Advanced Signal Enhancement Techniques with Application to Speech and Hearing

A DISSERTATION SUBMITTED TO THE CENTRE FOR EXCELLENCE IN SIGNAL AND IMAGE PROCESSING, DEPARTMENT OF ELECTRONIC AND ELECTRICAL ENGINEERING AND THE COMMITTEE FOR POSTGRADUATE STUDIES OF THE UNIVERSITY OF STRATHCLYDE IN PARTIAL FULFILLMENT OF THE REQUIRMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

BY

Navin Chatlani

2011

Declaration

I declare that this Thesis embodies my own research work and that it is composed by me. Where appropriate, I have made acknowledgments to the work of others.

Signed: Navin Chatlani

Digitally signed by Navin Chatlani
DN: cn=Navin Chatlani,
o=University of Strathclyde,
ou=CeSIP, email=navin.
chatlani@eee.strath.ac.uk, c=GB
Date: 2011.05.26 16:35:04 +01'00'

Date:

Acknowledgements

I owe my deepest gratitude to my PhD supervisor, Professor John J. Soraghan, for his support throughout the duration of my PhD. He imparted invaluable guidance, knowledge and sound advice, and gave me inspiration and encouragement when I needed it the most. He shared my enthusiasm on this research topic and helped to facilitate my PhD experience by steering my research in the right direction. I am also very grateful for his continual support relating to personal matters. I had a challenging time being away from home, especially given that my parents were ill at different points during my PhD. I am greatly indebted to Professor Soraghan for his support during these periods. One could not ask for a better *person* as a supervisor.

I would also like to thank Siemens Audiologische Technik (SAT) for giving me the opportunity to undertake my three month internship at their R&D facilities in Germany. This internship allowed me to develop my new noise reduction solutions on actual hearing aids. I thank Dr. Stephan Weiss for his referrals and for directing me to the right personnel at SAT, as this was critical for a timely start to this internship.

Most importantly, I thank my parents and my sisters, who always give me strength and encouragement. I would be lost without their support. I dedicate this thesis to them.

Abstract

Advanced signal enhancement techniques with application to speech and hearing are presented that are applied to areas including adaptive noise cancellation (ANC) in noisy speech signals, single channel noise reduction for speech enhancement, voice activity detection (VAD) and noise reduction in binaural hearing aids. The performance enhancement of the new techniques over competing approaches is presented.

For the domains of ANC and single channel noise reduction, the use of Empirical Mode Decomposition (EMD) is the underpinning technique employed. A novel approach to dual-channel speech enhancement using Adaptive Empirical Mode Decomposition (SEAEMD) is also presented, when a noise reference is available. The new SEAEMD system incorporates the multi-resolution approach EMD with ANC for effective speech enhancement in stationary and non-stationary noise environments.

Two novel Empirical Mode Decomposition based filtering (EMDF) algorithms are presented for single channel speech enhancement. The first system is designed to be particularly effective in low frequency noise environments. The second generalized EMDF system is designed to operate under other noisy conditions, with results presented for babble noise, military vehicle noise and car interior noise. It is shown that the proposed EMDF techniques enhance the speech more effectively than current speech enhancement approaches that use effective noise estimation routines.

Speech systems such as hearing aids require fast and computationally inexpensive signal processing technologies. A new and computationally efficient 1-dimensional local binary pattern (1-D LBP) signal processing procedure is designed and applied to (i) signal segmentation and (ii) the VAD problem. Both applications use the underlying features extracted from the 1-D LBP. The simplicity and low computational complexity of 1-D LBP processing are demonstrated.

A novel binaural noise reduction system is presented for steering the focus direction of a hearing aid (2 microphones per hearing aid) to additional directions as well as 0/180 degrees. The system places a spatial null in the direction of the target speaker to obtain a noise estimate. The noisy speech signal is then filtered to perform noise reduction, and thus focus on the target speaker located at the desired direction. The results demonstrate its performance at attenuating multiple directional interferers.

List of Symbols

$s[n]$	original noise-free speech signal
$d[n]$	noise source
$x[n]$	noisy speech
$S(k,i)$	STFT of clean speech
$D(k,i)$	STFT of noise source
$X(k,i)$	STFT of noisy speech
$w[n]$	analysis window
W	analysis window size
R	frame step size
Ψ	mother wavelet basis function
$I_j[n]$	j^{th} IMF
$r[n]$	residual signal of Empirical Mode Decomposition
$P_f\left(k,i\right)$	smoothed noisy speech spectrum
$\xi\left(k,i\right)$	a priori SNR
$\gamma(k,i)$	a posteriori SNR
$\tilde{\alpha}_d\left(k,i\right)$	time- varying, frequency dependent smoothing factor
α_d	smoothing parameter
$B_{\min}$	noise estimate bias
$\hat{\lambda}_d\left(k,i\right)$	Noise variance estimate
$p\left(k,i\right)$	conditional speech presence probability
$G_{LSA}\left(k,i\right)$	Log-Spectral Amplitude (LSA) gain
$x_D[n]$	denoised signal
segSNR	segmental SNR
WSS	Weighted Spectral Slope
$\mathrm{C_{SIG}}$	composite measure for signal quality
$\mathrm{C_{BAK}}$	composite measure for background distortion

$\mathrm{C}_{\mathrm{OVL}}$	composite measure for overall quality
$VAD(i)$	Output VAD decision
H_0	Hypothesis for speech absence
H_1	Hypothesis for speech presence
η	VAD decision threshold
$HR1$	Speech presence hit-rate
$HR0$	Speech absence hit-rate
$FAR1$	False alarm rate for speech presence
$FAR0$	False alarm rate for speech absence
$d_0[n]$	contaminating noise for adaptive noise cancellation
$d_1[n]$	Reference noise signal for adaptive noise cancellation
$V[j]$	IMF variance
D_{KL}	Kullback-Leibler Distance
D_{RAD}	Resistor Average Difference
$NUthresh$	threshold for unvoiced speech feature
T_{NV}	minimum time length threshold for unvoiced speech
T_V	minimum time length threshold for voiced speech
τ_{bc}	length of the detected speech burst
τ_{hb}	hangbefore length threshold
τ_{ho}	hangover length threshold
θ_{steer}	steering direction
θ_d	interferer direction
$c_R[n]$	anti-cardioid beamformer output
$c_F[n]$	cardioid beamformer output
θ_{null}	spatial notch direction
Φ_{focus}	power of the signal from the focus side
Φ_{int}	power of the signal from the interferer side
$f_{spatial}$	spatial sampling frequency
SIR_{gain}	Signal to interference gain

List of Acronyms

1-D	One Dimensional
2-D	Two Dimensional
ADMA	Adaptive Differential Microphone Array
ANC	Adaptive Noise Cancellation
AWGN	Additive White Gaussian Noise
BSS	Blind Source Separation
BTE	Behind The Ear
CASA	Computational Auditory Scene Analysis
CWT	Complex Wavelet Transform
DESPRIT	DUET-ESPRIT
DFT	Discrete Fourier Transform
DMA	Differential Microphone Array
DOA	Direction of Arrival
DUET	Degenerate Unmixing Estimation Technique
DWT	Discrete Wavelet Transform
EM	Electromagnetic
EMD	Empirical Mode Decomposition
EMDF	EMD-based Filtering
ENET	EMD-based Noise Estimation and Tracking
ESPRIT	Estimation of Signal Parameters via Rotational Invariance Techniques
ETSI	European Telecommunications Standards Institute
FIR	Finite Impulse Response
GAET	Geometrically Adaptive Energy Threshold
GSM	Global System for Mobile Communications
HMM	Hidden Markov Model
HOS	Higher Order Statistics
IBM	Ideal Binary Masks
ICA	Independent Component Analysis

ILD	Interaural Level Difference
IMCRA	Improved Minima Controlled Recursive Averaging
IMF	Intrinsic Mode Function
IP	Internet Protocol
ISTFT	Inverse Short Time Fourier Transform
ITD	Interaural Time Difference
ITU	International Telecommunication Union
KLD	Kullback-Leibler Distance
LBP	Local Binary Pattern
LLR	Log-likelihood Ratio
LPC	Linear Predictive Coding
LSA	Log-Spectral Amplitude
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MS	Minimum Statistics
MSE	Mean Square Error
MVDR	Minimum Variance Distortionless Response
MWF	Multi-channel Wiener Filtering
NLMS	Normalized Least Mean Square
OMLSA	Optimally-Modified Log-Spectral Amplitude
OSF	Over Subtraction Factor
PESQ	Perceptual Evaluation of Speech Quality
PSD	Power Spectral Density
RAD	Resistor Average Difference
ROC	Receiver Operating Characteristics
SEAEMD	Speech Enhancement using Adaptive Empirical Mode Decomposition
SIR	Signal to Interference Ratio
SNR	Signal to Noise Ratio
STFT	Short Time Fourier Transform
TDOA	Time Difference of Arrival
VAD	Voice Activity Detection

VoIP	Voice Over IP
WDO	W-Disjoint Orthogonal
WGN	White Gaussian Noise
WSS	Weighted Spectral Slope
WT	Wavelet Transform
ZCR	Zero-Crossing Rate

List of Figures

List of Tables

Table of Contents

1

Introduction

Speech enhancement and Voice Activity Detection (VAD) are fundamental to any noise reduction system which processes speech and audio signals to improve speech intelligibility and quality. Therefore, the technology spans commercial, industrial and even military sectors. Mobile phones, hearing aids, voice transmission devices and voice recognition systems are some examples where speech enhancement is applied. This thesis aims at solving some of the deficiencies which exist in systems for noise reduction for speech enhancement and VAD.

A common problem encountered in speech enhancement systems is the removal or reduction of unwanted disturbances, i.e. noise from desired speech signals. These noise sources may be present in the form of background, environmental noise sources or interfering speakers. Depending on the application scenario, one or multiple microphones may be available for acquiring the noisy speech signal. Noise reduction techniques for speech enhancement are important for improving speech quality. It is also known that a denoised signal is better attended to by a listener, as it delays the effect of listener fatigue which arises with noisy speech. Therefore, during the course of the PhD research, speech enhancement techniques have been focused on. These are enabled using underlying techniques in the domains of: noise estimation; filtering methods; adaptive noise cancellation (ANC); and directional signal processing. This is summarized in Figure 1.1 below.

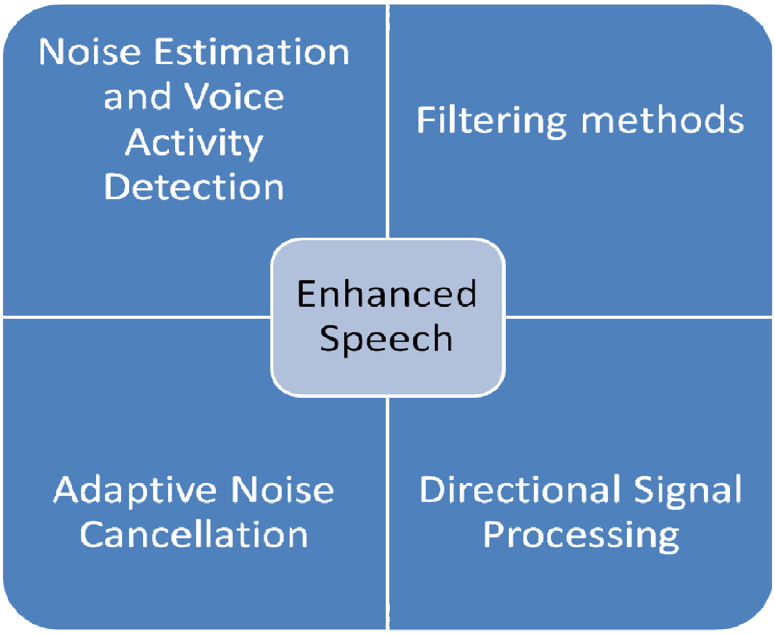

Noise Estimation and Voice Activity Detection
Filtering methods
Enhanced Speech
Adaptive Noise Cancellation
Directional Signal Processing

Figure 1.1: Summary of underlying techniques focused on to enable speech enhancement

1.1 Research Motivation

In this thesis, we investigate noise reduction methods for speech enhancement. A range of noise reduction techniques for speech enhancement have been proposed and many of them have been extensively reviewed in (Benesty et al., 2005, Loizou, 2007c, Ephraim and Cohen, 2006). Speech enhancement systems commonly employ multi-resolution techniques to analyze the noisy speech signal. When analyzing noisy signals, it is useful to decompose the signal using a spectral analysis technique which is able to provide a compact representation of the analyzed signal, as well as attain some degree of separation of the desired signal and the noise sources. In the case of non-stationary noisy speech, the short-time Fourier transform (STFT) is a popular choice for performing spectral analysis. Empirical Mode Decomposition (EMD) is a relatively new technique for multi-resolution, spectral analysis, which has been shown to be effective with stationary and non-stationary environments (Huang et al., 1998). Our aim is to investigate novel EMD-based speech enhancement techniques for single microphone (no noise reference) and dual microphone (noise reference available) systems.

Efficient voice activation detection is central to many application domains. In speech coding or speech processing systems, it is important to be able to distinguish between speech segments and speech pauses. By locating the pauses between speech components, one can access the characteristics of the environmental noise that exist, and one can subsequently reduce this noise with the additional information. Two examples of benefits associated with incorporating a VAD are higher speech quality and effective use of the limited bandwidth available with VoIP and mobile conversations. In wireless/mobile communication systems, the VAD process may be carried out using ITU G.729 standard. We investigate methods for improving on the G.729 VAD.

Binaural hearing aids are a relatively new class of hearing aids which are configured to have a wireless transmission link between the left and the right hearing aids. These hearing aids use directional signal processing to improve speech intelligibility. However, due to size constraints, two microphones are commonly utilized in each device. Traditionally, such hearing aids use the locally obtained microphone signals using first-order differential microphone arrays (DMA). Therefore, they have maximum sensitivity to target sources located directly in front of directly behind the user. We investigate methods to perform direction dependent noise reduction to focus on sources located at azimuths other than these two predefined directions, under the constraint of having two microphones in each hearing aid. An example of such a noisy scenario with multiple interferers is depicted in the figure below where two hearing aid users are conversing.

Can you hear me with all this noise?
Yes, my hearing aid reduces the noise level so I can hear you even though you are *not* in front of me.

Figure 1.2: Depiction of a noisy scenario with two speakers where one is a binaural hearing aid user employing the novel noise reduction system presented in this thesis

It must be noted that in our investigations, we consider noisy speech signals generated with additive noise in anechoic environments.

1.2 Summary of Original Contributions

The main research contributions can be divided into the two groups of single microphone and dual microphone approaches. These novel techniques are shown in Figure 1.3 and Figure 1.4 respectively, and outlined in this section.

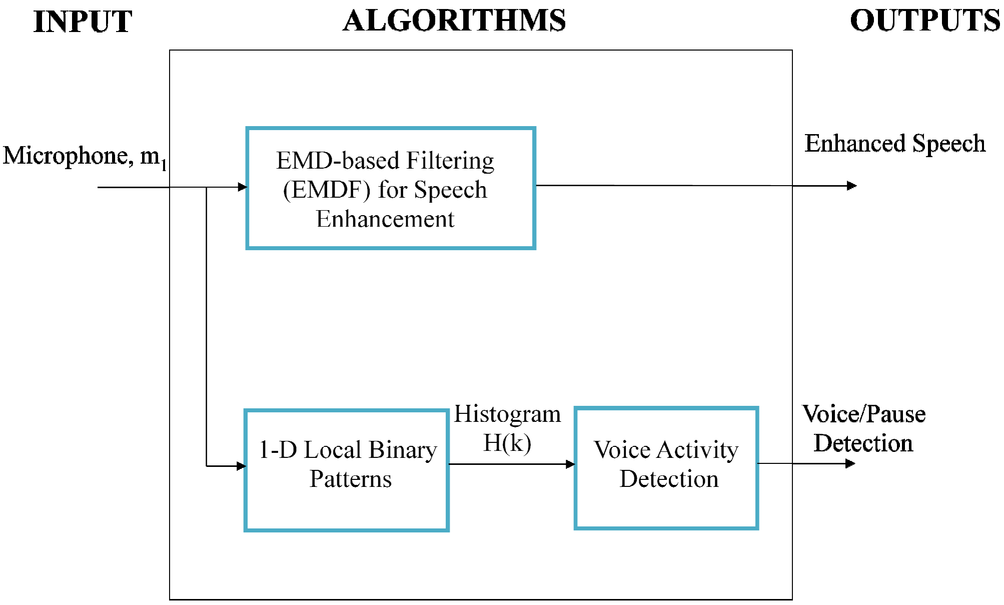


Figure 1.3: Proposed single microphone approaches

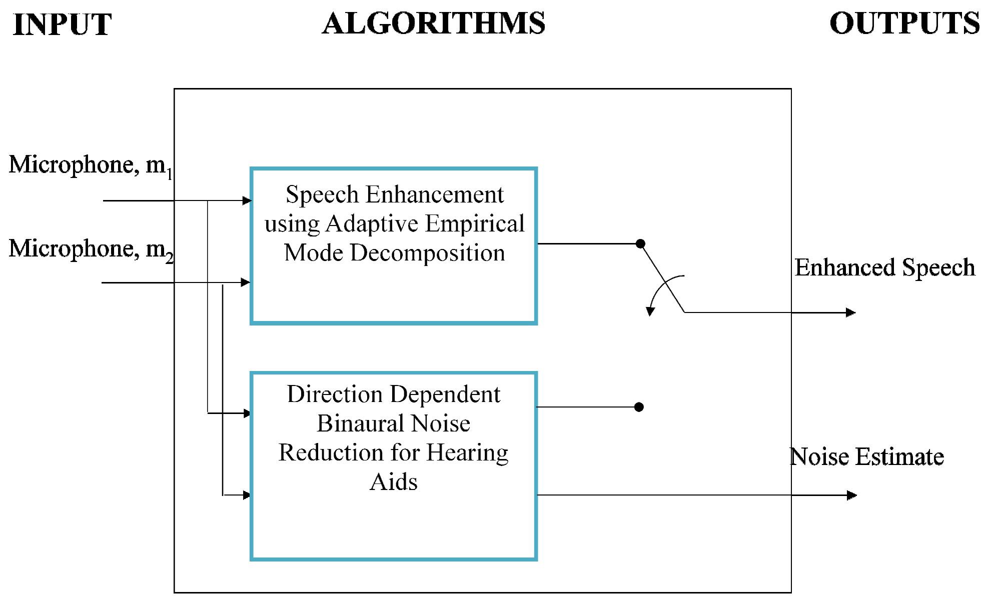

INPUT
ALGORITHMS
OUTPUTS
Microphone, m1
Microphone, m2
Speech Enhancement using Adaptive Empirical Mode Decomposition
Direction Dependent Binaural Noise Reduction for Hearing Aids
Enhanced Speech
Noise Estimate

Figure 1.4: Proposed dual microphone approaches

The single microphone contributions from Figure 1.3 are summarized below:

1) EMD-based Filtering (EMDF) for speech enhancement

Two novel EMD-based filtering (EMDF) methods for speech enhancement are designed. The first system is designed to be particularly effective in low frequency noise environments; whereas the second generalized EMDF technique is designed to operate in other noisy environments. It is shown that the proposed EMDF techniques enhance the speech more effectively than conventional optimally-modified log-MMSE approach which uses an IMCRA noise estimate. Comparative performance studies are conducted that demonstrate the superiority of the EMDF systems for speech enhancement in babble noise, car interior noise and military vehicle noise environments.

