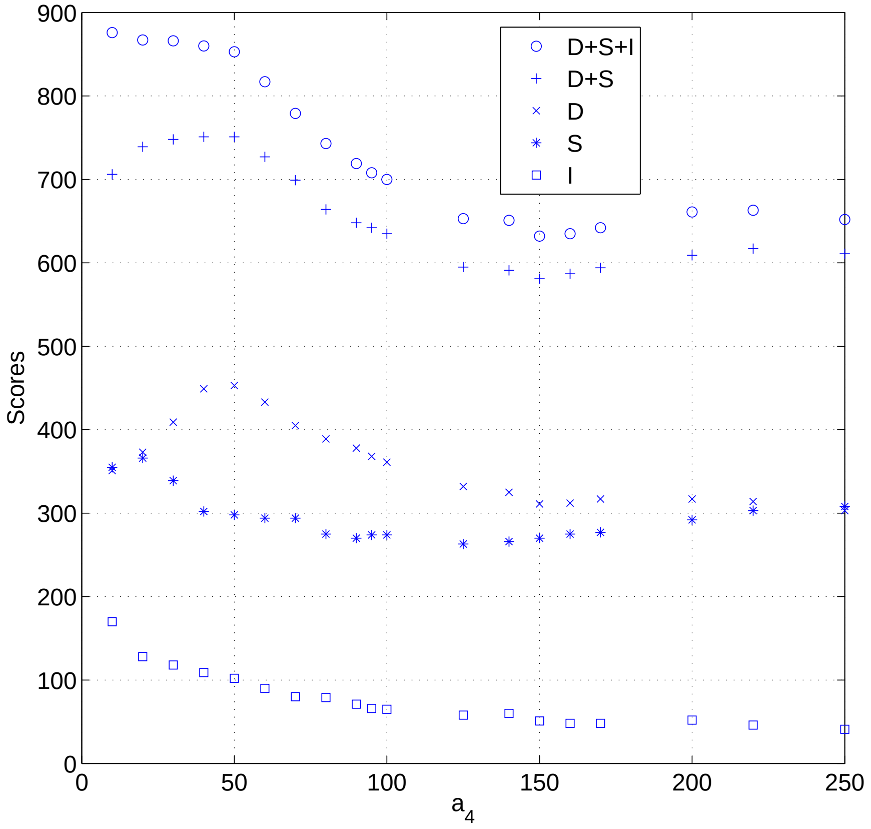

Figure E.2: Tendencies of D/S/I obtained with SFE in without retraining mode, when varying factor a_4 of frequency weighting η_k^{SR}.

E.3 Recognition performance

Table E.2 to E.4 reports the test results derived from the baseline (without denoising for SFE, with WF for AFE), and from the use of our proposed PSWF as a pre-processing for SFE and as a replacement for WF in AFE.

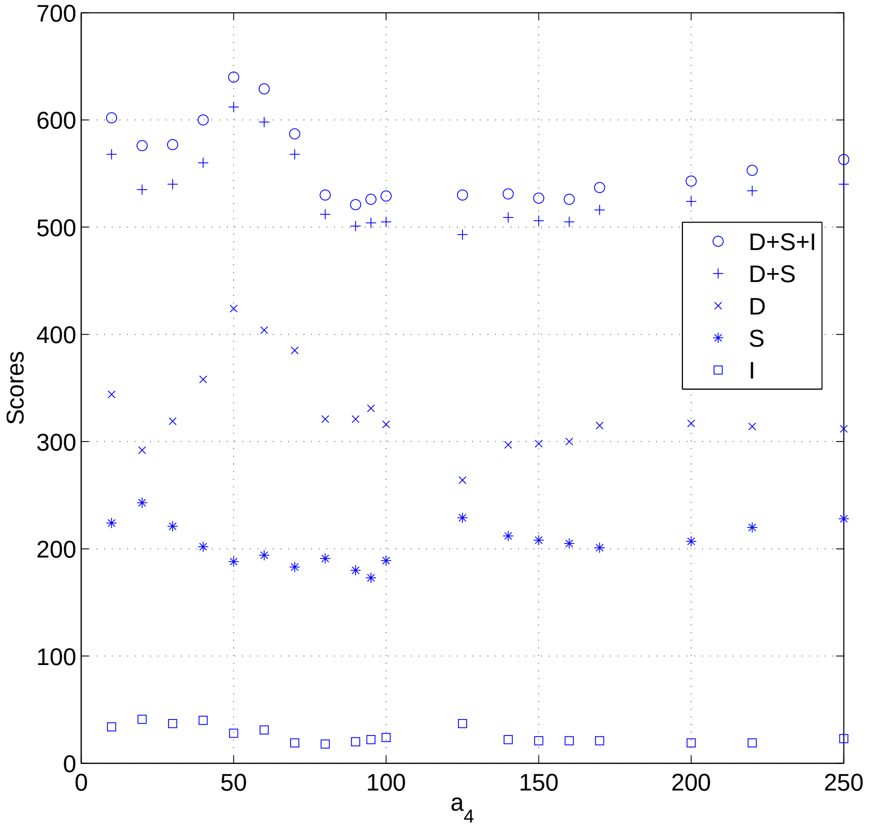

Figure E.3: Tendencies of D/S/I obtained with SFE in retraining mode, when varying factor a_4 of frequency weighting η_k^{SR}.

Table E.2: Recognition performance obtained from the baseline.