2) 1-D Local Binary Patterns with application to Voice Activity Detection

Local Binary Patterns (LBP) have been used in 2-D image processing for applications such as texture segmentation and feature detection. A new 1-dimensional local binary pattern (LBP) signal processing method is designed. Speech systems such as hearing aids require fast and computationally inexpensive signal processing. The practical use of LBP based speech processing is demonstrated on two signal processing problems: - (i) signal segmentation and (ii) voice activity detection (VAD). Both applications use the underlying features extracted from the 1-D LBP. The proposed VAD algorithm demonstrates the simplicity of 1-D LBP processing with low computational complexity. It is also shown that distinct 1-D LBP features are obtained and combined with a local noise power estimate, to identify the unvoiced components, voiced components and pauses from the measured noisy speech signals.

The dual microphone contributions from Figure 1.4 are summarized below:

3) **Speech Enhancement using Adaptive Empirical Mode Decomposition (SEAEMD)**

A novel approach to dual-channel speech enhancement using Adaptive EMD (SEAEMD) is presented. Spectral analysis of non-stationary signals can be performed by employing techniques such as the STFT and the Wavelet transform (WT), which use predefined basis functions. EMD performs very well in such environments as shown in (Huang et al., 1998). EMD decomposes a signal into a finite number of data-adaptive basis functions, called Intrinsic Mode Functions (IMFs). The new SEAEMD system incorporates this multi-resolution approach with adaptive noise cancellation (ANC) for effective speech enhancement on an IMF level, in stationary and non-stationary noise environments. A comparative performance study is conducted that compares the competitive method of conventional ANC to the robust SEAEMD system. The results demonstrated that the new system achieves improved speech quality with a lower level of residual noise.

4) **Direction Dependent Binaural Noise reduction for Hearing Aids**

A novel binaural noise reduction system is presented for steering the focus direction of the hearing aid to additional directions as well as 0/180 degrees. The system places a spatial null in the direction of the target speaker to obtain a noise estimate. The results demonstrate its performance at attenuating multiple directional interferers. This system satisfies the hearing aid’s size constraints by using two microphones in each device and the power consumption constraints since it requires only one microphone signal to be transmitted from each hearing aid to the other using a wireless data link. The new system is tested in a real-time environment to confirm the results using hearing aids.

1.3 Organization of the Thesis

Chapter 1 provided an introduction to this thesis. It covered the research motivations and introduced the novel techniques which will be presented in the following chapters. A summary of the original contributions was also provided. This thesis is decomposed into two parts. In the first part of the thesis, novel speech enhancement techniques for dual-microphone and single-microphone scenarios are presented, along with novel VAD techniques. Therefore, as seen in the description of the thesis organization below, relevant speech enhancement and VAD backgrounds are first presented in Chapter 2, followed by the novel speech enhancement and VAD systems in Chapters 3 and 4. In the second part of the thesis, we focus on noise reduction techniques for hearing aids, in environments with spatially localized interfering speakers. Therefore, the relevant background for this research work is given in Chapter 5 followed by the novel system for direction dependent binaural noise reduction in hearing aids in Chapter 6.

Chapter 2 reviews speech enhancement techniques. Several spectral analysis techniques for analyzing noisy speech signals are first reviewed, including the recently proposed EMD. When no noise reference is available, noise estimation techniques estimate the noise spectrum. This estimate is used in noise reduction methods for speech enhancement. Some relevant techniques for noise estimation and speech enhancement are therefore highlighted, including a section on EMD based denoising algorithms. This includes the Minimum Statistics (MS) based noise estimates, which use speech presence probability in their estimation routine. This inspired our work on our novel single-microphone speech enhancement technique. Finally, we present a review of VAD algorithms, including the ITU-T G.729 VAD which was used as the benchmark for performance comparison.

In Chapter 3, novel EMD-based techniques for speech enhancement are presented. The SEAEMD technique is proposed for noise cancellation in scenarios when a noise reference is available. A performance evaluation is given for speech which is contaminated with different noise types. Two new EMDF systems for speech enhancement in single-microphone environments are then described, and results are presented for different noisy scenarios.

Chapter 4 introduces the novel and computationally efficient 1-D LBP technique which is applied for signal segmentation and VAD. Novel VAD systems are presented, which combine the features obtained from the 1-D LBP operator, with a local noise power estimate, to distinguish voiced speech, unvoiced speech and speech pauses. A new VAD technique is presented which is able to update the noise statistics even during speech activity. The 1-D LBP VAD system's performance is evaluated using various noise sources using Receiver Operating Characteristics (ROC).

One of the research objectives in this thesis was to design a novel system for hearing aid users, which performs noise reduction and speech enhancement of signals received in environments with spatially localized interfering speakers. Chapter 5 describes the relevant background for noise reduction techniques, which are applicable to the problem where the desired speech signal is contaminated with spatially separated interfering speakers. First, target speaker separation and speech demixing techniques are detailed, followed by a brief background on a hearing aid's signal processing. A review of recent techniques is then given for achieving noise reduction in multi-talker scenarios, which are applicable to binaural hearing aids.

In Chapter 6, new techniques are presented for binaural systems in the hearing aid which can focus on additional directions as well as 0° and 180°. These systems use differential microphone arrays (DMAs) and filtering techniques. The performance of these systems is demonstrated using directivity plots, and Signal to Interference gains, in noisy scenarios with multiple interfering speakers. These techniques are computationally efficient and have been tested on actual hearing aids to confirm their performance.

Chapter 7 provides conclusions for this thesis based on the novel contributions and also presents extensions for future work. All references can be found at the end of this thesis.

2

Speech Enhancement Techniques

2.1 Introduction

A common problem encountered in speech enhancement systems is the removal or reduction of unwanted disturbances, i.e. noise from desired speech signals. These noise sources may be present in the form of background, environmental noise sources or interfering speakers. Speech enhancement and Voice Activity Detection (VAD) are fundamental to any system which processes speech and audio signals to improve speech intelligibility and quality. Adaptive noise cancellation (Widrow and Stearns, 1985) is commonly performed when enhancing speech sequences using an available noise reference. However, in many practical cases, no noise reference is available. In such scenarios, single-channel speech enhancement must be employed and a generalized system is shown in Figure 2.1. The noisy speech $x[n]$ is given by:

$$x[n]=s[n]+d[n] \tag{2.1}$$

where $s[n]$ is the original noise-free speech signal and $d[n]$ is the noise source which is assumed to be independent of the speech.

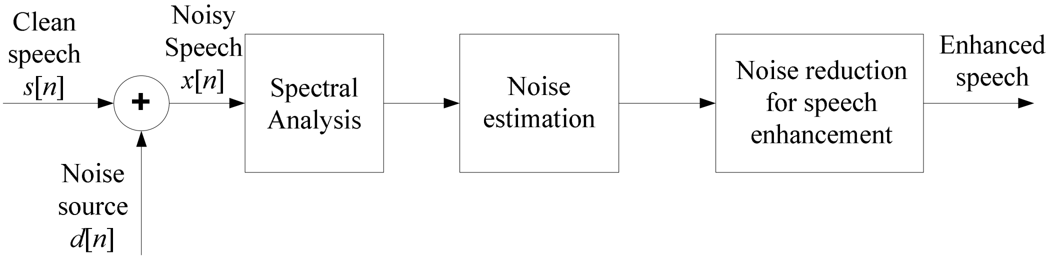


Figure 2.1: Generalized single-channel speech enhancement system

In the case of non-stationary signals such as speech, multi-resolution spectral analysis is first performed as shown in Figure 2.1. When analyzing noisy signals, it is useful to decompose the signal using a spectral analysis technique which is able to provide a compact representation of the analyzed signal, as well as attain some degree of separation of the desired signal and the noise sources. For example, the fractional Fourier transform has been shown to be effective at filtering noisy chirp signals due to its projection of the time-domain signal onto the time/frequency-domain at an angle between 0 (time-domain) and $\frac{\pi}{2}$ (Fourier transform). In Section 2.2, several techniques will be reviewed for performing spectral analysis of non-stationary signals, such as speech.

As depicted in Figure 2.1, a noise estimate is typically obtained from the transformed signal. Traditionally, VAD is employed to determine the noise statistics during silent segments. However, newer noise estimation techniques estimate the noise even during speech activity and this will be discussed in Section 2.3. The final block in Figure 2.1 shows the use of this noise estimate in order to perform noise reduction for speech enhancement. Some relevant techniques will be presented in Section 2.4 and EMD-based denoising in Section 2.5. Different subjective and objective assessment measures can be used when performing noise reduction for speech enhancement. Section 2.6 highlights several objective measures for assessing the performance of speech enhancement systems.

Efficient VAD is central to many application domains. In speech coding or speech processing systems, it is important to be able to distinguish between speech activity and non-speech (i.e. speech pauses). Examples of speech processing applications where VAD is employed are discussed in Section 2.7. A review of VAD techniques for noisy speech signals is also provided in this section. Finally, conclusions are given in Section 2.8.

2.2 Spectral Analysis Techniques

The non-stationary nature of the input signal that must be enhanced implies that relevant spectral analysis algorithms must be investigated (giving frequency, time, and amplitude). This section will give an overview of the Short Time Fourier Transform (STFT), Wavelet Transform, Empirical Mode Decomposition (EMD) and Cepstral Analysis.

2.2.1 Short Time Fourier Transform (STFT)

The well-known STFT breaks up a signal into short pieces by taking a window of sample points and performing the Fourier Transform of each piece. Therefore, the Fourier Transform of each successive section of the signal can be plotted next to each other to show the evolution of the frequency content of the signal over time in the resulting spectrum. The STFT of the noisy speech $x[n]$ from Equation (2.1) may be expressed as:

$$X(k,i)=\sum_{n=0}^{W-1}x[iR+n]w[n]e^{-j\frac{2\pi}{W}nk} \tag{2.2}$$

where k is the frequency bin, i denotes the time frame, $w[n]$ is the analysis window of size W. The frame step size R is given such that $R<W$ for overlapping time frames and $R=W$ for contiguous segments. The signal can be reconstructed using the Inverse STFT (ISTFT) by using the Overlap-Add method described in (Smith, 2008). For the ISTFT, a synthesis window which is biorthogonal to the analysis window is used (Wexler and Raz, 1990).

The STFT method suffers greatly from the time-frequency uncertainty effect, where it is possible to obtain either good frequency resolution (Narrowband Analysis) at the expense of time resolution or good time resolution (Wideband Analysis) at the expense of losing frequency resolution.

2.2.2 Real-Valued Wavelet Transform

The Wavelet Transform was developed from the family of sub-band coding and was introduced to overcome the deficiencies found with the STFT. Hence it also provides the desired time-frequency representation required.

The Discrete Wavelet Transform (DWT) is a sampled version of the Continuous Wavelet Transform. The Continuous Wavelet Transform is given by the general definition:

$$X_{WT}(\tau, c) = (|c|)^{-\frac{1}{2}} \int_{-\infty}^{+\infty} x(t)\Psi(\frac{t-\tau}{c})dt \tag{2.3}$$

where $x(t)$ is the analyzed signal, Ψ is the basis function, c is a variable scaling constant and τ is a constant of translation. The mother wavelet's width varies with each spectral component. Thus, at low frequencies the Wavelet Transform gives better frequency resolution but poorer time resolution, whereas at high frequencies, better time resolution is obtained but with inferior frequency resolution.

The DWT is efficiently computed using discrete-time filter banks by recursively applying a discrete-time low-pass filter and high-pass filter, in order to convert the source *X* into a finite number of filtered versions of *X*. This can be better illustrated by the multiple-level decomposition, termed the *N*-level wavelet decomposition tree, for simple 1-D signals, as shown in Figure 2.2. The number of decomposition levels *N* chosen is based on the analyzed signal and the application.

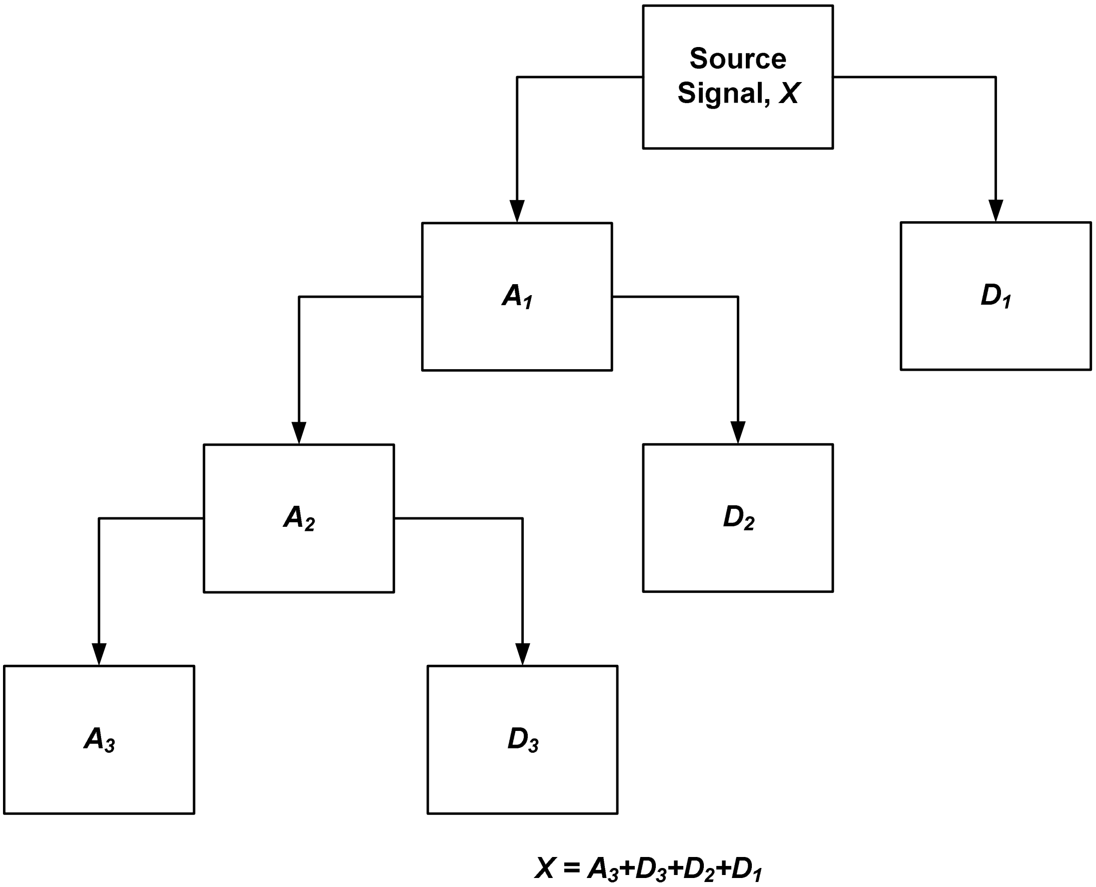
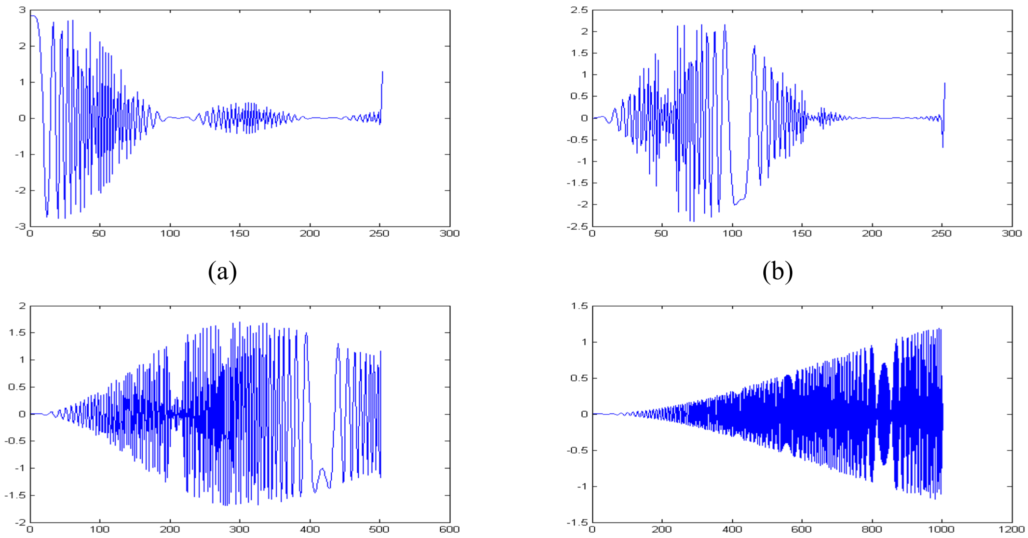


Figure 2.2: 1-D 3-Level Wavelet Decomposition Tree

From Figure 2.2, note that at each branch, low-pass filtering gives the approximation A_n, whereas high-pass filtering gives the detail D_n. Decimation occurs after each filtering stage. Two dimensional filtering, which occurs in images is an extension of this, and uses separable low-pass and high-pass filters. The DWT (using "Daubechies 2" family of wavelets) was performed on a non-stationary, chirp signal (which increases in frequency from 0 to 300 Hz at 2 seconds, with sampling frequency f_s = 1 kHz – hence with length 2000) using three levels of decomposition as was illustrated above. Figure 2.3 (a)-(d) show the different levels of approximation and detail that were obtained from the 3-level DWT decomposition. These decomposed components could be used to reconstruct the original signal through the inverse DWT process. This resynthesis is basically the reverse of the decomposition shown in Figure 2.3 and involves interpolating each sub-band by a factor of two and then using synthesis filters before summation (Sripathi, 2003).



(a) (b)

(c) (d)

Figure 2.3: DWT of chirp signal (a) DWT Level 3 Approximation Coefficients (b) DWT Level 3 Detail Coefficients (c) DWT Level 2 Detail Coefficients (d) DWT Level 1 Detail Coefficients

2.2.3 Complex-Valued Wavelet Transform

The real-valued DWT has some deficiencies such as shift sensitivity, poor directionality in higher dimensions and the inherent lack of phase information as detailed in (Selesnick et al., 2005). The complex Wavelet Transform (CWT) has been proposed in an attempt to solve these issues. The CWT has a complex-valued scaling function and a complex-valued wavelet. The complex-valued wavelet is given by:

$$\Psi_c[t] = \Psi_r[t] + j\Psi_i[t] \qquad \textbf{(2.4)}$$

where $\Psi_r[t]$ and $\Psi_i[t]$ are a Hilbert transform pair. The dual-tree approach proposed in (Kingsbury, 2001) and (Selesnick et al., 2005) is a possible implementation of the CWT which has two filter bank branches instead of the one shown in Figure 2.2 for the real-valued DWT, to produce the real and the imaginary parts of the complex coefficients. MATLAB simulations using the Q-shift, dual-tree filter CWT implementation proposed in (Selesnick et al., 2005) gave the results shown in Figure 2.4 to demonstrate the shift invariance using the CWT. Sixteen unit pulses were turned on at different times and analyzed using both the real-valued

DWT and the CWT. From Figure 2.4, it is seen that the CWT is not shift sensitive unlike the real-valued DWT.

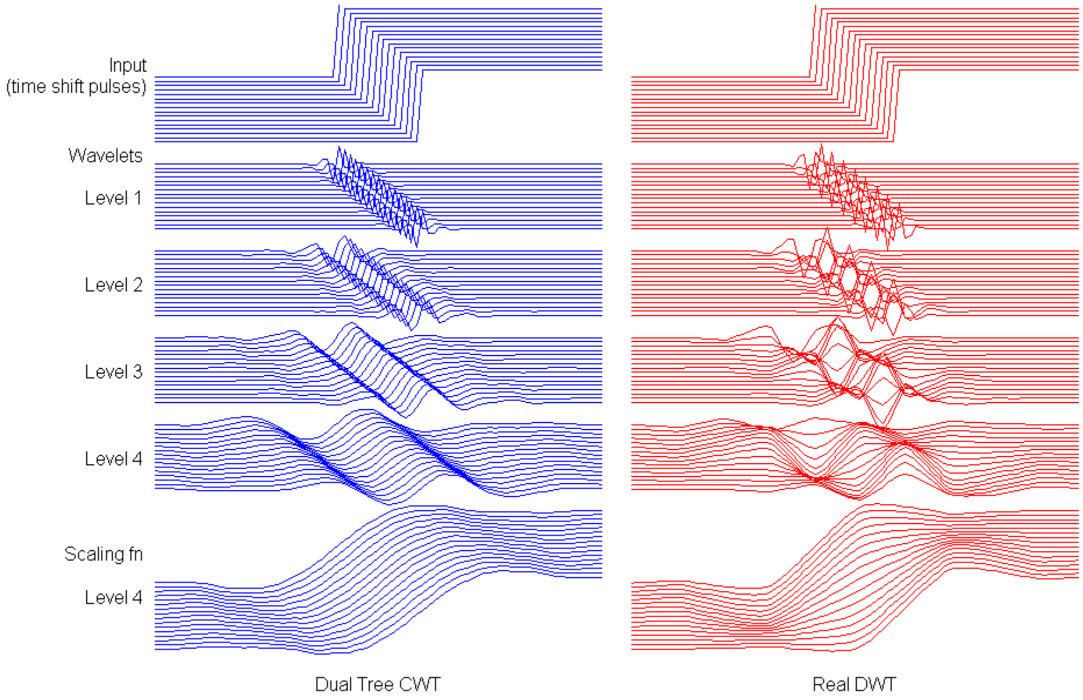


Figure 2.4: Comparison of shift-sensitivity using the CWT and real DWT

2.2.4 Empirical Mode Decomposition (EMD)

EMD (Huang et al., 1998, Rilling et al., 2003) is a non-linear technique for analyzing and representing non-stationary signals. EMD is data-driven and decomposes a time domain signal into a complete and finite set of adaptive basis functions which are defined as Intrinsic Mode Functions (IMFs). EMD does not use predefined basis functions. The IMFs formed by the EMD are oscillatory functions that have no DC component. Figure 2.5 illustrates the main stages in the EMD algorithm. EMD examines the signal between two consecutive extrema (e.g. minima) and picks out the high frequency component that exists between these two points. The remaining local, low frequency component can then be found. The motivation behind the EMD is to perform this procedure on the entire signal and then to iterate on the residual low frequency parts. This allows identification of the different oscillatory modes that exist in the signal. The IMFs found must be symmetric with respect to local zero means and have the same number of zero crossings and extrema, or differ at most by one. The IMF is considered as zero-mean based on some stopping criteria such as the standard deviation between consecutively sifted functions as detailed in (Rilling et al., 2003).

Frequency information is embedded in the IMFs. These data-adaptive basis functions give physical meaning to the underlying process. The reconstruction process is given in Equation (2.5), which involves combining the N IMFs formed from the EMD and the residual $r[n]$:

$$x[n] = \sum_{j=1}^{N} I_j[n] + r[n] \tag{2.5}$$

where $I_j[n]$ is the j^{th} IMF.

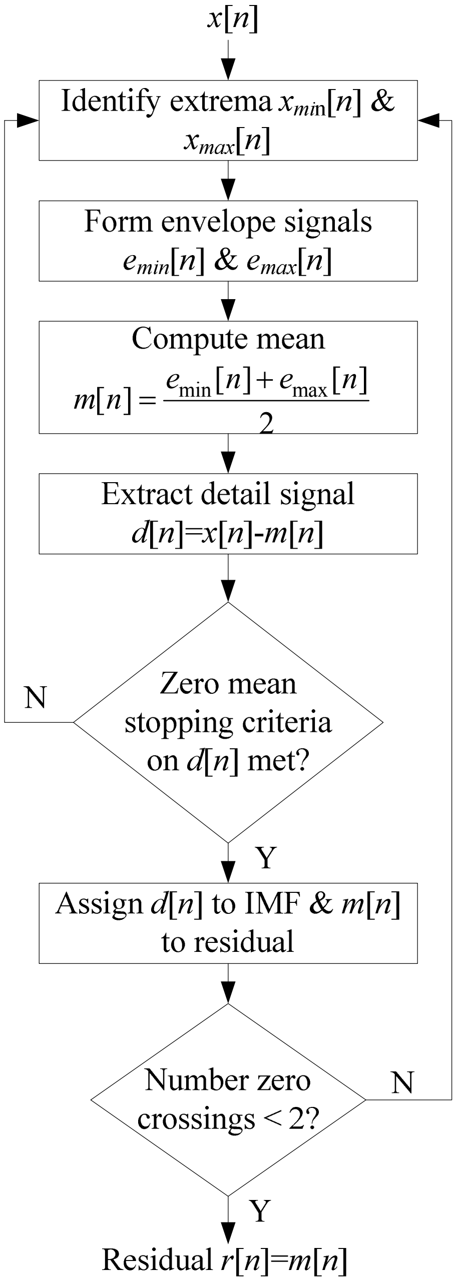

$x[n]$
Identify extrema $x_{min}[n]$ & $x_{max}[n]$
Form envelope signals $e_{min}[n]$ & $e_{max}[n]$
Compute mean $m[n] = \frac{e_{\min}[n] + e_{\max}[n]}{2}$
Extract detail signal $d[n]=x[n]-m[n]$
Zero mean stopping criteria on $d[n]$ met?
N
Y
Assign $d[n]$ to IMF & $m[n]$ to residual
Number zero crossings < 2?
N
Y
Residual $r[n]=m[n]$

Figure 2.5: EMD algorithm

The EMD of the same chirp signal (which increases in frequency from 0 to 300 Hz at 2 seconds, with f_s = 1 kHz) used in Section 2.2.2 on wavelets, results in the generation of the IMFs where the first IMF contains most of the signal energy. The use of the Hilbert transform (known as the Hilbert Huang Transform (Huang et al., 1998)) on the first IMF only, then extracts instantaneous frequencies as a function of time as shown in the time-frequency spectrum in Figure 2.6.

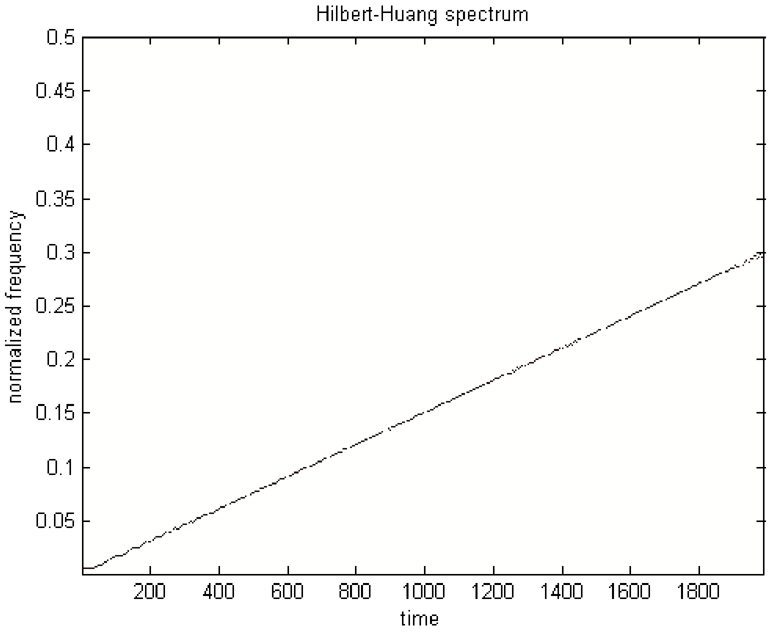


Figure 2.6: Hilbert-Huang spectrum of the chirp signal

From Figure 2.6, the usefulness of this spectral analysis technique is obvious since it presents an effective way for extracting the instantaneous frequencies in the signal being analyzed. The above spectrum clearly shows how the frequency of the chirp signal changes over time from 0 Hz to 300 Hz after two seconds.

2.2.5 Cepstral Analsysis

Various signal processing applications including speech processing utilize the non-linear technique cepstrum analysis (Oppenheim and Schafer, 1989). The complex cepstrum $\hat{x}[n]$ of a signal $x[n]$ is defined as:

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\ln |X(\Omega)| + j\angle X(\Omega) \right) e^{j\Omega n} d\Omega \qquad \textbf{(2.6)}$$

where $X(\Omega)$ is the Discrete Fourier Transform (DFT) of $x[n]$. The complex cepstrum exists if $\ln|X(\Omega)|$ has a convergent power series. The DFT $\hat{X}(\Omega)$ of the complex cepstrum must be a continuous periodic function of Ω.

The transformation of a signal into its complex cepstrum is homomorphic where a homomorphic system is one which satisfies a generalization of the principle of superposition. Therefore, the complex cepstrum of the convolution of two signals $x_1[n]*x_2[n]$ is equivalent to the summation of the complex cepstra $\hat{x}_1[n]+\hat{x}_2[n]$. This can be generalized in Figure 2.7 below, where L is any linear system, $D_*\left[.\right]$ is the complex cepstrum operator and $D_*^{-1}\left[.\right]$ is its inverse.

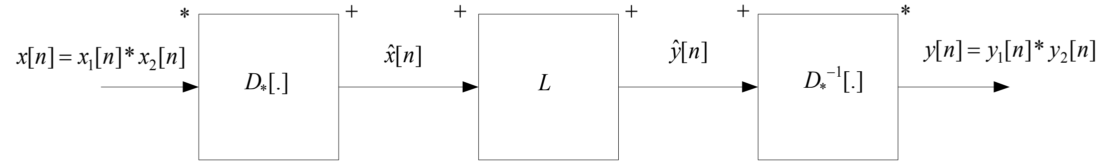


Figure 2.7: Generalization of the complex cepstrum in a homomorphic system

One common application of this homomorphic system decomposition is in deconvolution. For example, the recovery of $x_1[n]$ could be performed by using a frequency-invariant system for L in order to perform the deconvolution. This frequency-invariant linear system is termed a “lifter” which operates on the complex cepstrum and further details are given in (Oppenheim and Schafer, 1989).

2.2.6 Discussion

The above algorithms are effective for spectral analysis of non-stationary signals. The EMD is relatively new, and is still being tested. In (Huang et al., 1998), it is shown that the EMD performs well when analyzing non-stationary signals, due to the data-dependent extraction of its IMFs in the algorithm. Since the IMFs are extracted in the time domain, the basis functions are the same length as the original signal and this allows for preservation of varying frequency/time. As shown in Section 2.2.4, it is very effective in extracting a signal’s instantaneous frequencies.

2.3 Noise Estimation

Single-channel speech enhancement systems traditionally employ Voice Activity Detection (VAD) to estimate the statistics of the noise signal during silent segments. If the VAD approach is conservative, then it will attempt to reduce false alarms for silence detection, which results in less frequent noise power updates. In highly non-stationary environments, the noise power must be tracked even during speech activity. Noise estimation techniques which operate in the STFT domain are very popular, including newer noise estimation systems such as the Minimum Statistics (MS) (Martin, 2001) and the Improved Minima Controlled Recursive Averaging (IMCRA) (Cohen, 2003). These techniques estimate the noise spectrum based on the observation that the noisy signal power decays to values characteristic of the contaminating noise during speech pauses. The main challenge faced by these techniques is tracking the noise power during speech segments. This would result in poor estimates during long speech segments with few pauses. In this thesis, we focus on the IMCRA noise estimation technique and its background is given in the following sub-section.

2.3.1 Improved Minima Controlled Recursive Averaging

Improved Minima Controlled Recursive averaging (IMCRA) combines Minimum Statistics (MS) with recursive averaging to perform noise spectrum estimation. A summary of the IMCRA algorithm is provided in Figure 2.8. Consider the noisy speech model $x[n]$ described by Equation (2.1). The STFT of $x[n]$ may be written as:

$$X(k,i)=S(k,i)+D(k,i) \tag{2.7}$$

for frequency bin k and time frame i. It is assumed that the STFT coefficients of both the speech and the noise have asymptotically independent, complex Gaussian distributions (Ephraim and Malah, 1984). The first iteration involves smoothing the noisy speech spectrum $P(k,i)=|X(k,i)|^2$ in frequency and time to form the smoothed power spectrum $P_f(k,i)$. The minima values of $P_f(k,i)$ termed $P_{f,\min}(k,i)$ are tracked using the MS approach, over a specified finite window of length L. Rough voice activity detection is performed after smoothing. A minimum tracker is then used to produce an indicator function $\delta(k,i)$ for speech presence. This speech presence decision is based on conditions (Cohen, 2003) set on the following ratios for the a posteriori SNR $\gamma_{\min}(k,i)$ and the a priori SNR $\xi(k,i)$ as defined by:

$$\gamma_{\min}(k,i) \triangleq \frac{P(k,i)}{B_{\min} P_{f,\min}(k,i)} \qquad \xi(k,i) \triangleq \frac{P_f(k,i)}{B_{\min} P_{f,\min}(k,i)} \tag{2.8}$$

where $B_{\min}$ is the bias of the minimum noise estimate.