Recognizers	German SpeechDat-Car/HTK			
Algorithms	Baseline		Baseline	
Front-ends	SFE		AFE	
Measures	WRR/ACC/SRR	D/S/I	WRR/ACC/SRR	D/S/I
Conditions				
hm	66.70/63.23/38.83	443/277/75	89.78/89.45/72.08	121/100/7
mm	78.48/76.43/52.70	149/145/28	89.53/89.02/70.54	59/84/7
wm	90.48/87.92/67.89	203/374/128	95.55/94.65/82.05	87/136/45

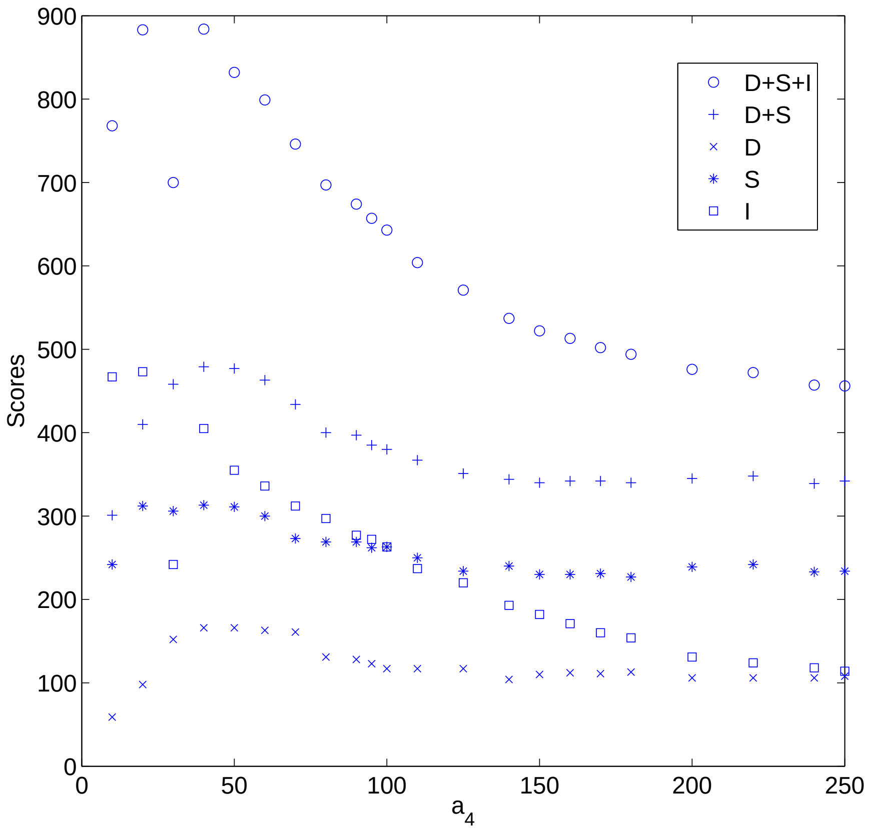

Figure E.4: Tendencies of D/S/I obtained with AFE in without retraining mode, when varying factor a_4 of frequency weighting η_k^{SR}.

Table E.3: Recognition performance obtained from standard front-end SFE.

Recognizers	German SpeechDat-Car/HTK			
Algorithms	PSWF		PSWF	
Mode	without retraining		with retraining	
Measures	WRR/ACC/SRR	D/S/I	WRR/ACC/SRR	D/S/I
Conditions				
hm	73.13/70.77/48.73	311/270/51	77.71/76.73/58.12	280/202/21
mm	69.03/56.88/30.71	191/232/166	81.92/78.99/52.28	75/172/40
wm	90.08/84.33/59.31	142/355/288	92.91/91.20/72.91	117/238/86

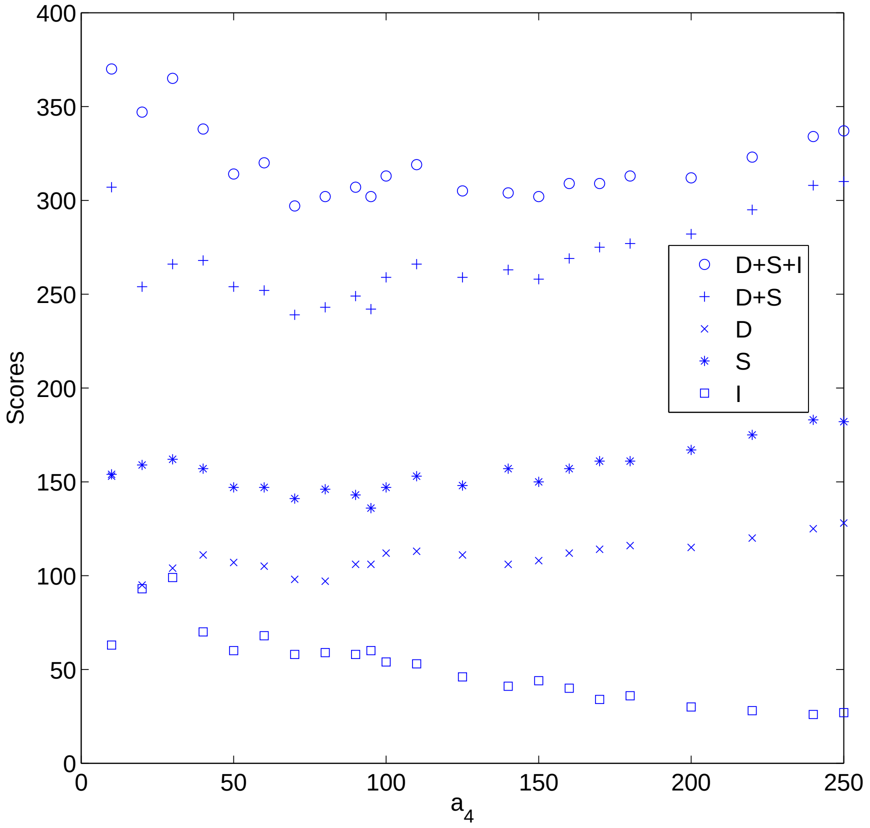

Figure E.5: Tendencies of D/S/I obtained with AFE in retraining mode, when varying factor a_4 of frequency weighting η_k^{SR}.

Table E.4: Recognition performance obtained from advanced front-end AFE.