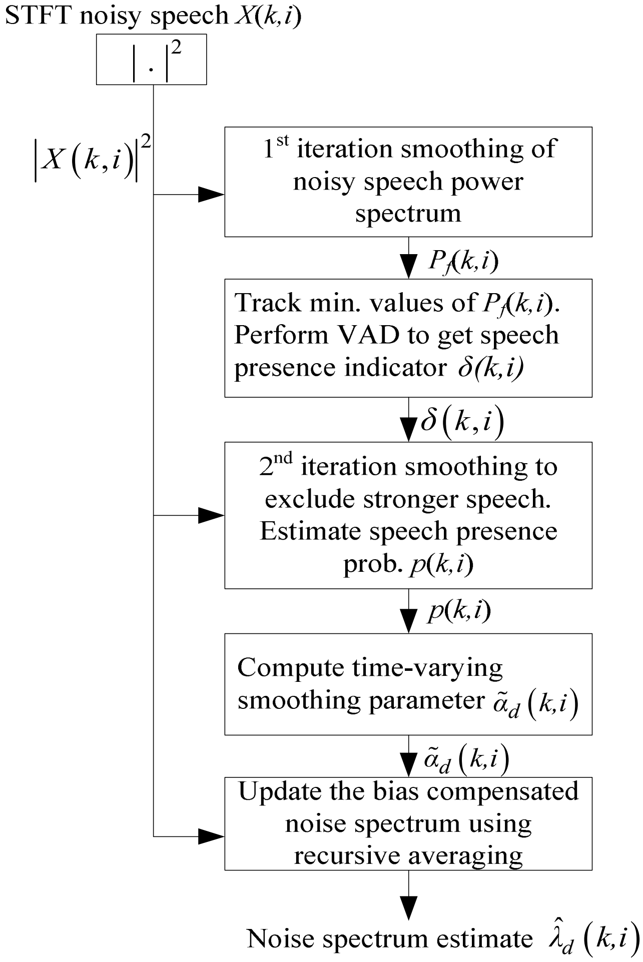

STFT noisy speech $X(k,i)$
$|\cdot|^2$
$|X(k,i)|^2$
1st iteration smoothing of noisy speech power spectrum
$P_f(k,i)$
Track min. values of $P_f(k,i)$. Perform VAD to get speech presence indicator $\delta(k,i)$
$\delta(k,i)$
2nd iteration smoothing to exclude stronger speech. Estimate speech presence prob. $p(k,i)$
$p(k,i)$
Compute time-varying smoothing parameter $\tilde{\alpha}_d(k,i)$
$\tilde{\alpha}_d(k,i)$
Update the bias compensated noise spectrum using recursive averaging
Noise spectrum estimate $\hat{\lambda}_d(k,i)$

Figure 2.8: Block diagram of IMCRA noise estimation

The speech presence indicator, $\delta(k,i)$, is used in a second smoothing stage to eliminate strong speech components from the short term spectrum $P(k,i)$ before time-domain recursive averaging. This exclusion enables improved minima tracking among the power components primarily associated with the contaminating noise source. Following (Cohen, 2003), the conditional speech presence probability, $p(k,i)$ is then estimated and used to compute the time-varying, frequency dependent smoothing factor $\tilde{\alpha}_d(k,i)$ given by:

$$\tilde{\alpha}_d(k,i) = \alpha_d + (1\text{-}\alpha_d)\,p(k,i) \qquad \textbf{(2.9)}$$

where the smoothing parameter α_d ranges from [0, 1]. Recursive averaging of the power spectral values, $P(k,i)$, is then performed to obtain an estimate of the noise spectrum $\hat{\lambda}_d(k,i)$. IMCRA was shown in (Cohen, 2003) to be more robust than the basic MS method in (Martin, 2001) since the minimum tracking was not used directly in the noise estimation.

2.3.2 Review of Relevant Noise Estimation Techniques

The method proposed by (Rangachari and Loizou, 2006) follows the procedure in IMCRA by determining the speech presence probability $p(k,i)$, in order to compute the smoothing factor $\tilde{\alpha}_d(k,i)$ as in Equation (2.9). However, a frequency dependent threshold was used in determining $p(k,i)$, as opposed to the fixed value used in IMCRA. This proposed method excludes the second smoothing stage performed before computing $p(k,i)$, as proposed in IMCRA. This approach utilizes the minima tracking steps from (Doblinger, 1995) to continuously track the minima of the noisy speech in each frequency bin. It was shown that faster noise spectrum updates are possible when tracking rapidly changing noise levels. The results in (Rangachari and Loizou, 2006) achieves a lower normalized Mean Square Error (MSE) between the estimated and the actual noise spectra. The spectrograms presented in this work revealed that when this proposed technique is combined with a speech enhancement algorithm, there is greater attenuation of the contaminating noise as well as of weaker speech components.

A subspace algorithm that tracks the noise power spectrum in speech segments was proposed in (Hendriks et al., 2008). This method involves the eigenvalue decomposition of pre-whitened STFT correlation matrices of the noisy speech. It assumes a low-rank model to separate the noise subspace from the signal (plus noise) subspace. A model-order estimation technique is proposed to determine the dimensions of the two subspaces. However, this remains a challenge as there are obvious implications to over- and under-estimating the noise subspace dimension. A deficiency of the low-rank model assumption is that it is applicable to voiced speech components, and unvoiced components require a relatively higher rank approximation. This proposed technique was also shown to be computationally expensive and not suitable for real-time applications. Approaches for subspace decomposition based noise estimation in the time domain have also been proposed, and an example is given in (Bhunjun et al., 2006).

A wavelet-based approach for noise estimation from the DWT coefficients of the noisy speech is proposed in (Lei and Tung, 2008). The noise power estimate is used in the subsequent wavelet-thresholding step for speech enhancement.

A detailed review of other related noise estimation techniques are given in (Loizou, 2007b).

2.4 Speech Enhancement

Noise reduction techniques are performed for speech enhancement to improve speech quality. The improvement of speech quality in noisy signals reduces listener fatigue and therefore increases intelligibility for long speech duration. Spectral subtraction for noise removal and speech enhancement is well accepted. If the noise level can be estimated, then subtraction of the noise power spectrum from the noisy signal's power spectrum can reduce the level of noise to estimate the enhanced speech spectrum, $\left|\hat{S}(k,i)\right|^2$ by (Berouti et al., 1979):

$$\left|\hat{S}(k,i)\right|^2 = \begin{cases} \left|X(k,i)\right|^2 - \alpha\left|\hat{D}(k,i)\right|^2 & \text{if } \left|X(k,i)\right|^2 > (\alpha+\beta)\left|\hat{D}(k,i)\right|^2 \\ \beta\left|\hat{D}(k,i)\right|^2 & \text{otherwise} \end{cases} \quad \textbf{(2.10)}$$

In Equation (2.10), α is the over-subtraction factor ($\alpha \geq 1$) and β is the spectral floor parameter used to reduce musical noise. Note that when $\alpha = 1$, Equation (2.10) corresponds to the standard spectral subtraction rule. After noise reduction, the estimate of the enhanced speech signal $\hat{S}(k,i)$ is obtained by:

$$\hat{S}(k,i) = \left|\hat{S}(k,i)\right| e^{j\theta_x(k,i)} \quad \textbf{(2.11)}$$

where $\theta_x(k,i)$ is the observable noisy speech phase.

The statistically optimum Wiener filter enhances speech in the Minimum Mean Square Error (MMSE) sense. The Wiener filter gain is easily derived as:

$$G(k,i) = \frac{\xi(k,i)}{\xi(k,i)+1} \quad \textbf{(2.12)}$$

where $\xi(k,i)$ is the a priori SNR and is defined as:

$$\xi(k,i) \triangleq \frac{R_{SS}(k,i)}{R_{DD}(k,i)} \quad \textbf{(2.13)}$$

and $R_{ss}(k,i)$ is the clean speech spectrum and $R_{DD}(k,i)$ is the noise spectrum. The Log-Spectral Amplitude (LSA) estimator was proposed in (Ephraim and Malah, 1985) as an alternative speech enhancement technique to the statistically optimum Wiener filter. The LSA estimator minimizes the MSE of the LSA of the speech signal as given by:

$$\mathrm{E}_{\min}\left[\left\{\lg S(k,i) - \lg \hat{S}(k,i)\right\}^2\right] \quad \textbf{(2.14)}$$

The corresponding LSA gain function denoted as $G_{LSA}(k,i)$ to be applied to the noisy signal $X(k,i)$ is (Ephraim and Malah, 1985):

$$G_{LSA} \triangleq \frac{\hat{\xi}(k,i)}{1+\hat{\xi}(k,i)} \exp\left(\frac{1}{2}\int_{v(k,i)}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{2.15}$$

where $v(k,i)$ is a function of the a priori and a posteriori SNR. The a priori SNR $\hat{\xi}(k,i)$ may be estimated using the decision directed approach in (Ephraim and Malah, 1984). This decision directed approach is able to minimize musical noise by producing a smoothed estimate of the true a priori SNR. Therefore, the attenuation of the noise from $X(k,i)$ does not change drastically in adjacent frames which is key to musical noise reduction. Investigations have been conducted for varying the smoothing factor used in the decision directed a priori SNR estimation rather than using a fixed smoothing factor. Other methods adaptively update the smoothing factor as well as determine it in the MMSE sense using the previous frame's a priori SNR estimate (Hasan et al., 2004, Soon and Koh, 2000). More recently, this problem of varying the smoothing factor and trading off the amount of speech distortion and musical noise using a different a priori SNR estimate was covered in (Gerkmann et al., 2008).

Subspace based speech enhancement approaches have been proposed to estimate the noise suppression gain and these are extensively reviewed in (Loizou, 2007d). However, challenges exist when coloured noises contaminate the speech signal. Also, a low-rank model for speech is usually assumed in order to estimate the signal and the noise subspace dimensions, as discussed in Section 2.3.2.

2.5 EMD-Based Denoising

Empirical Mode Decomposition (EMD) is an effective multi-resolution approach for analyzing non-stationary signals. By performing a sifting process, the EMD decomposes a desired signal into Intrinsic Mode Functions (IMFs) which are data-adaptive as opposed to other transforms such as the DWT (Section 2.2.2) which use predefined basis functions. As detailed in (Flandrin and Rilling, 2004), the IMFs formed from EMD are locally orthogonal. Furthermore, EMD does not correspond to pre-determined sub-band filtering. The frequency content of the IMFs varies from high frequency to low frequency as the IMF order increases. In (Flandrin and Rilling, 2004), the EMD of fractional Gaussian Noise (fGN) was shown to result in a filter bank like structure with overlapping pass-bands for each IMF mode. The first IMF has a high-pass characteristic but also contains some lower energy, low frequency content. The higher order modes were also shown to have this overlapping band-pass characteristic. Similar results as those for fGN were found for the EMD of white Gaussian Noise (WGN) as detailed in (Wu and Huang, 2004).

New EMD based methods for noise suppression and signal enhancement include single-channel speech enhancement methods. EMD-based denoising (Flandrin et al., 2004), EMD-MMSE (Khaldi et al., 2008) and EMD-based thresholding (Kopsinis and McLaughlin, 2009) of signals contaminated with stationary White Gaussian Noise (WGN) or fGN are based on an empirically observed noise model derived from a study of IMF statistics in noise-only situations. EMD-based denoising (Flandrin et al., 2004) involves decomposing a noisy signal using EMD and performing a partial reconstruction with those IMFs composed of the desired signal. In this work, a study was carried out on the IMF statistics of fGN signals which resulted in an empirically observed noise model for noise-only situations. This noise-only model allows an estimation of the energy of the IMF modes. The noisy signal $x[n]$ considered for denoising comprised the desired signal and fGN. For denoising, the energy of each IMF of the noisy signal is computed and compared to the noise-only model's IMF energy. The IMF order for which the computed IMF energy deviates from a predefined threshold is determined and denoted as $M+1$. The denoised signal $x_D[n]$ is then obtained from the partial reconstruction of the IMFs:

$$x_D[n] = \sum_{j=M+1}^{N} I_j[n] + r[n] \tag{2.16}$$

This reconstructed signal $x_D[n]$ corresponds to a slower-varying signal that was superimposed on the fGN signal which dominates the first M IMFs.

The two cases of a desired signal contaminated with fGN or WGN are special since the first few IMFs are predominantly composed of the noise signal and this led to successful speech denoising strategies such as in (Khaldi et al., 2008, Kopsinis and

McLaughlin, 2009, Zou et al., 2006). In (Khaldi et al., 2008), EMD-MMSE is performed by filtering the IMFs formed from the decomposition of speech contaminated with WGN. EMD-based thresholding methods were presented in (Kopsinis and McLaughlin, 2009) for signals contaminated with WGN. These proposed techniques followed successful wavelet thresholding methods. The EMD-MMSE and the EMD-based thresholding methods both estimate the noise statistics using the empirically observed noise model presented in (Flandrin and Rilling, 2004). In (Zou et al., 2006), enhancement is achieved for speech signals corrupted with WGN by an algorithm based on partial reconstruction of the higher order IMFs which are less affected by WGN. These techniques focus their enhancement efforts on the lower-order IMFs and therefore, for speech contaminated with additive WGN, it is expected that the high-frequency unvoiced components of the speech signal that exist in these IMFs will be filtered.

In (Hasan and Hasan, 2009), an optimum gain function is estimated for each IMF to suppress residual noise that may be retained after single channel speech enhancement algorithms.

2.6 Speech Quality Measures

Different subjective listening tests and objective measure can be used to evaluate the performance of speech enhancement algorithms. Subjective measures include Mean Opinion Score (MOS) and Diagnostic Acceptability Measure (DAM), however they require numerous listeners to rate the speech quality on a predetermined scale. Objective measures include the widely used Segmental SNR (SegSNR), Weighted Slope Spectral (WSS) distance, Log-Likelihood Ratio (LLR) and the more recent Perceptual Evaluation of Speech Quality (PESQ).

Segmental SNR is one of the most widely used objective measures and is defined as:

$$\text{segSNR} = \frac{10}{L}\sum_{i=0}^{L-1}\log_{10}\frac{\sum_{n=Wi}^{Wi+W-1}\left|s[n]\right|^2}{\sum_{n=Wi}^{Wi+W-1}\left|\hat{s}[n]-s[n]\right|^2} \tag{2.17}$$

where $\hat{s}[n]$ is the enhanced speech signal, L is the number of segments/frames, and W is the segment length. Segments which lie outside a specified range of segSNR (lower limit of -10 dB and upper limit of 35 dB typically) are excluded from this measurement.

The WSS (Klatt, 1982) is perceptually motivated, and it measures the weighted difference in spectral slope over B critical frequency bands. WSS is evaluated for each frame as:

$$\text{WSS} = \sum_{k=1}^{B} W(k)\left(C_s(k)-C_{\hat{s}}(k)\right) \tag{2.18}$$

where $W(k)$ is a weight calculated as in (Klatt, 1982) to emphasize slopes near spectral peaks or valleys. $C_s(k)$ and $C_{\hat{s}}(k)$ are the spectral slopes of the clean and enhanced speech respectively. WSS measures are usually averaged over all frames. Smaller WSS values indicate better performance.

The LLR is a Linear Predictive Coding (LPC) based measure and is given by:

$$d_{LLR}\left(\mathbf{a}_s,\mathbf{a}_{\hat{s}},i\right) = \log\frac{\mathbf{a}_{\hat{s}}{}^T\mathbf{R}_{\hat{s}\hat{s}}\mathbf{a}_{\hat{s}}}{\mathbf{a}_s{}^T\mathbf{R}_{ss}\mathbf{a}_s} \tag{2.19}$$

where $\mathbf{a}_s$ and $\mathbf{a}_{\hat{s}}$ are the LPC coefficients of the clean and the enhanced signals respectively and $\mathbf{R}_{ss}$ is the clean speech autocorrelation matrix. d_{LLR} is interpreted as ratio of the energies of the prediction residuals after filtering with the corresponding LPC coefficients and is always positive. Smaller values of this measure indicate better speech enhancement performance.

PESQ is an ITU-T recommendation (P.862) and was proposed to measure distortions to the speech signal over a telecommunications network. Hence, it is evaluated over the frequency range (350 Hz – 3250 Hz) and this quality index is scored over a

MOS-like scale. A typical PESQ scale ranges from 1 to 4.5. More details about its evaluation are given in (Rix et al., 2001).

The segSNR, WSS, LLR and PESQ have been combined into composite measures for measuring signal quality (C_{SIG}), background distortion (C_{BAK}) and overall quality (C_{OVL}) in (Hu and Loizou, 2008). These quality indices which lie in the range [1 5] are given by:

$$C_{SIG} = 3.093 - 1.029\,LLR + 0.603\,PESQ - 0.009\,WSS \quad (2.20)$$

$$C_{BAK} = 1.634 + 0.478\,PESQ - 0.007\,WSS + 0.063\,segSNR \quad (2.21)$$

$$C_{OVL} = 1.594 + 0.805\,PESQ - 0.512\,LLR - 0.007\,WSS \quad (2.22)$$

2.7 Voice Activity Detection

Voice Activity Detection (VAD) is a common process in speech processing systems. VAD involves the segmentation of audio speech signals into periods of speech and pauses. A review of some recent VAD fundamentals and techniques is given in (Ramírez et al., 2007). The speech/non-speech classification task in VAD is useful in many applications and some examples of speech processing domains where it is used are given below:

1) Speech enhancement and noise reduction techniques

   Traditionally VAD is required for speech enhancement and noise reduction techniques so that the noise statistics could be estimated during periods of speech inactivity.

2) Speech coding

   In communication systems, speech coders use VAD to transmit detected speech segments only and then generate comfort noise on the receiver end during silence. This results in efficient bandwidth use.

3) Speech recognition

   The use of VAD reduces errors in the output of speech recognition systems.

4) Adaptive noise cancellation (ANC)

In ANC, if there is leakage of speech in the measured noise reference signal, then a VAD can be used to perform adaptation only during the pauses to prevent cancellation of the speech.

In VAD, the detection of speech boundaries is commonly achieved by first extracting features from the speech signal and then performing the speech/pause classification. The VAD must be robust to both the presence of different kinds of noise and varying levels of noise power. The extracted features can be compared to predefined threshold values for making the VAD decisions, and therefore, many VAD algorithms are sensitive to parameter tuning.