Recognizers	German SpeechDat-Car/HTK			
Algorithms	PSWF		PSWF	
Mode	without retraining		with retraining	
Measures	WRR/ACC/SRR	D/S/I	WRR/ACC/SRR	D/S/I
Conditions				
hm	86.63/68.87/34.77	59/230/384	89.45/86.63/64.21	87/141/61
mm	86.75/57.83/19.92	25/156/395	88.65/85.29/63.49	41/114/46
wm	94.29/81.31/52.40	49/237/650	95.07/93.25/78.37	61/186/91

Bibliography

[AC99] A. Agarwal and Y. M. Cheng. Two-stage mel-warped wiener filter for robust speech recognition. In *Proceedings of the International Workshop on Automatic Speech Recognition and Understanding*, pages 67–70, 1999.

[Ace90] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* PhD thesis, Carnegie Mellon University, 1990.

[Agb96] J. I. Agbinya. Discrete wavelet transform techniques in speech processing. In *Proceedings of the Region Ten Conference (TENCON) - Digital Signal Processing Applications*, volume 2, pages 514–519, 1996.

[AH01] A. N. Akansu and R. A. Haddad. *Multiresolution Signal Decomposition: Transform, Subbands, and Wavelets.* Academic Press, UK, 2001.

[AK05] D. Arifianto and Takao Kobayashi. Voiced/unvoiced determination of speech signal in noisy environment using harmonicity measure based on instantaneous frequency. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 877–880, 2005.

[AR76] B. Atal and L. R. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):201–212, 1976.

[AR82] B. S. Atal and J. R. Remde. A new model of LPC excitation for producing natural-sounding speech at low bit rates. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 614–617, 1982.

[AS84] L. Almeida and F. Silva. Variable-frequency synthesis: an improved harmonic coding scheme. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 2751–2754, 1984.

[AS99] S. Ahmadi and A.S. Spanias. Cepstrum-based pitch detection using a new statistical v/uv classification algorithm. *IEEE Transactions on Speech and Audio Processing*, 7(3):333–338, 1999.

[ASS98] F. Abramovich, T. Sapatinas, and B. W. Silverman. Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society*, 60(4):725–749, 1998.

[aur01] AURORA Project Database - Subset of SpeechDat-Car German database (AURORA/CD0003-03). Technical report, Evaluations and Language resources Distribution Agency, 2001.

[BE97] E. Bernstein and W. Evans. Wavelet based noise reduction for speech recognition. In *Proceedings of the Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 111–114, 1997.

[Bea00] F. Bimbot and et al. Speaker verification using adapted gaussian mixture models. In *Digital Signal Processing*, number 10, pages 19–41, 2000.

[BEK00] M. Baum, G. Erbach, and G. Kubin. SPEECHDAT-AT: A telephone speech database for Austrian German. LREC Workshop Very Large Telephone Databases, 2000.

[BG95] A. Bruce and H. Y. Gao. WaveShrink: Shrinkage functions and thresholds. In *Proceedings of the SPIE Conference on Wavelet Applications in Signal and Image Processing*, volume 2569, pages 270–281, 1995.

[BG96] A. G. Bruce and H.Y. Gao. Understanding WaveShrink: Variance and bias estimation. *Biometrika*, 83(4):727–745, 1996.

[BG97] A. G. Bruce and H. Y. Gao. WaveShrink with firm shrinkage. *Statistica Sinica*, 7:855–874, 1997.

[BG02] D. Burshtein and S. Gannot. Speech enhancement using a mixture-maximum model. *IEEE Transactions on Speech and Audio Processing*, 10(6):341–351, 2002.

[BJF94] R. L. Bouquin-Jeannes and G. Faucon. Proposal of a voice activity detector for noise reduction. *IEE Electronics Letters*, 30:930–932, 1994.

[BJF95] R. L. Bouquin-Jeannes and G. Faucon. Study of a voice activity detector and its influence on a noise reduction system. *Speech Communication*, 16:245–254, 1995.

[BM03] C. Breithaupt and R. Martin. MMSE estimation of magnitude-squared DFT coefficients with supergaussian priors. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 896–899, 2003.

[Bol79] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27:113–120, 1979.

[BR01] M. Bahoura and J. Rouat. Wavelet speech enhancement based on the teager energy operator. In *IEEE Signal Processing Letter*, volume 8, pages 10–12, 2001.

[Bra68] P. T. Brady. A statistical analysis of on-off patterns in 16 conversations. *The Bell Systems Technical Journal*, 47(1):73–91, 1968.

[BSM79] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 208–211, 1979.

[BSN03] F. Basbug, K. Swaminathan, and S. Nandkumar. Noise reduction and echo cancellation front-end for speech codecs. *IEEE Transactions on Speech and Audio Processing*, 11(1):1–13, 2003.

[BSS$^+$97] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit. ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *IEEE Communications Magazine*, 35(9):64–73, 1997.

[Cap94] O. Cappe. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, 2:345–349, 1994.

[CD99] B. Carnero and A. Drygajlo. Perceptual speech coding and enhancement using frame-synchronizedfast wavelet packet transform algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 47:1622–1635, 1999.

[CG05] G.F. Choueiter and J.R. Glass. A wavelet and filter bank framework for phonetic classification. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 933–936, 2005.

[Che99] R. Chengalvarayan. Robust energy normalization using speech/non-speech discriminator forGerman connected digit recognition. In *Proceedings of Eurospeech*, pages 61–64, 1999.

[Chi00] D. G. Childers. *Speech processing and synthesis toolboxes.* John Wiley & Sons, USA, 2000.

[CHL89] D. G. Childers, M. Hahn, and J. N. Larar. Silence and voiced/unvoiced/mixed excitation classification of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1771–1774, 1989.

[CK01] Y. D. Cho and A. Kondoz. Analysis and improvement of a statistical model-based voice activity detector. In *IEEE Signal Processing Letter*, volume 8, pages 276–278, 2001.

[CKM97] H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch. Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, 92:1413–1421, 1997.

[CKYK02] S. Chang, Y. Kwon, S. Yang, and I. Kim. Speech enhancement for non-stationary noise environment by adaptive wavelet packet. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 561–564, 2002.

[CMU06] CMU - Speech Group, School of Computer Science, Carnegie Mellon University. *Learning to use the CMU SPHINX Automatic Speech Recognition system*, 2006. http://www.speech.cs.cmu.edu/sphinx/tutorial.html.

[CO91] Y. M. Cheng and D. O'Shaughnessy. Speech enhancement based conceptually on auditory evidence. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 39(9):1943–1954, 1991.

[Coh03] I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5):466–475, 2003.

[CW02] S. H. Chen and J. F. Wang. A wavelet-based voice activity detection algorithm in noisy environments. In *Proceedings of the International Conference on Electronics, Circuits and Systems*, volume 3, pages 995–998, 2002.

[CW04] S. H. Chen and J. F. Wang. Speech enhancement using perceptual wavelet packet decomposition and teager energy operator. *Journal of VLSI Signal Processing Systems*, 36(2):125–139, 2004.

[Dav02] G. M. Davids. Noise reduction in speech applications. *CRC Press*, USA, 2002.

[DB02] H. Demuth and M. Beale. *Neural Network Toolbox.* The MathWorks, 2002.

[DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[DJ95] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[DJ98] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(3):879–921, 1998.

[DJC03] O. Donnellan, E. Jung, and E. Coyle. Speech-adaptive time-scale modification for computer assisted language-learning. In *on Advanced Learning Technologies*, pages 165–169, 2003.

[DM80] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition. *Trans. Acoust., Speech, Signal Processing*, Vol. 28, pp. 357-366, Aug. 1980.

[Don95] D. L. Donoho. De-noising by soft thresholding. *IEEE Trans. Information Theory*, 41:613–627, 1995.

[Dru68] H. Drucker. Speech processing in a high ambient noise environment. *IEEE Transactions on Audio Electroacoustics*, 16(2):165–168, 1968.

[DS94] L. Deng and . D. Sun. Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of english sounds. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 45–48, 1994.

[EM83] Y. Ephraim and D. Malah. Speech enhancement using optimal non-linear spectral amplitude estimation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 1118–1121, USA, 1983.

[EM84] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32:1109–1121, 1984.

[EM85] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33:443–445, 1985.

[ENS02] F. B. Ertem, S. Nandkumar, and K. Swaminathan. Method of noise reduction for speech codecs. United States Patent 6453289, September 2002.

[Eph92a] Y. Ephraim. A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Transactions on Signal Processing*, 40:725–735, 1992.

[Eph92b] Y. Ephraim. Statistical-model-based speech enhancement systems. In *Proceedings of the IEEE*, volume 80, pages 1526–1555, 1992.

[ET95] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 3:251–266, 1995.

[ets99] GSM 06.94. Digital cellular telecommunication system (Phase 2+); voice activity detector for adaptive multi rate (AMR) speech traffic channels. General description, tech. rep. v.7.0.0, ETSI, 1999.

[ETS00] ETSI. *ETSI ES 201 108 V1.1.1 Speech Processing, Transmission and Quality Aspects (STQ), Distributed speech recognition, Front-end feature extraction algorithm, Compression algorithms*, 2000.

[ETS03] ETSI. *ETSI ES 202 050 V1.1.3 Speech Processing, Transmission and Quality Aspects (STQ), Distributed speech recognition, Advanced front-end feature extraction algorithm, Compression algorithms*, 2003.

[Eva01] N. W. D. Evans. Noise estimation without explicit speech, non-speech detection: a comparison of mean, modal and median based approaches. In *Proceedings of Eurospeech*, pages 893–896, 2001.

[FD01] O. Farooq and S. Datta. Mel filter-like admissible wavelet packet structure for speech recognition. In *IEEE Signal Processing Letter*, volume 8, pages 196–198, 2001.

[FD03] O. Farooq and S. Datta. Wavelet-based denoising for robust feature extraction for speech recognition. *Electronics Letters*, 39(1):163–165, 2003.

[FDGM$^+$93] W. M. Fisher, G. R. Doddington, K. M. Goudie-Marshall, C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. NTIMIT. Technical report, Linguistic Data Consortium, University of Pennsylvania, 1993. http://www.ldc.upenn.edu/.

[FGG97] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[FI93] U.M. Fayyad and K.B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993.

[FSBO76] R.H. Frazier, S. Samsan, L.D. Braida, and A.V. Oppenheim. Enhancement of speech by adaptive filtering. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 251–253, 1976.

[FW03] Q. Fu and E. A. Wan. Perceptual wavelet adaptive denoising of speech. In *Proceedings of Eurospeech*, pages 577–580, 2003.

[GAE04] T. S. Gunawan, E. Ambikairajah, and J. Epps. Perceptual wavelet packet audio coder. In *Proceedings of International Conference on Spoken Language Processing*, pages 283–286, 2004.

[GAM99] R. Gemello, D. Albesano, and F. Mana. Multi-source neural networks for speech recognition. In *International Joint Conference on Neural Networks (IJCNN)*, 1999.

[GCEJ91] T. Ghiselli-Crippa and A. El-Jaroudi. Voiced-unvoiced-silence classification of speech using neural nets. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2, pages 851–856, 1991.

[GG01] M. Gupta and A. Gilbert. Robust speech recognition using wavelet coefficient features. In *Proceedings of the International Workshop on Automatic Speech Recognition and Understanding*, 2001.

[GGPP93] J. Garofalo, D. Graff, D. Paul, and D. Pallett. Continous Speech Recognition (CSR-I) Wall Street Journal (WSJ0) news, complete. Linguistic Data Consortium, Philadelphia, 1993. http://ldc.upenn.edu/Catalog/.