Based on the application, a consideration such as the computational cost of the VAD system may be a factor in its design. With hearing aids, low latency and low power consumption are crucial and a computationally efficient VAD must be used. In this section, a review of selected approaches for VAD in noisy speech signals will be first presented. Then, different evaluation measures for VAD algorithms will be overviewed. Finally, a discussion will be given.

2.7.1 VAD in noisy speech signals

Voice activity detection can be considered as a solution of the statistical hypothesis problem, where discriminative features can be first extracted from a measured speech signal which is usually contaminated by noise. The combination of these extracted features is then used for performing speech/pause classification. The VAD decision is normally given on a frame basis where adjacent frames which are detected to contain speech compose a speech segment. Using the model for the noisy speech $x[n]$ given in Equation (2.1), the hypotheses for the i^{th} frame, $\mathbf{x}_i$, are given by:

$$\begin{aligned} &H_0: \text{ speech absent: } \mathbf{x}_i = \mathbf{d}_i \\ &H_1: \text{speech present: } \mathbf{x}_i = \mathbf{s}_i + \mathbf{d}_i \end{aligned} \qquad \textbf{(2.23)}$$

where $\mathbf{s}_i$ and $\mathbf{d}_i$ are the i^{th} frame of the speech and noise respectively.

Typically, decision smoothing is performed on the output of the initial classification stage to reduce misclassification errors which may occur due to the decision rule. Hangover algorithms such as the Hidden Markov Model (HMM) based scheme proposed in (Sohn et al., 1999) are employed to extend the detected speech segments. This extension allows inclusion of neighbouring speech frames which contain low energy speech at the start and at the end of utterances. Another example of a hangover scheme is presented in the ETSI standard (ETSI, 1996) which uses a fixed hangover extension for detected speech segments which are longer than a predefined length. The generalized VAD procedure is summarized in Figure 2.9 for the input noisy speech frame, $\mathbf{x}_i$.

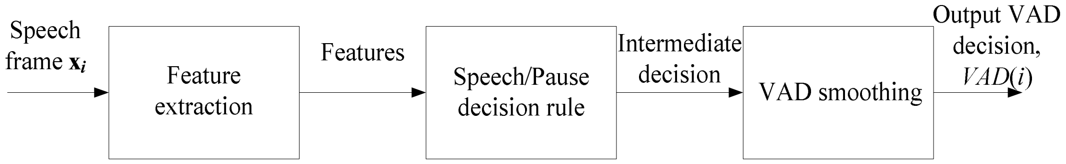

Speech frame $\mathbf{x}_i$
Feature extraction
Features
Speech/Pause decision rule
Intermediate decision
VAD smoothing
Output VAD decision, $VAD(i)$

Figure 2.9: Generalized VAD procedure

It must be noted that the output VAD decision, $VAD(i)$, may be a hard decision where the binary values of 1 and 0 denote the speech and pauses respectively. Soft decisions may be made for the VAD as shown in the frequency domain based VAD algorithms in (Sohn and Sung, 1998, Sohn et al., 1999). In (Tahmasbi and Rezaei, 2007), multiple observations of the soft decision are used to obtain a hard VAD decision.

2.7.1.1 Review of VAD techniques

VAD systems base their decisions on various speech features. Some of the traditional features used include periodicity (Tucker, 1992), cepstral features (Fukuda et al., 2010) and zero-crossing rate (ZCR) (Junqua et al., 1991). However, it is critical that these features are very robust in the presence of contaminating noise. The statistical hypothesis problem from Equation (2.23) was formalized (Sohn et al., 1999) in the frequency domain, and the DFT coefficients of the speech, the noise and the noisy speech were assumed to be asymptotically independent Gaussian random variables (Ephraim and Malah, 1984), as in Section 2.3.1. The statistical model based VAD decision was then formulated as:

$$p\left(H_1 \middle| X(k,i)\right) \underset{H_0}{\overset{H_1}{\gtrless}} p\left(H_0 \middle| X(k,i)\right) \tag{2.24}$$

where $X(k,i)$ is the k^{th} DFT coefficient of the i^{th} frame for the noisy speech and $p\left(H_a \middle| X(k,i)\right)$ is the a posteriori probability for speech presence (a=1) or speech absence (a=0). From Equation (2.24), Bayes rule gives:

$$p\left(X(k,i) \middle| H_1\right) p\left(H_1\right) \underset{H_0}{\overset{H_1}{\gtrless}} p\left(X(k,i) \middle| H_0\right) p\left(H_0\right) \tag{2.25}$$

This was reduced to the statistical likelihood ratio test $\Lambda(k,i)$ for the k^{th} DFT coefficient:

$$\Lambda(k,i) \triangleq \frac{p\left(X(k,i) \middle| H_1\right)}{p\left(X(k,i) \middle| H_0\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{p\left(H_0\right)}{p\left(H_1\right)} \tag{2.26}$$

From Equation (2.26), the final expression for the VAD decision rule in (Sohn et al., 1999) was given by the mean of the log likelihood ratios $\Lambda(k,i)$ over the K DFT bins as:

$$\frac{1}{K}\sum_{k=0}^{K-1} \log \Lambda(k,i) = \frac{1}{K}\sum_{k=0}^{K-1}\left[\frac{\gamma(k,i)\xi(k,i)}{1+\xi(k,i)} - \log\left(1+\xi(k,i)\right)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{2.27}$$

where the a priori SNR $\xi(k,i)$ was defined in Equation (2.13). The a posteriori SNR $\gamma(k,i)$ is defined as:

$$\gamma(k,i) \triangleq \frac{R_{XX}(k,i)}{R_{DD}(k,i)} \qquad \textbf{(2.28)}$$

and $R_{XX}(k,i)$ is the noisy speech spectrum and $R_{DD}(k,i)$ is the noise spectrum. The decision threshold is given by η. In (Sohn et al., 1999), $\xi(k,i)$ was estimated using the decision-direction approach from (Ephraim and Malah, 1984) to reduce the bias towards speech presence from the expression given in Equation (2.27). This VAD technique also requires the estimation of the noise statistics using one of the methods outlined in Section 2.3. The effect of using a different underlying distribution for the data was shown in (Gazor and Zhang, 2003).

Energy-based VAD techniques have been proposed which attempt to compare the power of the analyzed frame against a threshold value which is dependent on the noise power. The comparison of the analyzed frame's power against the threshold is used to make the VAD decision. However, in the presence of non-stationary noise, the noise power, $\hat{\lambda}_d(k,i)$, can be updated by recursive averaging as given by:

$$\hat{\lambda}_d(k,i+1) = \alpha_d \hat{\lambda}_d(k,i) + [1-\alpha_d]|X_a(k,i)|^2 \qquad \textbf{(2.29)}$$

where the smoothing factor α_d ranges from [0, 1] as in Section 2.3.1. Many VAD algorithms such as (Marzinzik and Kollmeier, 2002, Ramírez et al., 2005) initialize the noise power estimate by assuming an initial noise-only period, after which the noise spectrum is updated, usually during the detected speech absent frames. In (Tanyer and Özer, 2000), a geometrically adaptive energy threshold (GAET) was employed to estimate the noise power, and this measure was combined with periodicity estimation to perform VAD. In (Yu and Hansen, 2010), a microphone array based VAD was proposed for detecting the driver's speech in a vehicle. The a priori knowledge of the driver's spatial power distribution in a vehicle is used for obtaining an estimate of the driver's speech power and an estimate of the background noise power. A comparison of these power estimates was used in the VAD decision. In (Marzinzik and Kollmeier, 2002), the noisy speech signal was filtered to obtain the low-pass and the high-pass temporal power envelopes. The dynamics of these envelopes were the features used, and the speech absence classification was based on the frequency characteristics of the contaminating noise source, in order to improve the accuracy of the VAD decision. If the noise power dominates either the high-pass or the low-pass band of the noisy speech, the alternate frequency band is used for the detection. For broadband noises which do not dominate either the high-pass or the low-pass band, the speech absence decision is based on a combination of the detection in both frequency bands.

Higher order statistics (HOS) measures using 3rd and 4th order cumulants have been used for VAD. The 3rd and 4th order cumulants for a Gaussian random process are identically zero (Therrien, 1992). In (Nemer et al., 2001), the HOS of speech were shown to be non-zero and the higher order cumulants were used to distinguish speech from Gaussian noise. In (Li et al., 2005), the HOS metric was combined with the energy ratio of the low-pass and the high-pass frequency bands of the signal to improve VAD performance when non-Gaussian noises were present. Also, this improved detection due to the Gaussian-like nature of unvoiced speech components.

VAD algorithms which use multiple observations based on long term speech information are discussed in (Ramírez et al., 2007), and these are shown to offer improvements to some of the already mentioned techniques. Many proposed techniques were often compared to the well-known and computationally simple ITU-T standard G.729B VAD (ITU-T, 1996) which combines measures of spectral distortion, short-term energy in the full-band and low-pass band, and ZCR for a VAD decision on segmented speech with frames of length 10 ms. The individual features used in VAD algorithms have been shown in the literature to not always be robust to different levels of non-stationary noises. Therefore, VAD methods tend to fuse individual features for overall improvement in detection accuracy.

2.7.2 Performance Evaluation for VAD Techniques

The performance of VAD techniques can be evaluated on frame-level detection or utterance-level detection as detailed in (Kitaoka et al., 2007). In this thesis, frame-level performance of VAD will be considered for noisy speech contaminated by different types of noise over a range of SNRs. The speech hit-rate (*HR*1) and speech absence hit-rate (*HR*0) will be used to measure the VAD accuracy. These measures are defined as:

$$HR1 = \frac{N_{1,1}}{N_1^{ref}} \qquad \textbf{(2.30)}$$

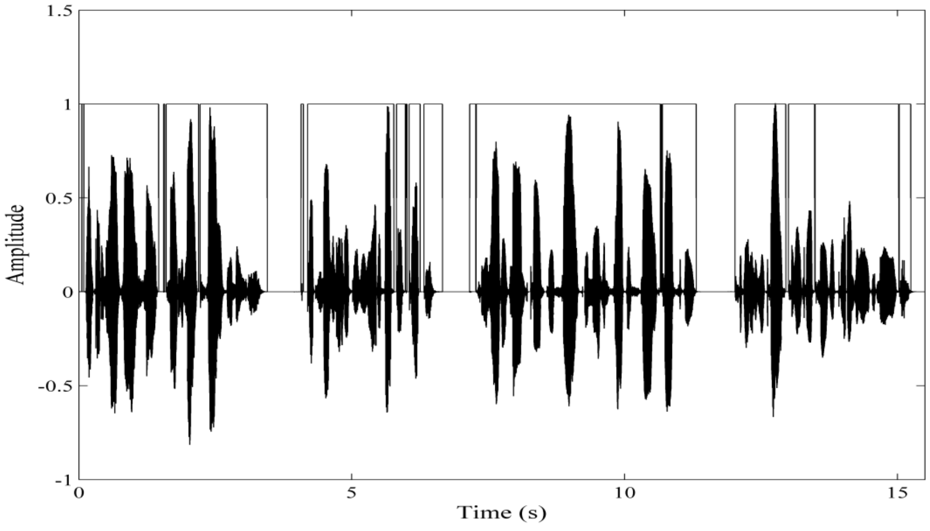
$$HR0 = \frac{N_{0,0}}{N_0^{ref}} \qquad \textbf{(2.31)}$$

where $N_{1,1}$ denotes the number of frames correctly classified as speech and $N_{0,0}$ denotes the number of frames correctly classified as speech absent frames. N_1^{ref} and N_0^{ref} are the actual number of speech and non-speech frames respectively. The false alarm rate for speech presence, *FAR*1 and for speech absence *FAR*0 can be correspondingly determined by:

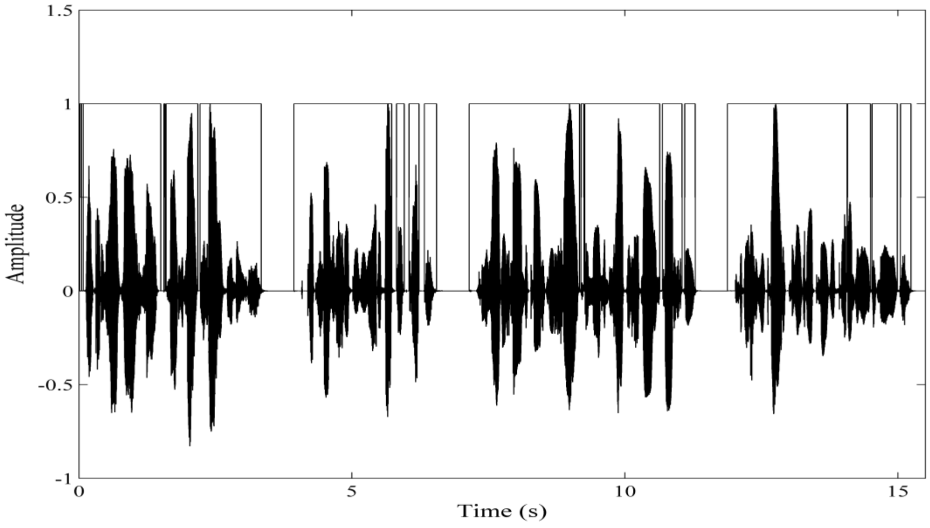
$$FAR1 = 1 - HR0 \qquad \textbf{(2.32)}$$

$$FAR0 = 1 - HR1 \qquad \textbf{(2.33)}$$

These aforementioned measures require the actual number of speech and non-speech frames. Hand-labelling of the speech pauses is often done, however, in (Marzinzik and Kollmeier, 2002) the actual speech pauses were labelled using the G.729B VAD algorithm on the clean speech sequences. It was found that many VAD algorithms tend to have a conservative approach to speech pause detection, where, the *FAR*1 measure is allowed to increase as the SNR of the noisy speech decreases. This increase in the incorrect detection of speech pauses as speech frames has an adverse effect on techniques which update the noise spectrum estimate during speech absence. In the presence of non-stationary noise, the noise power estimate will not be updated regularly and there will be poor tracking of changes in the noise power, which will result in a decline in VAD performance.



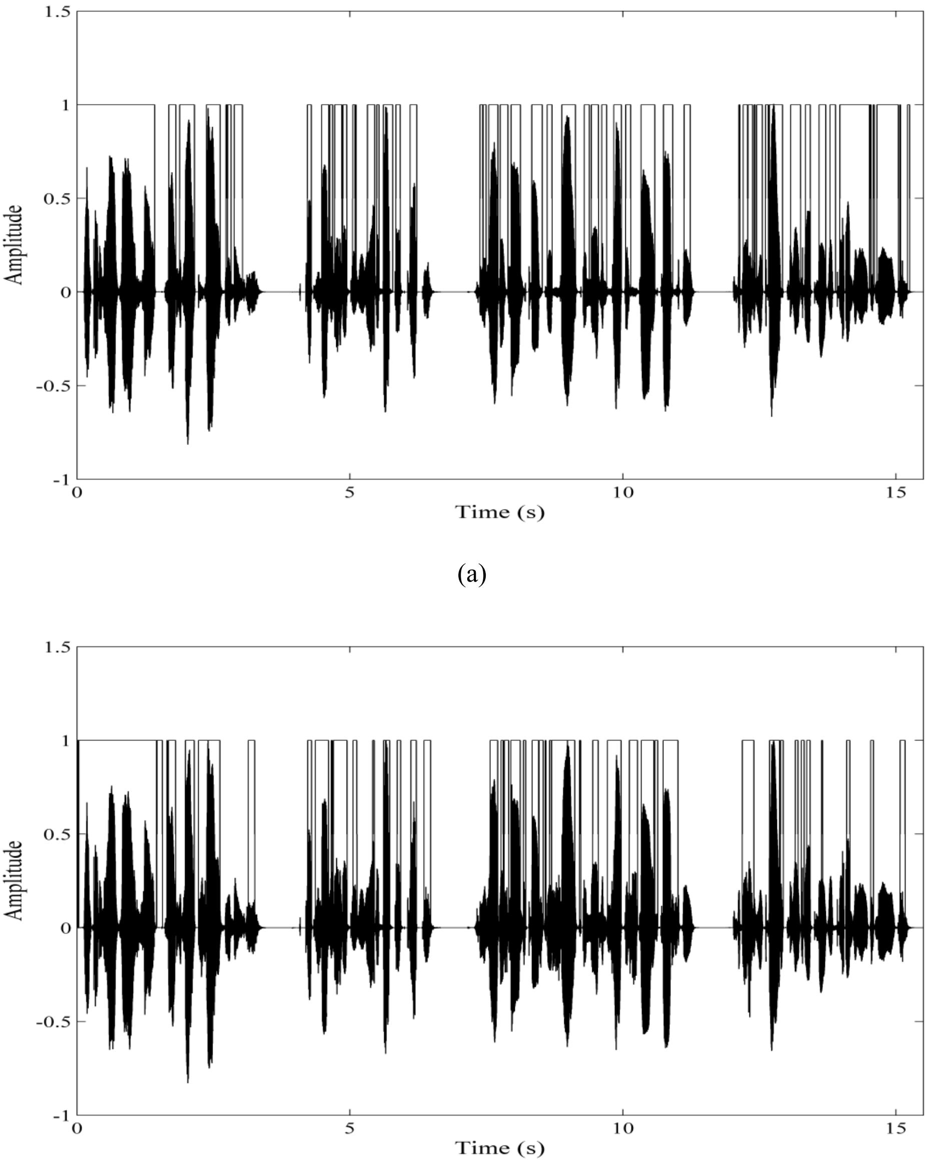
(a)



(b)

Figure 2.10: VAD decisions on the clean speech sequence using (a) G.729B VAD and (b) Sohn's VAD

The results of the VAD detection using G.729B on a clean speech sequence can be seen in Figure 2.10(a). The test speech of approximate duration of 15 seconds, was artificially concatenated using three female and one male speech sentences from the TIMIT database (John S. Garofolo et al., 1993). Figure 2.10(b) shows the detection results on the same clean speech sequence using the statistical model-based VAD from (Sohn et al., 1999) which was described in Section 2.7.1.1. The detection performance for both of these VAD algorithms was evaluated using the same speech sequence contaminated with F16 cockpit noise and car interior noise from the Noisex-92 database (Varga and Steeneken, 1993) at 5 dB SNR in Figure 2.11 and Figure 2.12 respectively. A minimum statistics based approach (Martin, 2001) was employed to obtain the noise estimate required for Sohn's VAD. The VAD decisions from both techniques are superimposed on the original clean speech sequence to indicate the algorithm's performance.