[GKG91] J. D. Gibson, B. Koo, and S. D. Gray. Filtering of colored noise for speech enhancement and coding. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 39:1732–1742, 1991.

[GL88] D. Grin and J. Lim. Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(8), 1988.

[GLF+93] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. TIMIT acoustic - phonetic - continuous speech corpus. Technical report, National Institute of Standards and Technology (NIST), 1993. http://www.ldc.upenn.edu.

[GMM04] R. Gemello, F. Mana, and R. D. Mori. A modified Ephraim-Malah noise suppression rule for automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 957–960, 2004.

[GMW97] D. Gibbon, R. Moore, and R. Winski. Handbook of standards and resources for spoken language systems. *Mouton de Gruyter*, Berlin & New York, 1997.

[GP06] K. Gupta and T. V. Pham. Parametric optimization of statistical wavelet filtering for robust ASR. Technical report, Signal Processing and Speech Communication Laboratory, Graz University of Technology, August 2006.

[Gro00] JPEG Group. *Guide to the practical implementation of JPEG 2000*, 2000. http://www.jpeg.org/jpeg2000/index.html.

[GT00] J. N. Gowdy and Z. Tufekci. Mel-scaled discrete wavelet coefficients for speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1351–1354, 2000.

[GY96] M. J. F. Gales and S. J. Young. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, 4(5):352–359, 1996.

[GZ02] R. Greiner and W. Zhou. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. In *Proceedings of the 18th National conference on Artificial Intelligence*, pages 167–173, 2002.

[HB96] A.J. Hunt and A.W. Black. Unit selection in a concatenative speech synthesis system using alarge speech database. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 373–376, 1996.

[HC95] B. M. Helf and P. L. Chu. Reduction of background noise for speech enhancement. United States Patent 5550924, March 1995.

[Hil04] J. M. Hilario. Discriminative connectionist approaches for automatic speech recognition in cars. *PhD thesis*, Cottbus University of Technology, Germany, Aug. 2004.

[HM93] J. A. Haigh and J. S. Mason. Robust voice activity detection using cepstral features. In *Proceedings of the Region Ten Conference (TENCON) - Digital Signal Processing Applications*, pages 321–324, 1993.

[HN97] J. Hernando and C. Nadeu. Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. *IEEE Transactions on Speech and Audio Processing*, 5(1):80–84, 1997.

[HT90] S. Teager H. Teager. Evidence for nonlinear production mechanisms in the vocal tract. In *Speech Production and Speech Modeling*, volume 55, pages 241–261. NATO Advanced Study Institute, Kluwer Academic Pub., 1990.

[Ita75] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23:67–72, 1975.

[Jan00] M. Jansen. *Noise reduction by wavelet thresholding.* PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2000.

[JB99] M. Jansen and A. Bultheel. *Bayesian inference in wavelet based models*, chapter Geometrical priors for noise-free wavelet coefficient configurations in image de-noising, pages 223–242. Springer-Verlag, 1999.

[JC03] F. Jabloun and B. Champagne. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 11(6):700–708, 2003.

[JFF99] J. C. Junqua, S. Fincke, and K. Field. The Lombard effect: a reflex to better communicate with others in noise. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 2083-2086, 1999.

[JM03] E. Jafer and A. E. Mahdi. Wavelet-based perceptual speech enhancement using adaptive threshold estimation. In *Proceedings of Eurospeech*, pages 569–572, 2003.

[Joh88] J.D. Johnson. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal Selected Areas in Communication*, 6(2):314–323, 1988.

[Joh99] I. M. Johnstone. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. In *Technical Report*, Stanford University, 1999.

[Jun93] J. C. Junqua. The Lombard reflex and its role on human listerners and automatic speech recognizers. *The Journal of the Acoustical Society America*, 93(1):510–524, 1993.

[Kai90a] J. F. Kaiser. On a simple algorithm to calculate the energy of a signal. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 381–384, 1990.

[Kai90b] J.F. Kaiser. On Teager's energy algorithm, its generalization to continuous signals. In *Proceedings of the Workshop on Digital Signal Processing*, 1990.

[KAK93] G. Kubin, B.S. Atal, and W.B. Kleijn. Performance of noise excitation for unvoiced speech. In *Proceedings of the Workshop on Speech Coding for Telecommunications*, 1993.

[Ked86] B. Kedem. Spectral analysis and discrimination by zero-crossings. In *Proceedings of the IEEE*, volume 74, pages 1477–1493, 1986.

[KGG89] B. Koo, J. D. Gibson, and S. D. Gray. Filtering of colored noise for speech enhancement and coding. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 349–352, 1989.

[KK94] G. Kubin and W.B. Kleijn. Time-scale modification of speech based on a nonlinear oscillator model. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 453–456, 1994.

[KKH03] B. Kotnik, Z. Kacic, and B. Horvat. The usage of wavelet packet transformation in automatic noisy speech recognition systems. In *Proceedings of the International Conference on "Computer as a tool"*, volume 2, pages 131–134, 2003.

[Kle93] W.B. Kleijn. Encoding speech using prototype waveforms. *IEEE Transactions on Speech and Audio Processing*, 1(4):386–399, 1993.

[Kot04] B. Kotnik. *Robust speech parameterization based on joint wavelet packet decomposition and autoregressive modeling.* PhD thesis, Maribor University, Slovenia, 2004.

[KPK$^+$06] M. Kepesi, T. V. Pham, G. Kubin, L. Weruaga, A. Juffinger, and M. Grabner. Noise cancellation frontends for automatic meeting transcription. In *Proceedings of the European Conference on Noise Control*, 2006.

[KPN06] M. Kepesi, T. V. Pham, and M. Neffe. Audio unit results. MISTRAL Project homepage, December 2006.