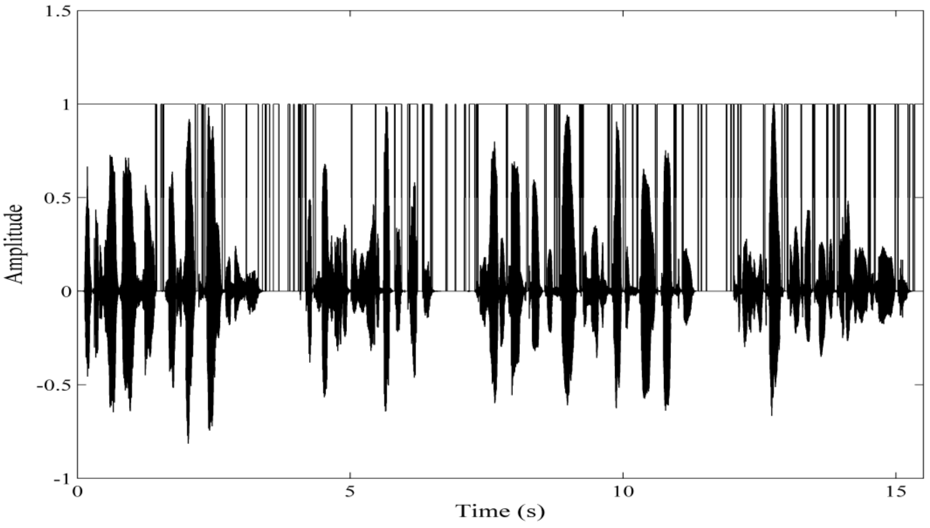


(a)

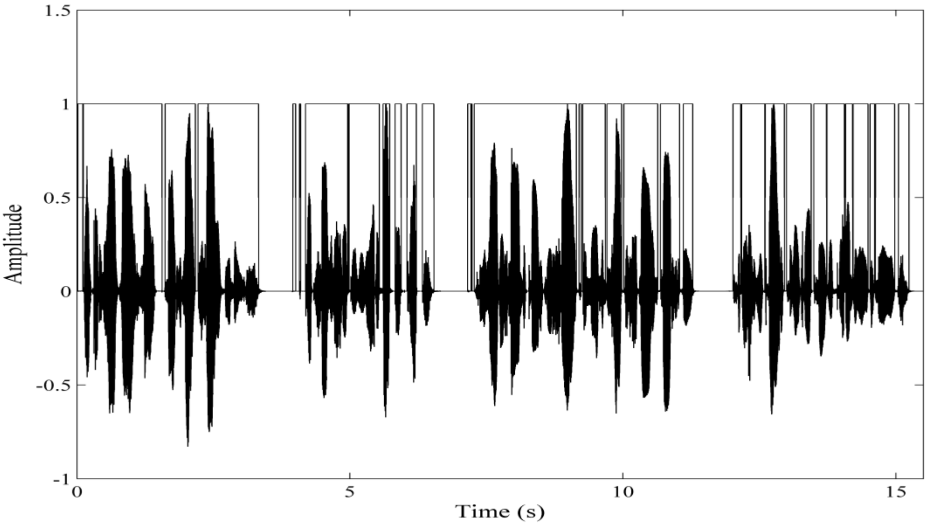
(b)

Figure 2.11: VAD decisions at 5db SNR in F16 cockpit noise using (a) G.729B VAD and (b) Sohn's VAD

Figure 2.11 shows that both techniques perform quite poorly at 5 dB SNR in a cockpit noise environment with many misdetections of speech absence which resulted in a high *FAR*0. Under these noisy conditions, the G.729B is seen to perform slightly better than Sohn’s VAD.



(a)



(b)

Figure 2.12: VAD decisions at 5db SNR in car interior noise using (a) G.729B VAD and (b) Sohn’s VAD

Figure 2.12 shows that the G.729B VAD has a high *FAR*1 at 5 dB SNR in a car interior noise environment with many misdetections of speech presence. Under these noisy conditions, Sohn's VAD is seen to have improved performance. It was discussed in 2.7.1.1 that Sohn's VAD relied on a noise estimate. Therefore, its performance also relies on the accuracy of the noise estimation routine employed. Car interior noise is relatively stationary compared to F16 cockpit noise and this should result in lower errors in the noise estimate obtained and lead to improvements obtained using Sohn's VAD results under this noisy condition.

2.8 Conclusion

This chapter reviewed relevant signal processing algorithms for speech enhancement and VAD. The relatively new EMD was described as a multi-resolution approach which sifts data dependent basis functions, and is suitable for analyzing non-stationary signals. New EMD-based speech enhancement techniques will be introduced in the next chapter. The MS-based IMCRA noise estimation procedure was detailed. With this background, new EMD-based systems for speech enhancement will be presented in the next chapter.

A review of selected VAD techniques for noisy speech was given, along with performance evaluation measures. This evaluation was presented to give an appreciation to the challenge faced by VAD of discriminating speech activity and speech pauses, while maintaining low false alarm rates. The computationally efficient G.729 VAD was included in this evaluation and will be used for comparison with the novel VAD technique presented in Chapter 4.

3

EMD Based Techniques for Speech Enhancement

3.1 Introduction

Adaptive signal processing techniques such as adaptive noise cancellation (ANC) are commonly used when enhancing speech sequences when a noise reference is available, as opposed to fixed linear filters such as the Wiener filter. This is the case because the adaptive filter does not require prior knowledge of either the signal or the contaminating noise source. When a noise reference is unavailable, single channel speech enhancement techniques are employed, as discussed in Sections 2.3, 2.4 and 2.5.

In the case of non-stationary signals such as speech, a multi-resolution approach which incorporates the Empirical Mode Decomposition (EMD) may be more effective than spectral analysis techniques that use predefined basis functions. EMD performs data-driven spectral decomposition of non-stationary multi-component signals such as speech into Intrinsic Mode Functions (IMFs). In this chapter, EMD-based techniques for dual-channel and single-channel speech enhancement are developed.

This chapter is organized as follows. In Section 3.2, the proposed dual-channel enhancement technique is presented which combines the basic ANC with the EMD for improved signal enhancement. This method uses the entire bandwidth of the noise as the reference in an attempt to work iteratively in refining the noisy input speech, by performing the adaptive noise cancellation on an IMF level. The goal of speech enhancement using Adaptive Empirical Mode Decomposition (SEAEMD) is to break up a full-band task into a set of smaller sub-band tasks. This has potential when performed in highly non-stationary and time-varying environments. It is also envisaged that this technique can be extended to other applications of signal enhancement. In this section, results obtained from testing and comparing the SEAEMD to basic ANC in non-stationary and varying SNR conditions are presented and discussed. The results demonstrate that the SEAEMD system achieves significantly improved speech quality with a lower level of residual noise.

In Section 3.3, new EMD based filtering (EMDF) techniques for single-channel speech enhancement are proposed. It is shown that the proposed EMDF techniques are able to enhance the speech more effectively than a conventional optimally-modified log-spectral amplitude (OMLSA) approach from (Cohen and Berdugo, 2001) which uses an IMCRA noise estimate. In EMDF, EMD is first used to decompose the noisy speech into its IMFs. A method is then developed to select the IMF index for separating the noise components from the speech based on the IMF statistics. The denoised speech signal is separated by a partial reconstruction from the IMFs formed from the noisy speech. The first proposed system has been designed to be particularly effective in low frequency noise environments such as car interiors. The second EMDF system is able to enhance speech signals in various non-stationary noise environments. Comparative performance studies are included that demonstrate the superiority of the EMDF system compared to the full-band OMLSA/IMCRA speech enhancement. It shows that the EMDF system has significant potential when performed in highly non-stationary and time-varying environments compared to a full-band OMLSA/IMCRA system. Finally, concluding remarks are presented in Section 3.5.

3.2 Speech Enhancement using Adaptive Empirical Mode Decomposition (SEAEMD)

ANC is a popular technique for noise removal from time-varying speech signals, by attempting to minimize the output error of the filter. It accomplishes this by adaptively filtering the noisy speech input, with a reference noise signal. The new SEAEMD is illustrated in Figure 3.1, where $s[n]$ is the original speech, $d_0[n]$ is the contaminating noise and $d_1[n]$ is the reference noise signal. The EMD of the speech signal is first performed to decompose the signal into its corresponding IMFs. The resulting IMFs contain a mixture of speech and noise and these IMFs are then adaptively filtered using the noise reference. The IMFs that are output from the adaptive filter in the SEAEMD model are then used to reconstruct the enhanced speech signal.

In this section, the analysis of the novel SEAEMD for speech is first presented. Next, results obtained from testing and comparing the SEAEMD to basic ANC in non-stationary and varying SNR conditions are presented and discussed.

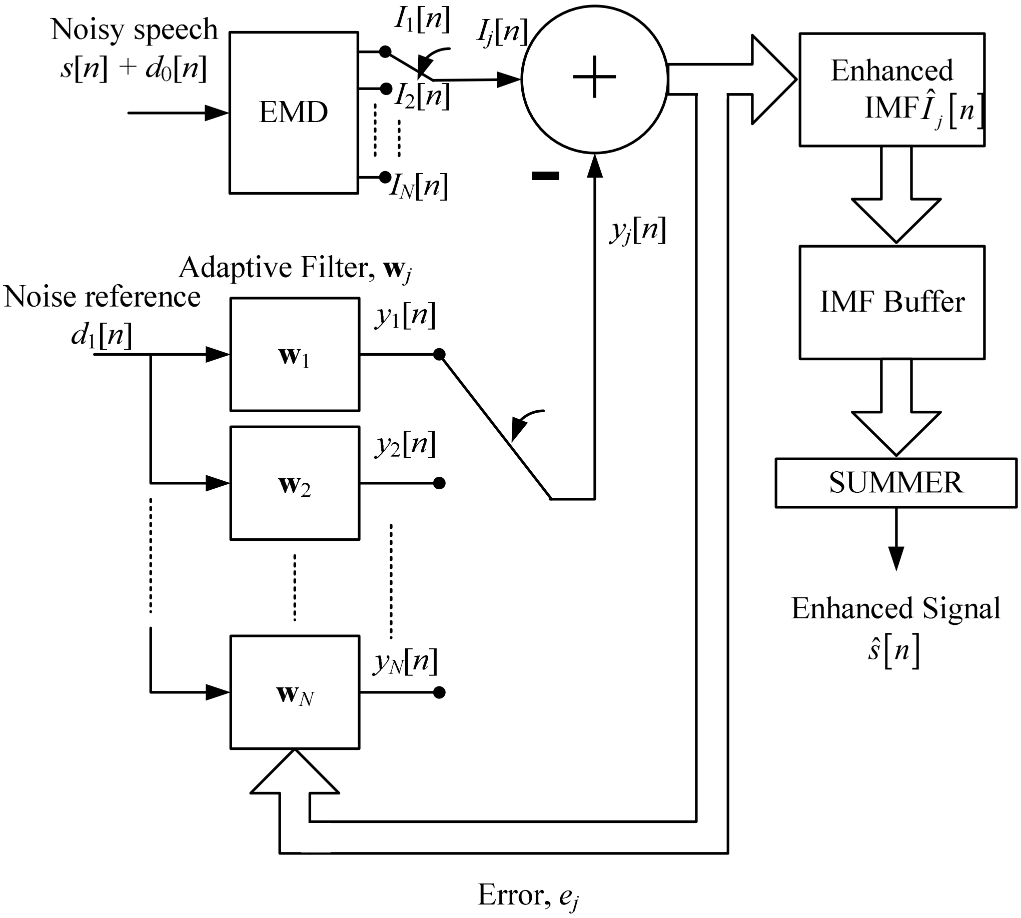


Figure 3.1: SEAEMD model

3.2.1 Analysis of SEAEMD

It is assumed in Figure 3.1 that $d_0[n]$ is correlated in some way to $d_1[n]$ and $s[n]$ is uncorrelated with $d_0[n]$ and $d_1[n]$. A total of N IMFs are formed from the EMD of the input noisy speech denoted by $I_j[n]$, $j=1..N$. Therefore, $I_j[n]$ is the j^{th} IMF of the noisy speech signal that contains both noise and speech components as described below by:

$$I_j[n]=s_j[n]+d_{0,j}[n] \tag{3.1}$$

$y_j[n]$ is the output of the j^{th} adaptive filter with coefficients $\mathbf{w}_j$. The error signal, $e_j[n]$, is given by:

$$e_j[n]=I_j[n]-y_j[n] \tag{3.2}$$

The MSE is therefore evaluated as:

$$E\left[e_j^2[n]\right]=E\left[\left(I_j[n]-y_j[n]\right)^2\right] \tag{3.3}$$

$$E\left[e_j^2[n]\right]=E\left[s_j[n]^2\right]+E\left[\left(d_{0,j}[n]-y_j[n]\right)^2\right]+2E\left[s_j\left(d_{0,j}[n]-y_j[n]\right)\right] \tag{3.4}$$

where $E[.]$ is the expectation operator. Since $d_{0,j}[n]$ is uncorrelated with s_j, the adaptive filter $\mathbf{w}_j$ updates its filter coefficients to minimize the MSE signal from Equation (3.4) as shown below in:

$$E_{\min}\left[e_j^2[n]\right]=E\left[s_j[n]^2\right]+E\left[\left(d_{0,j}[n]-y_j[n]\right)^2\right] \tag{3.5}$$

where the speech component in that IMF, $I_j[n]$, is unaffected and the other term is minimized to lower the MSE signal. The noise will be adaptively filtered to feed into the system, to produce an error signal with which it is uncorrelated. Hence, the error signal will be an enhanced version of the j^{th} IMF. The enhanced speech signal, $\hat{s}[n]$, can then be reconstructed using the enhanced IMFs, $\hat{I}_j[n]$ as shown in Equation (3.6):

$$\hat{s}[n]=\sum_{j=1}^{N}\hat{I}_j[n] \tag{3.6}$$

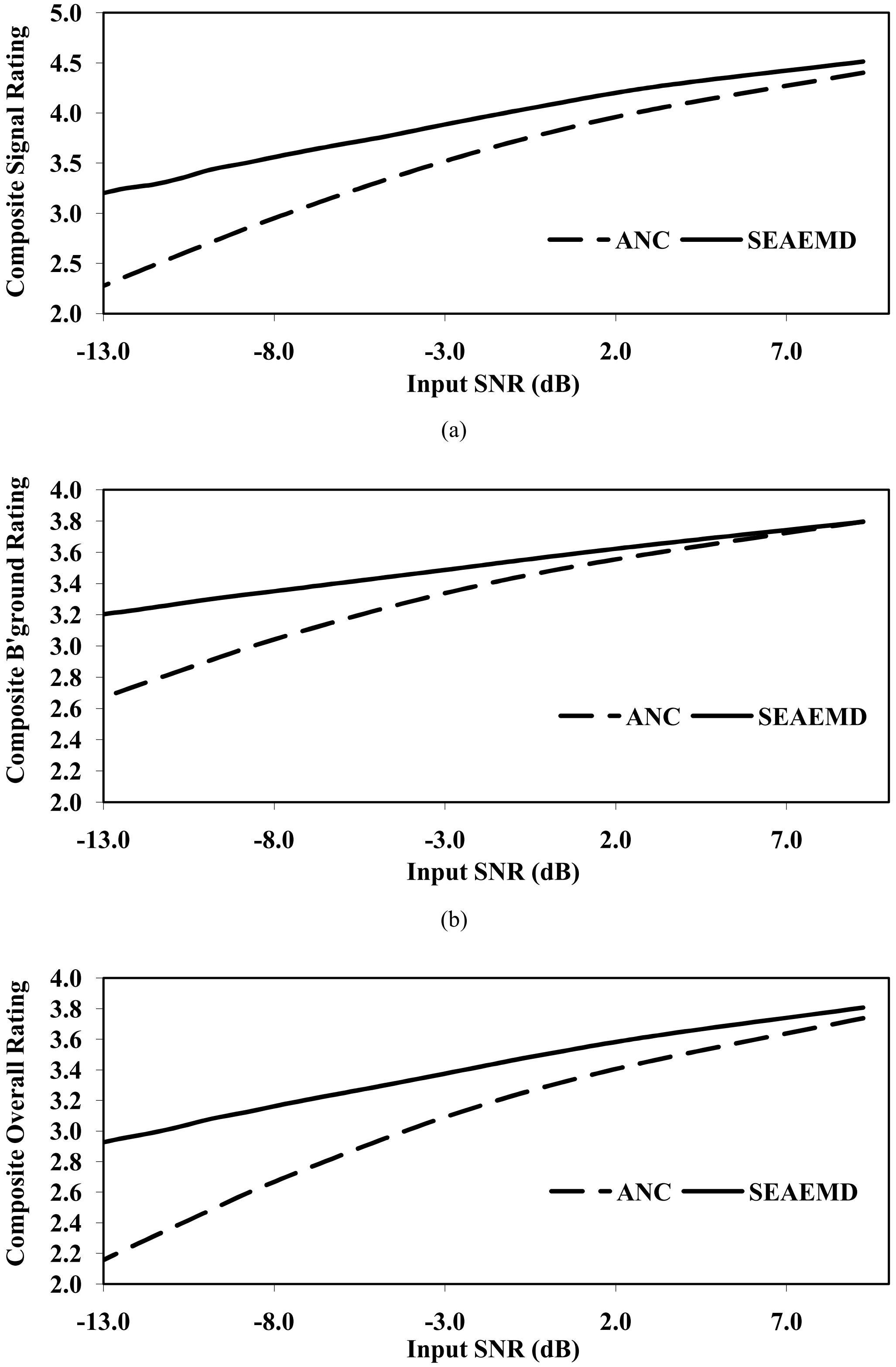
Note that the full-band noise was used as the reference in Figure 3.1 rather than its decomposed version. This was done because it is known that the corresponding IMFs from the noisy speech differ in content from those formed from the EMD of the additive noise signal.

3.2.2 Performance Evaluation

The performance of the SEAEMD was compared with the conventional ANC algorithm on speech signals contaminated with “real-life” non-stationary noises. A sampling frequency of 16 kHz was used and the noise reference signals (d_1 from Figure 3.1) input to the adaptive filter were obtained from the Noisex-92 database. They include factory noise, F16 cockpit noise and military vehicle noise. The contaminating noise d_0 from Figure 3.1 was generated by filtering d_1 with a high-pass FIR filter of order 40 and lower cut-off frequency of 0.05. The speech utterance was taken from the TIMIT database and degraded by the different noise types under varying SNR conditions. In both scenarios of ANC and SEAEMD, a 41 tap FIR adaptive filter, $\mathbf{w}_j$, with a standard NLMS (Normalized Least Mean Square) learning algorithm was used. The step size was varied from 0.1 to 0.005 and the filter taps were initialized to 1. The noisy speech signals were broken up into overlapping blocks of length 16,000 samples, and consecutively input into the relevant algorithm being tested.

In order to assess the performance of the SEAEMD algorithm, different subjective and objective assessment measures can be used as described in Section 2.6. The rating from the composite measure for measuring signal quality (C_{SIG}), background distortion (C_{BAK}) and overall quality (C_{OVL}) from Section 2.6, will be presented for evaluation of the SEAEMD algorithm. It must be recalled that these quality indices lie in the range 1 to 5.

Figure 3.2, Figure 3.3 and Figure 3.4 show examples of the improved quality of speech enhancement obtained by performing the SEAEMD as compared to ANC on speech signals contaminated with military noise, factory noise and F16 cockpit noise respectively. These results demonstrate that in low SNR, non-stationary adverse conditions, the performance of the conventional ANC degrades while the SEAEMD continues to provide consistent and superior levels of speech enhancement and lower levels of residual noise. Listeners appear to be particularly sensitive to speech distortion (Hu and Loizou, 2008) and as shown in Figure 3.2(a), Figure 3.3(a), and Figure 3.4(a), the SEAEMD algorithm gives the desired significant improvement, as the noise level increases. This occurs since the enhancement was performed on an IMF level with the SEAEMD technique, and therefore demonstrates that this approach to noise cancellation is more effective by focusing enhancement efforts on the noisy IMFs. Another advantage of the SEAEMD technique is that it retains high-frequency components of the denoised signal, as opposed to smoothing algorithms.

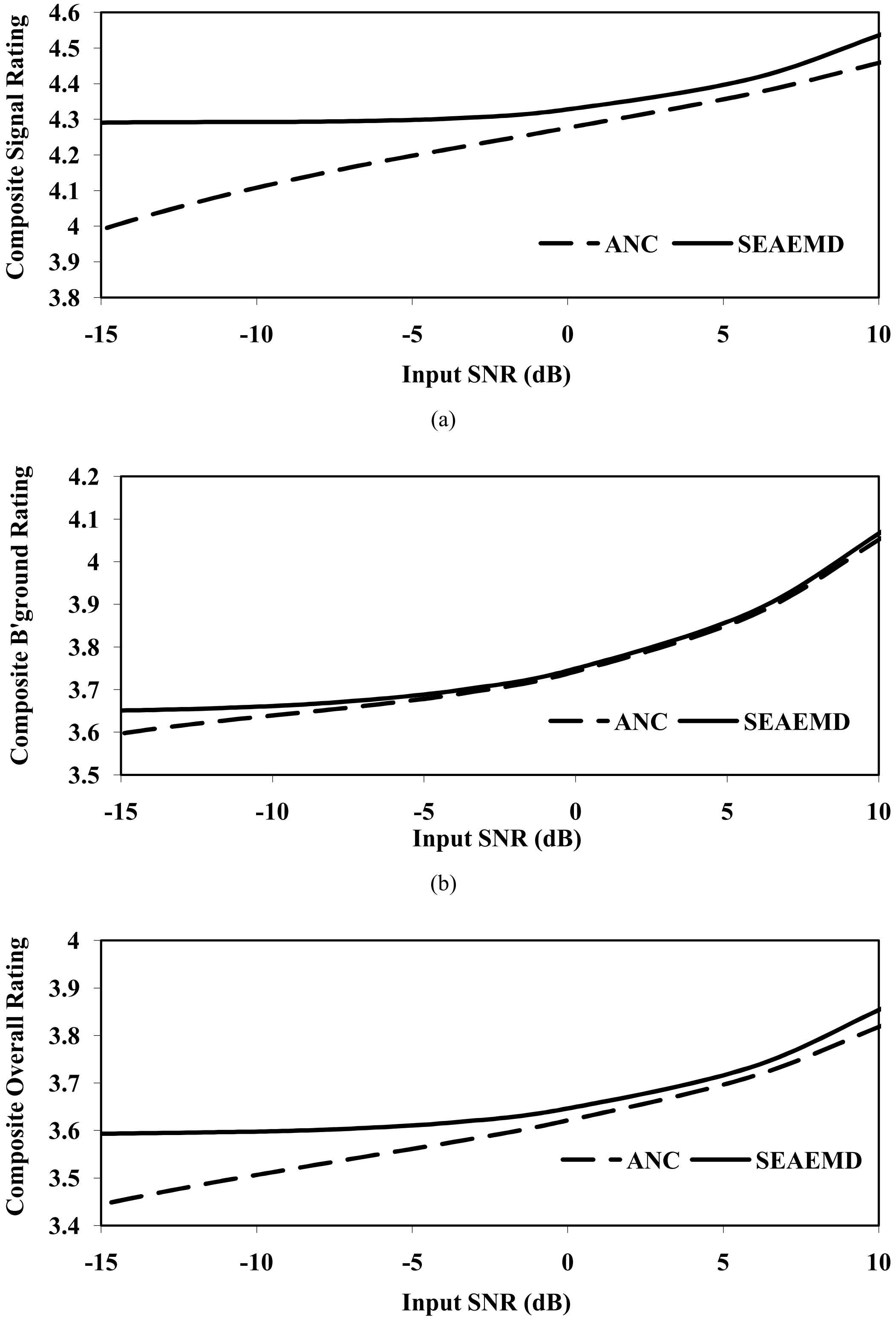


(a)

(b)

(c)

Figure 3.2: Comparison of speech enhancement using SEAEMD and ANC for speech contaminated with different levels of military vehicle noise. (a)C_{SIG} (b)C_{BAK} (c)C_{OVL}

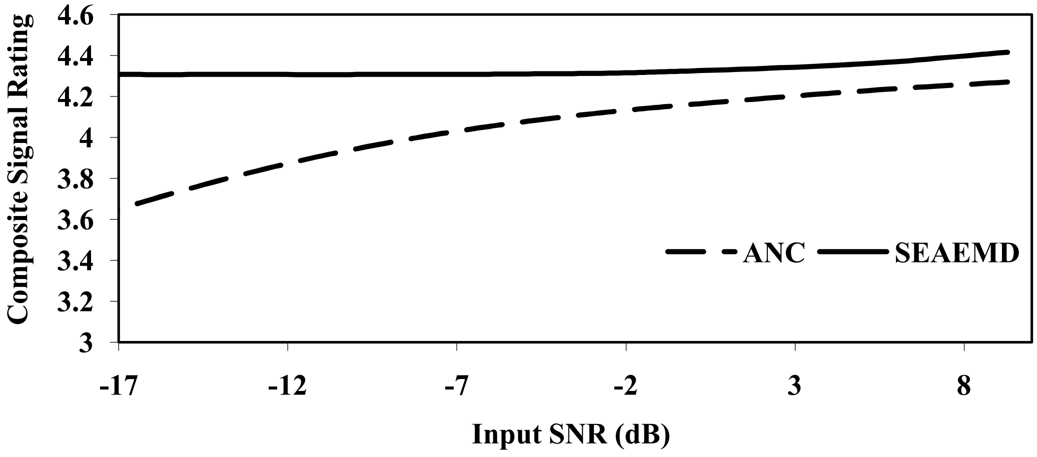


(a)

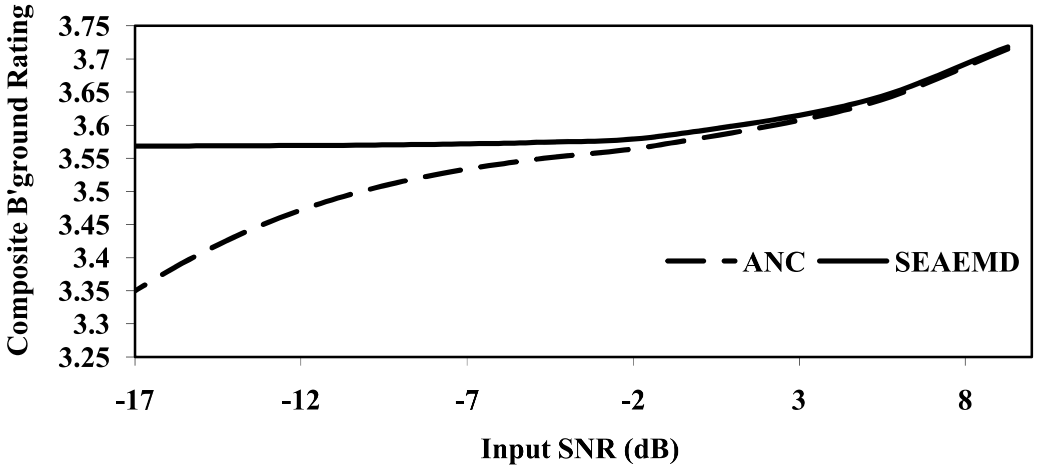
(b)

(c)

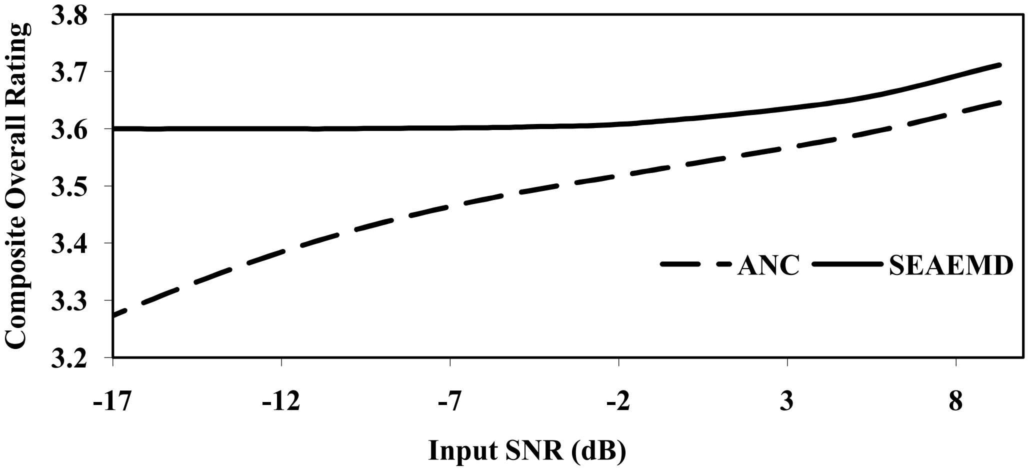
Figure 3.3: Comparison of speech enhancement using SEAEMD and ANC for speech contaminated with different levels of factory noise. (a)C_{SIG} (b)C_{BAK} (c)C_{OVL}



(a)

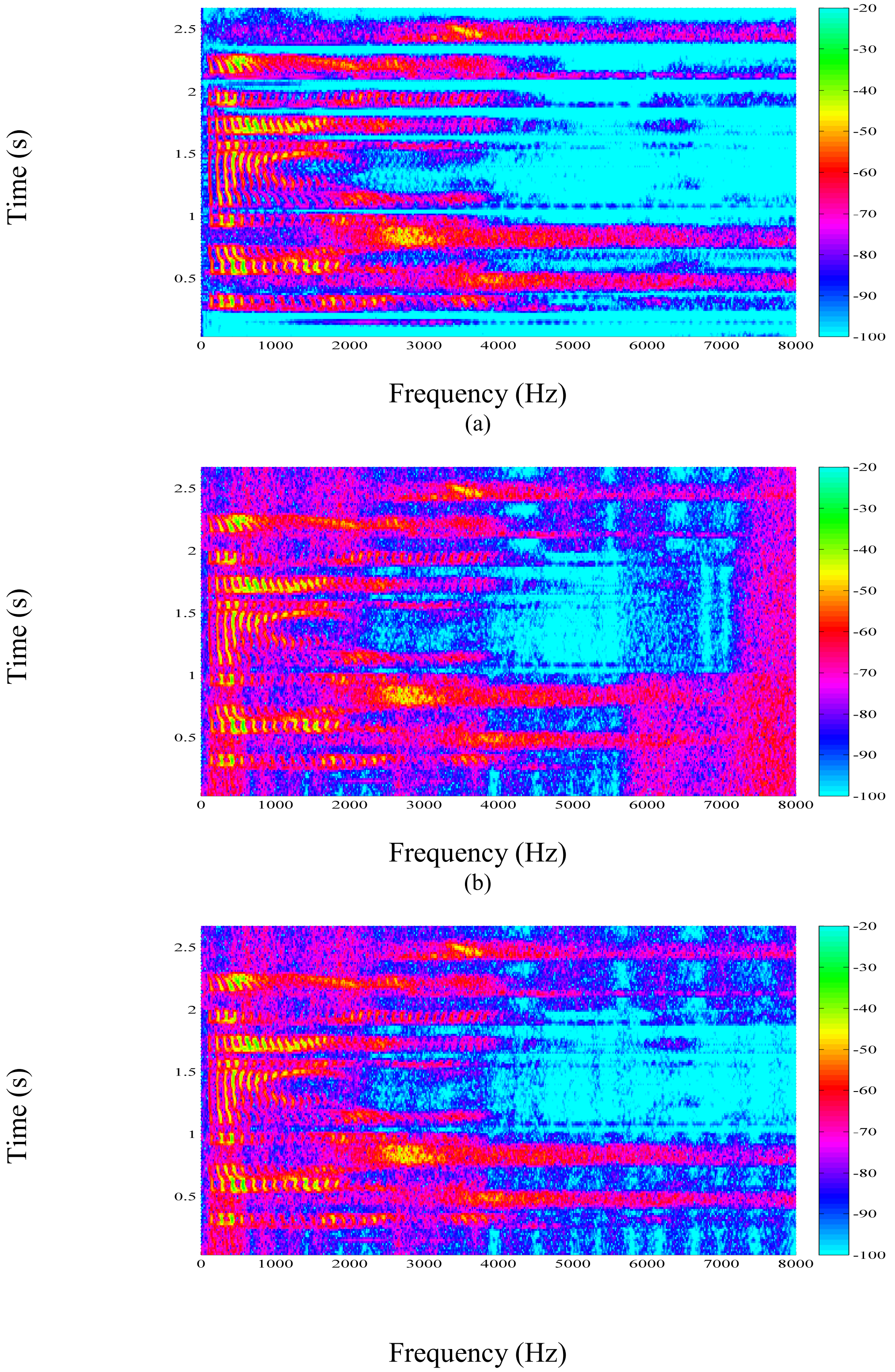


(b)



(c)

Figure 3.4: Comparison of speech enhancement using SEAEMD and ANC for speech contaminated with different levels of F16 cockpit noise. (a)C_{SIG} (b)C_{BAK} (c)C_{OVL}



(a)

(b)

(c)

Figure 3.5: Comparison of the spectrograms for speech enhanced by both ANC and SEAEMD methods in F16 cockpit noise at -20 dB. (a) Original clean speech (b) Speech enhanced by ANC (c) Speech enhanced by SEAEMD

The spectrogram of a clean male speech utterance is shown in Figure 3.5(a). This signal is contaminated with F16 cockpit noise at -20 dB SNR and enhanced using both conventional ANC and SEAEMD techniques and the spectrograms of the enhanced signals are shown in Figure 3.5(b) and Figure 3.5(c) respectively. These plots demonstrate the improved noise suppression of this high frequency cockpit noise using the new SEAEMD technique.

3.3 EMD-based Filtering for Speech Enhancement in Low Frequency Noise

Single channel speech enhancement algorithms rely on accurate noise spectrum estimation and speech estimation. IMCRA was introduced in Section 2.3.1, and in (Cohen, 2003), it was shown that eliminating strong speech segments from the second smoothing stage in IMCRA improves minima tracking and the estimation of the speech presence probability. The speech presence probability is dependent on the ratio of the instantaneous and smoothed signal power spectrum to the minima power spectral values. In frequency bins with low SNR, the noise power dominates the speech power. Using IMCRA in low frequency noise environments such as in car interiors, there is poor noise estimation from the noisy low frequency bins since the smoothing factor, $\tilde{\alpha}_d$ from Equation (2.9), tends to 1.