[KYK01] I-Jae Kim, Sung-Il Yang, and Y. Kwon. Speech enhancement using adaptive wavelet shrinkage. In *Proceedings of the International Symposium on Industrial Electronics*, volume 1, 2001.

[LA06] Y. C. Lee and S. S. Ahn. Statistical model-based vad algorithm with wavelet transform. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E89-A:1594–1600, 2006.

[Lav06] I. Lavrik. *Novel wavelet-based statistical methods with applications in classification, shrinkage, and nano-scale image analysis.* PhD thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, 2006.

[LCG93] H.C. Leung, B. Chigier, and J.R. Glass. A comparative study of signal representations and classification techniques for speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 657–664, 1993.

[LE03] Z. Lachiri and N. Ellouze. Speech classification in noisy environment using subband decomposition. In *Proceedings of the International Symposium on Signal Processing and its Applications*, volume 1, pages 409–412, 2003.

[LG99] L. Liao and M. A. Gregory. Algorithms for speech classification. In *ISSPA*, pages 623–627, 1999.

[LGG04] A. Lallouani, M. Gabrea, and C.S. Gargour. Wavelet based speech enhancement using two different threshold-based denoising algorithms. In *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, pages 315–318, 2004.

[LHR90] K.-F. Lee, H.-W. Hon, and R. Reddy. An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(1):35–44, 1990.

[Lit98] Jr Litwin, L.R. Speech coding with wavelets. *IEEE Potentials*, 17(2):38–41, 1998.

[LO79] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. In *Proceedings of the IEEE*, volume 67, pages 1586–1604, 1979.

[Loq06] Loquendo. Loquendo ASR brochure, visited: Jan. 2006. http://www.loquendo.com.

[LT71] H. Lane and B. Tranel. The Lombard sign and the role of hearing in speech. *Journal of Speech Hearing Research*, 14:677–709, 1971.

[LV03] T. Lotter and P. Vary. Noise reduction by maximum a posteriori spectral amplitude estimation with supergaussian speech modeling. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control*, pages 83–86, 2003.

[LW03] C.-T. Lu and H.-C. Wang. Enhancement of single channel speech based on masking property and wavelet transform. *Speech Communication*, 41(3):409–427, 2003.

[Mal89] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet presentation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 11:674–693, 1989.

[Mal99] Stephane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[Mar93] R. Martin. An efficient algorithm to estimate the instantaneous SNR of speech signals. In *Proceedings of Eurospeech*, pages 1093–1096, 1993.

[Mar94] R. Martin. Spectral subtraction based on minimum statistics. In *Proceedings of Eurospeech*, pages 1182–1185, 1994.

[Mar01] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9:504–512, 2001.

[Mar02] R. Martin. Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 253–256, 2002.

[Mar05] R. Martin. Speech enhancement based on minimum square error estimation and supergaussian priors. *IEEE Transactions on Speech and Audio Processing*, 13(5):845–856, 2005.

[MBB96] J. Minghu, Y. Baozong, and L. Biquin. The consonant/vowel speech classification using high-rank function neural network. In *Proceedings of the International Conference on Signal Processing*, volume 2, pages 1469–1472, 1996.

[MCGB01] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet. Overview of the 2000-2001 ELISA consortium research activities. pages 67–72, June 2001.

[MDK$^+$97] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech*, pages 1895–1898, 1997.

[MH92] S. Mallat and W. L. Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38(2):617–643, 1992.

[mis] Measurable intelligent and reliable semantic extraction and retrieval of multimedia data (MISTRAL Project). http://mistral-project.tugraz.at/.

[Mit97] T. M. Mitchell. *Machine Learning*. The McGraw-Hill, 1997.

[MKQ93] P. Maragos, J.F. Kaiser, and T.F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on Signal Processing*, 41(10):3025–3051, 1993.

[MM80] R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(2):137–145, 1980.

[MMOP02] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi. *Wavelet Toolbox*. The MathWorks, 2002.

[MQ86] R. McAulay and T. Quatieri. Speech analysis-synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:744–754, 1986.

[NGM01] E. Nemer, R. Goubran, and S. Mahmoud. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing*, 9(3):217–231, March 2001.

[NPHKon] M. Neffe, T. V. Pham, H. Hering, and G. Kubin. *Lecture Notes in Artificial Intelligence*, chapter Speaker Segmentation for Air Traffic Control. Springer, accepted for publication.

[NPK07] M. Neffe, T. V. Pham, and G. Kubin. Robust speaker verification in air traffic control using improved voice activity detection. In *Proceedings of the International Conference on Signal Processing, Pattern Recognition and Applications*, 2007.

[NS02] P. Niyogi and M. M. Sondhi. Detecting stop consonant in continuous speech. *The Journal of the Acoustical Society America*, 111:1063–1076, 2002.

[NSJ$^+$01] B. Noe, J. Sienel, D. Jouvet, L.Mauuary, L. Boves, J. de Veth, and F. de Wet. Noise reduction for noise robust feature extraction for distributed speech recognition. In *Proceedings of Eurospeech*, pages 433–436, 2001.

[OGC93] J. P. Olive, A. Greenwood, and J. Coleman. *Acoustic of American English.* Springer, 1993.

[Ohs93] Y. Ohshima. *Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing.* PhD thesis, Carnegie Mellon University, 1993.

[O'S00] D. O'Shaughnessy. *Speech Communications: Human and Machine.* IEEE Press, 2000.

[Par76] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *The Journal of the Acoustical Society America*, 60:911–918, 1976.

[PB06] F. Pernkopf and J. Bilmes. Ordering-based discriminative structure learning for Bayesian network classifiers. In *submitted*, 2006.

[Pea88] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

[Pea00] David Pearce. An overview of the ETSI standards activities for distributed speech recognition front-ends. Technical report, Motorola Labs & ETSI STQ-Aurora DSR Working Group, 2000. http://www.cavs.msstate.edu/hse/ies/projects/aurora/.

[PH00] D. Pearce and H.-G Hirsch. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of International Conference on Spoken Language Processing*, volume 4, pages 29–32, 2000.

[Pha05] T. V. Pham. Demonstration of denoised sounds. Signal Processing and Speech Communication Lab, September 2005. http://www.spsc.tugraz.at/people/tuan/SE.

[PK04] T. V. Pham and G. Kubin. DWT-based classification of acoustic-phonetic classes and phonetic units. In *Proceedings of International Conference on Spoken Language Processing*, pages 985–988, 2004.

[PK05a] T. V. Pham and G. Kubin. DWT-based phonetic groups classification using neural networks. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 401–404, 2005.

[PK05b] T. V. Pham and G. Kubin. WPD-based noise suppression using nonlinearly weighted threshold quantile estimation and optimal wavelet shrinking. In *Proceedings of Interspeech*, pages 2089–2092, 2005.

[PK06a] T. V. Pham and G. Kubin. Comparison of models using Time-Frequency features for speech classification. In *Proceedings of the International Conference on Research, Innovation and Vision For The Future*, pages 117–125, 2006.

[PK06b] T. V. Pham and G. Kubin. Low-complexity and efficient classification of voiced/unvoiced/silence for noisy environments. In *Proceedings of Interspeech*, pages 661–664, 2006.

[PKW$^+$06] T. V. Pham, M. Kepesi, L. Weruaga, G. Kubin, M. Sigmund, and T. Dostal. Time-frequency analysis for voice activity detection. In *Proceedings of the International Conference on Signal Processing, Pattern Recognition and Applications*, pages 244–249, 2006.

[PM98] M. Przybocki and A. Martin. NIST speaker recognition evaluations. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 331–335, 1998.

[PNK07] T. V. Pham, M. Neffe, and G. Kubin. Robust voice activity detection for improving narrow-band speaker verification. In *unpublished*, 2007.

[PP06a] F. Pernkopf and T. V. Pham. Bayesian networks for phonetic classification using time-scale features. In *Proceedings of Interspeech*, pages 2198–2201, 2006.

[PP06b] F. Pernkopf and T. V. Pham. Discriminative Bayesian networks for phonetic classification using time-scale features. *Submitted to IEEE Transactions on Audio, Speech, and Language Processing*, 1(12), 2006.

[PRK07] T. V. Pham, E. Rank, and G. Kubin. Perceptual statistical wavelet filtering for robust speech recognition in adverse environments. *in preparation*, 2007.

[PY04] S.R.M. Prasanna and B. Yegnanarayana. Extraction of pitch in adverse conditions. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 109–112, 2004.

[QH93] Y. Qi and B. R. Hunt. Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. *IEEE Transactions on Speech and Audio Processing*, 1:250–255, 1993.

[RHM$^+$02] M. Rahurkar, J.H.L. Hansen, J. Meyerhoff, G. Saviolakis, and M. Koenig. Frequency band analysis for stress detection using a Teager energy operator based feature. In *Proceedings of International Conference on Spoken Language Processing*, volume 3, pages 2021–2024, 2002.

[RJ93] L. Rabiner and B. H. Juang. *Fundamental of Speech Recognition.* Prentice-Hall, 1993.

[RPK06] E. Rank, T. V. Pham, and G. Kubin. Noise suppression based on wavelet packet decomposition and quantile noise estimation for robust automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 477–480, 2006.

[RS97] L. R. Rabiner and M. R. Sambur. Voiced-unoiced-silence detection using Itakura LPC distance measure. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 323–326, 1997.

[RSB$^+$04] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, and A. Rubio. A new voice activity detector using subband order-statistics filters for robust speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 849–852, 2004.

[RSB$^+$05] J. Ramirez, J.C. Segura, C. Benitez, A. de la Torre, and A. Rubio. An effective subband OSF-based VAD with noise reduction for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(6):1119–1129, 2005.

[SA85] M. R. Schroeder and B. S. Atal. Code-excited linear prediction CELP: High quality speech at very low bit rates. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 937–940, 1985.

[SA96] E. P. Simoncelli and E. H. Adelson. Noise removal via Bayesian wavelet coring. In *Proceedings of the International Conference on Image Processing*, 1996.

[SA01] H. Sheikhzadeh and H. R. Abutalebi. An improved wavelet-based speech enhancement system. In *Proceedings of Eurospeech*, pages 1855–1858, 2001.

[San98] H. Sanneck. Concealment of lost speech packets using adaptive packetization. In *Proceedings of the International Conference on Multimedia Computing and Systems*, pages 140–149, 1998.

[SB97] J. W. Seok and K. S. Bae. Speech enhancement with reduction of noise components in the wavelet domain. In *Proceedings of the International*

Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 1323–1326, 1997.

[SCJ$^+$02] A. Sangwan, M.C. Chiranth, H.S. Jamadagni, R. Sah, R. Venkatesha Prasad, and V. Gaurav. VAD techniques for real-time speech transmission on the internet. In *Proceedings of the International Conference on High Speed Networks and Multimedia Communications*, pages 46–50, 2002.

[Ser03] E. Serdarevic. Speech enhancement algorithms: Simulation and quality evaluation. Master's thesis, Graz University of Technology, Austria, 2003.

[SF96] P. Scalart and J. V. Filho. Speech enhancement based on a priori signal to noise estimation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 629–632, 1996.

[SFB00] V. Stahl, A. Fischer, and R. Bippus. Quantile based noise estimation for spectral subtraction and Wiener filtering. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1875–1878, 2000.

[SGFK05] M. Siafarikas, T. Ganchev, N. Fakotakis, and G. Kokkinakis. Overlapping wavelet packet features for speaker verification. In *Proceedings of Interspeech*, pages 3121–3124, 2005.

[She96] Y. Sheng. *The transforms and applications handbook*, chapter Wavelet transform, pages 747–827. The Electrical Engineering Handbook Series. CRC Press, 1996.

[SMM02] Zilany M. S., Hasan Md., and Khan M. Efficient hard and soft thresholding for wavelet speech enhancement. In *Proceedings of the European Signal Processing Conference*, 2002.

[sno] Services for NOmadic Workers. European Commission. `http://www.snow-project.org/`.

[SNO06] SNOW consortium. Publishable final report. Technical report, IST-FP6-511587 Project SNOW "Services for NOmadic Workers", December 2006.

[SPH98] R. Sarikaya, B. Pellom, and J. Hansen. Wavelet packet transform features with application to speaker identification. In *Proceedings of the Nordic Signal Processing Symposium*, pages 81–84, 1998.

[Sri05] S. Srinivasan. *Knowledge-based speech enhancement.* PhD thesis, KTH Electrical Engineering, 2005.

[SS97a] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/musicdiscriminator. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1331–1334, 1997.

[SS97b] J. Stegmann and G. Schroeder. Robust voice activity detection based on the wavelet transform. In *Proceedings of the Workshop on Speech Coding Telecommunications*, pages 99–100, 1997.

[SS98] J. Sohn and W. Sung. A voice activity detector employing soft decision based noise spectrum adaption. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 365–368, 1998.

[SSDB98] H. Sameti, H. Sheikhzadeh, Li Deng, and R.L. Brennan. HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, 6(5):445–455, 1998.

[SSKon] S. Srinivasan, J. Samuelsson, and W. B. Kleijn. Codebook driven short-term predictor parameter estimation for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, accepted for publication.

[StNSF95] Signal Processing Society and the National Science Foundation. Signal processing information base - noise data, 1995.

[Tas00] C. Taswell. The what, how, and why of wavelet shrinkage denoising. *Computing in Science and Engineering*, 2(3):12–19, 2000.

[Tea90] H. Teager. Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(5):599–601, 1990.

[TLS$^+$94] B. Tan, R. Lang, H. Schroder, A. Spray, and P. Dermody. Applying wavelet analysis to speech segmentation and classification. In *Proceedings of the SPIE Conference on Wavelet Applications in Signal and Image Processing*, pages 750–761, 1994.

[TMK97] D. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis. Speech enhancement based on audible noise suppression. *IEEE Transactions on Speech and Audio Processing*, 5(6):497–514, 1997.

[TO00] S. G. Tanyer and H. Ozer. Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, 8:478–482, 2000.

[Tuc92] R. Tucker. Voice activity detection using a periodicity measure. In *Inst. Elect. Eng.*, volume 139, pages 377–380, 1992.

[Val04] C. Valens. A really friendly guide to wavelets. Online tutorial, 2004.

[Vas00] S. V. Vaseghi. Advanced digital signal processing and noise reduction. *John Wiley & Sons*, UK, 2000.

[Vet01] R. Vetter. Single channel speech enhancement using MDL-based subspace approach in bark domain. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 641–644, 2001.

[VG03] D. Veselinovic and D. Graupe. A wavelet transform approach to blind adaptive filtering of speech from unknown noises. *IEEE Transaction on Circuits and Systems II: Analog and Digital Signal Processing*, 50:150–154, 2003.

[VH92] M. Vetterli and C. Herley. Wavelets and filter banks: theory and design. *IEEE Transactions on Signal Processing*, 40:2207–2232, 1992.

[Vid98] B. Vidakovic. Non-linear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association*, 93:173–179, 1998.

[Vir99] N. Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing*, 7(2):126–137, 1999.

[VK95] M. Vetterli and J. Kovacevic. *Wavelets and subband coding.* Prentice Hall, 1995.

[WG01] P. Wolfe and S. Godsill. Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. In *Proceedings of the Workshop on Statistical Signal Processing*, pages 496–499, 2001.

[WW06] B. F. Wu and K. C. Wang. Voice activity detection based on autocorrelation function using wavelet transform and Teager energy operator. *Journal of Computational Linguistics and Chinese Language Processing*, 11:87–100, 2006.

[XH02] Z. Xiong and T. Huang. Boosting speech/non-speech classification using averaged mel-frequency cepstrum. In *Proceedings of the Pacific-Rim Conference on Multimedia*, 2002.

[YEG$^+$05] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book (for HTK Version 3.3).* Microsoft Corporation-Cambridge University, Engineering Department, Cambridge University, 2005. http://htk.eng.cam.ac.uk/.

[YMPSR02] B. Yegnanarayana, S.R. Mahadeva Prasanna, and K. Sreenivasa Rao. Speech enhancement using excitation source information. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 541–544, 2002.

[You96] S. Young. A review of large-vocabulary continuous-speech. *IEEE Signal Processing Magazine*, 13:45–57, 1996.

[YV04] B. J. Yoon and P. P. Vaidyanathan. Wavelet-based denoising by customized thresholding. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 925-928, Montreal, Canada, 2004.

[YVH99] H. Yang, S. V. Vuuren, and H. Hermansky. Relevancy of time-frequency features for phonetic classification measured by mutual information. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 225–228, 1999.

[YWS92] X. Yang, K. Wang, and S.A. Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, 38(2):824–839, 92.

[ZD97] X. Zhang and M. Desai. Nonlinear adaptive noise suppression based on wavelet transform. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[ZL99] X. Zhang and Z. Luo. A new time-scale adaptive denoising method based on wavelet shrinkage. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1629–1632, 1999.

[ZSW97] S. A. Zahorian, P. Silsbee, and X. Wang. Phone classification with segmental features and a binary-pair partitioned neural network classifier. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1011–1014, 1997.

[ZT80] E. Zwicker and E. Terhardt. Analytical expression for critical band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society America*, 68:1523–1525, 1980.

[ZWC97] L. Zhang, T. Wang, and V. Cuperman. A CELP variable rate speech codec with low average rate. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 735–738, 1997.