A technique for EMD-based noise estimation and tracking (ENET) with application to speech enhancement is now presented. The ENET system is illustrated in Figure 3.6. Consider the model described by:

$$x[n] = s[n] + d[n] \tag{3.7}$$

where $x[n]$ is the noisy speech signal, $s[n]$ is the original noise-free speech, and $d[n]$ is the noise source which is assumed to be independent of the speech. The EMD pre-processing stage from Figure 3.6 decomposes the signal into two signal spaces which are useful for noise tracking and speech estimation. It can be interpreted as:

$$x[n]=\sum_{j=1}^{M} I_j\left[n\right]+\sum_{j=M+1}^{N} I_j\left[n\right]=\sum_{c=1}^{2} B_c\left[n\right] \qquad \textbf{(3.8)}$$

where the EMD of the noisy signal $x[n]$ produces N IMFs. From Equation (3.8), when c=1, let $B_c\left[n\right]$ denote the band that contains stronger speech components as well as some noise. When c=2, let $B_c\left[n\right]$ contain the residual noise as well as the weaker speech. Correspondingly, let $B_c\left(k,i\right)$ denote the STFT of $B_c\left[n\right]$. ENET then estimates the noise power spectrum in each band individually using the IMCRA routine, and enhances each band independently using the OMLSA speech estimate.

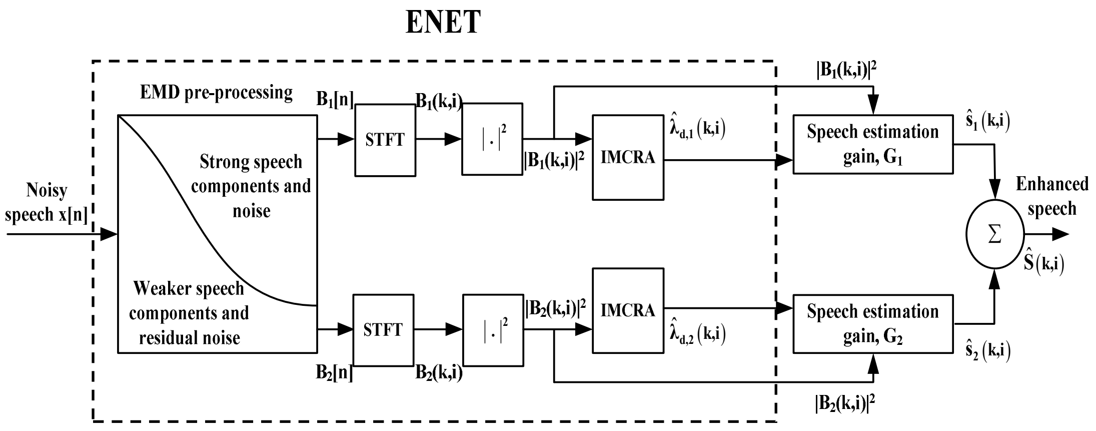


Figure 3.6: Block diagram of ENET with application to speech enhancement

The frequency characteristics of the contaminating noise determine the value of M from Equation (3.8). However, no method has been formally proposed for the selection of M. The ENET system is refined to form the new EMD-based Filtering (EMDF) system for speech enhancement in *low frequency noise environments* as illustrated in Figure 3.7. Note that in our EMDF system, the lower branch of ENET has been removed, in order to reduce the low frequency noise components.

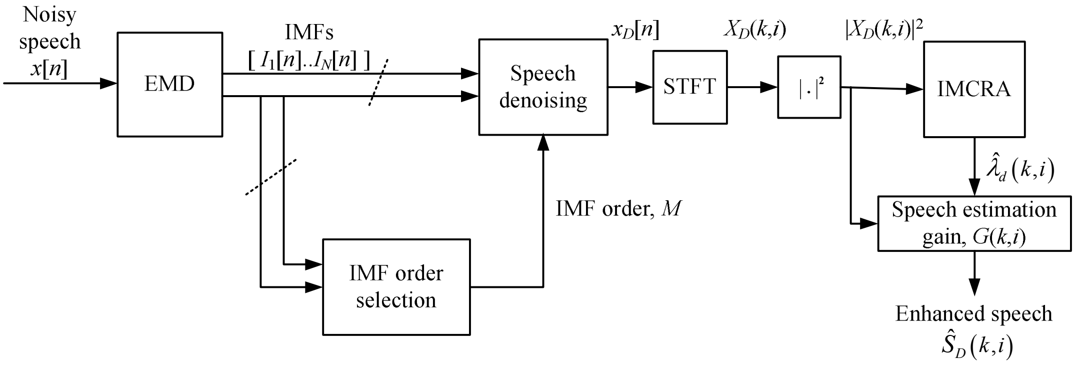


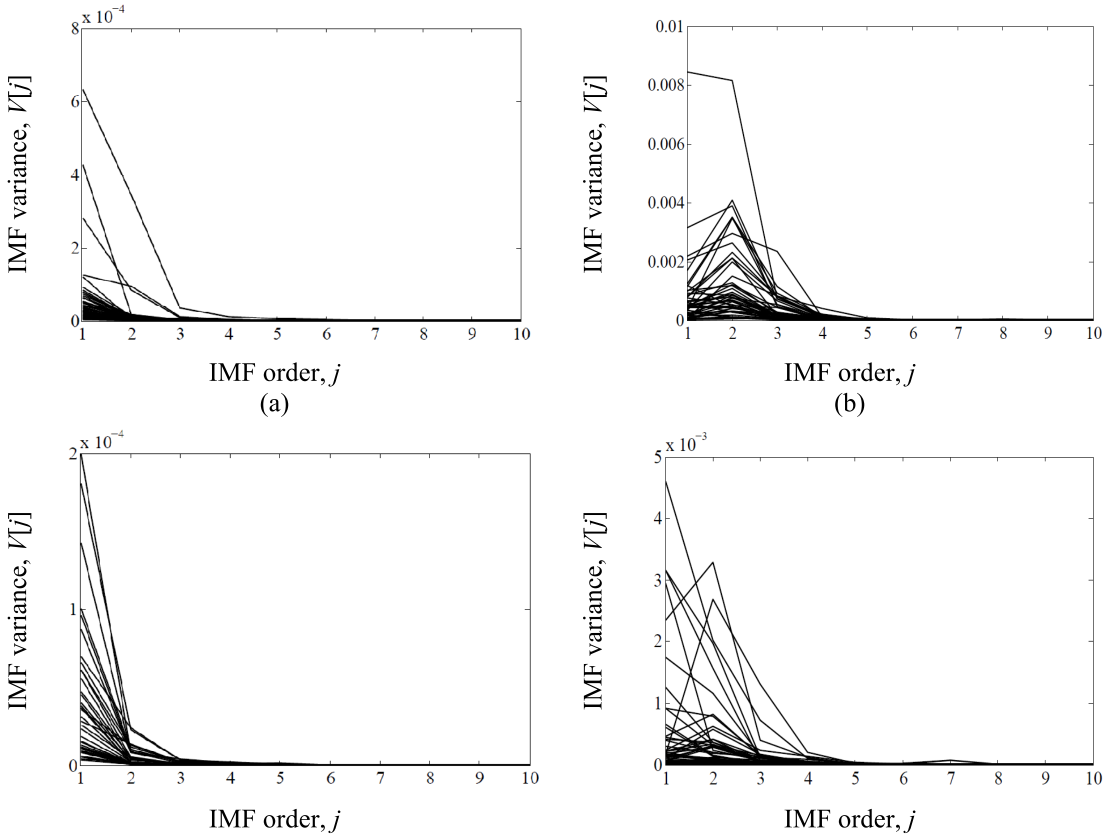
Figure 3.7: Block diagram of the EMDF system for speech enhancement in low frequency noise environments

The system analysis of the proposed EMDF system is first presented. It is seen that the noisy speech signal $x[n]$ is first denoised using our EMDF method before estimating the residual noise power components and the speech components. This is followed by the performance evaluation of the EMDF system, when compared with basic full-band OMLSA/IMCRA speech estimation. These tests are performed in non-stationary and varying SNR car interior noise conditions.

3.3.1 EMDF System Analysis

As seen in Figure 3.7, the EMD decomposes the noisy speech signal $x[n]$ into N IMFs. Consider the IMF variance plots shown in Figure 3.8 for random clean unvoiced and voiced speech components extracted from utterances spoken by various males and females from the TIMIT database. In these plots, the IMF order is denoted as j and the IMF variance is denoted as $V[j]$ where:

$$V[j]=\frac{1}{L}\sum_{n=1}^{L}I_j^2[n],\ j=1,2..N \tag{3.9}$$



(a) (b) (c) (d)

Figure 3.8: IMF variance plots of (a) clean male unvoiced speech components (b) clean male voiced speech components (c) clean female unvoiced speech components and (d) clean female voiced speech components

Partial reconstruction of these speech signals is redefined by:

$$x_D[n]=\sum_{j=1}^{M} I_j[n] \qquad \textbf{(3.10)}$$

where $I_j[n]$ denotes the j^{th} IMF. The SNR is used to objectively evaluate the resynthesis error of $x_D[n]$ compared to the original speech components. The SNR of the partially reconstructed signals using Equation (3.10) for clean unvoiced and voiced components spoken by a female, is given in Table 3.1(a) and Table 3.1(b) respectively. It can be seen that in both cases, signal reconstruction with the first 4 IMFs (i.e. M=4 in Equation (3.10)) is sufficient for good speech resynthesis. This is consistent with the low-rank approximation used in subspace algorithms as discussed in Section 2.4, which consider 9-15 dB SNR sufficient for reconstruction. Figure 3.8

shows that the IMF variance for clean speech signals significantly decreases after the fourth IMF, as the IMF order increases. It was found experimentally that the IMF statistics for a speech signal contaminated with a low frequency noise has a peak IMF energy in a higher IMF order $I_j[n]$, where $j>4$.

IMF order, *j*	SNR (dB) of $x_D[n]$
1	12.7
2	16.9
3	20.5
4	22.9
5	25.5
6	28.0
7	28.9
8	30.8

(a)

IMF order, *j*	SNR (dB) of $x_D[n]$
1	1.3
2	7.5
3	14.0
4	17.5
5	21.9
6	28.5
7	31.9

(b)

Table 3.1: SNR of partially reconstructed signals using *j* IMFs for (a) clean unvoiced speech segment and (b) clean voiced speech segment

An example of the IMF variance plot for a clean voiced speech female utterance $s[n]$ contaminated with car interior noise $d[n]$ at 0 dB SNR is shown in Figure 3.9. The peak a_1 (considering IMFs of order $j>4$) is highlighted along with its associated variance build-up b_1 to that peak. Identification of this IMF variance deviation is used to select the IMF order, *M*, to use in the speech reconstruction. The remaining IMFs from $M+1$ to *N* are assumed to be dominated by the noise whereas in Equation (2.16), these IMFs were used to reconstruct the desired signal which was contaminated by fGN. Therefore, the denoised signal $x_D[n]$ is obtained from the partial reconstruction in Equation (3.10).

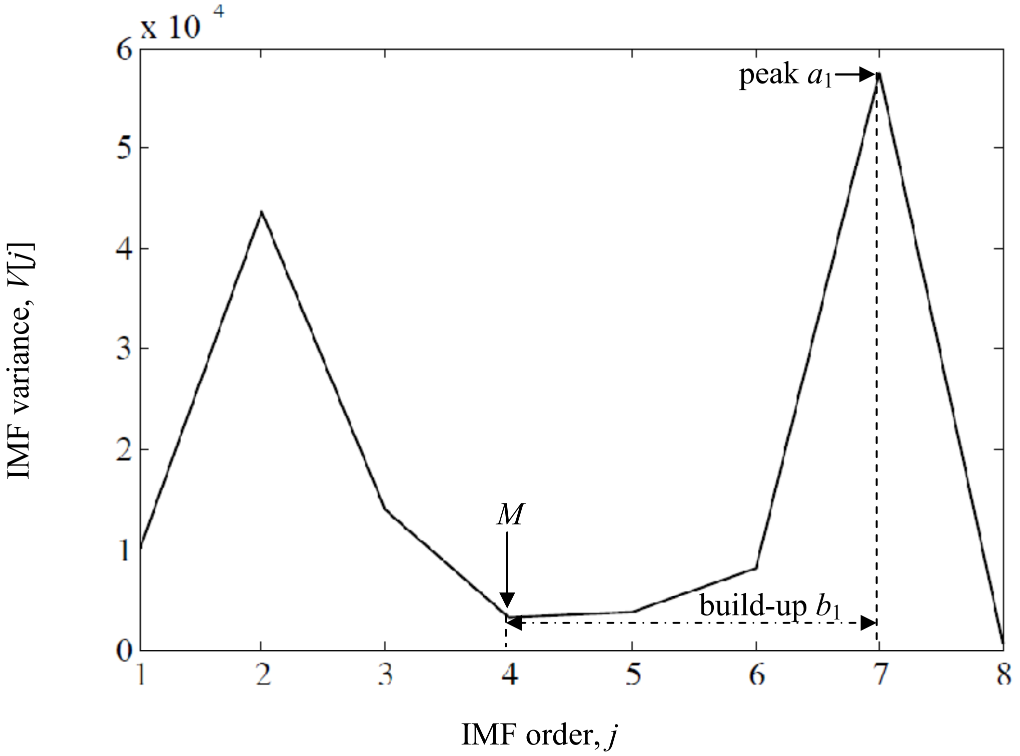


Figure 3.9: IMF variance plot of clean speech contaminated with car interior noise at 0 dB SNR

Our method to select the IMF index M is described as follows:

1. Compute the variance $V[j]$ of the j^{th} IMF, I_j from Equation (3.9). Identify the indices of the peaks, $\mathbf{a}=a_1,a_2...$, in $V[j]$ for $j>4$ and compute the corresponding variance build-up, $\mathbf{b}=b_1,b_2...$ to those peaks.
2. Determine the largest build-up b_m in **b** and select the corresponding peak a_m in **a**.
3. The IMF index M is determined by examining the valley in $V[j]$ prior to the peak a_m. M is constrained to be greater than or equal to 4 and is evaluated as follows:

$$M = \max\left(4, a_m - b_m\right) \tag{3.11}$$

The IMF variance plot of a noisy speech shown in Figure 3.9 is used as an example to demonstrate the above algorithm for selecting M. The peak $\mathbf{a} = [a_1] = [7]$ and the build-up $\mathbf{b} = [b_1] = [3]$ are first computed. The value for M is then evaluated as in Equation (3.11) from the algorithm above. In this example, the IMF index M is 4.

Four speech utterances (equal male and female) were obtained from the TIMIT database and contaminated with car interior noise from the Noisex-92 database under varying SNR conditions. The above procedure for selecting the IMF index *M* was performed on the noisy speech signals and the average value for *M* varies with SNR level as shown in Table 3.2. It can be seen that the value of *M* decreases as the noise level increases. The minimum value of 4 is obtained for *M* at very low SNR levels.

Input SNR (dB)	**Average value of *M***
10	7
8	7
6	7
4	6
2	6
0	5
-2	5
-4	5
-6	5
-8	4
-10	4

Table 3.2: Average value of *M* for speech contaminated with car interior noise under varying SNR conditions

In (Flandrin and Rilling, 2004) and (Wu and Huang, 2004), it was stated that the IMFs of a Gaussian sequence is Gaussian. The denoised signal $x_D[n]$ is formed from a partial reconstruction of the IMFs of the noisy signal and it is Gaussian distributed. Therefore its energy density function must be Chi-squared distributed (Wu and Huang, 2004). From Figure 3.7, $x_D[n]$ is the denoised speech signal with some residual noise. Correspondingly, let $X_D(k,i)$ denote the STFT of $x_D[n]$. The IMCRA noise estimation routine described in Section 2.3.1 is performed on the short-time power spectrum $|X_D(k,i)|^2$. The residual noise power spectrum $\hat{\lambda}_d(k,i)$ is estimated using recursive averaging as follows:

$$\hat{\lambda}_d(k,i+1) = \tilde{\alpha}_d(k,i)\hat{\lambda}_d(k,i) + \left[1-\tilde{\alpha}_d(k,i)\right]\left|X_D(k,i)\right|^2 \tag{3.12}$$

Following Equation (2.14), the noise estimate $\hat{\lambda}_d(k,i)$ in our proposed system is used to perform speech enhancement on the denoised signal by minimizing the optimal LSA estimator as follows:

$$\mathrm{E}_{\min}\left[\left\{\lg S_D(k,i)-\lg \hat{S}_D(k,i)\right\}^2\right] \tag{3.13}$$

where $S_D(k,i)$ is the speech amplitude component that exists in the denoised signal and $\hat{S}_D(k,i)$ is the optimal speech estimate. The a priori SNR $\xi(k,i)$ is estimated using the modified, decision directed approach in (Ephraim and Malah, 1984). The corresponding LSA gain function denoted as $G_{LSA}(k,i)$ to be applied to $X_D(k,i)$ was expressed in Equation (2.23). The OMLSA estimator incorporates speech presence uncertainty to produce the gain function $G(k,i)$ given by:

$$G(k,i)=G_{LSA}(k,i)^{p(k,i)}\,G_{\min}{}^{1-p(k,i)} \tag{3.14}$$

where $p(k,i)$ is the conditional speech presence probability, and the threshold $G_{\min}$ is based on a subjective criteria. The enhanced speech signal is then estimated as follows:

$$\hat{S}_D(k,i)=G(k,i)\left|X_D(k,i)\right| \tag{3.15}$$

where $\hat{S}_D(k,i)$ is the enhanced speech from the proposed system. This speech estimate $\hat{S}_D(k,i)$ will be used to compare the performance of speech enhancement of the EMDF system with that obtained from the full-band OMLSA/IMCRA system.

3.3.2 Performance Evaluation

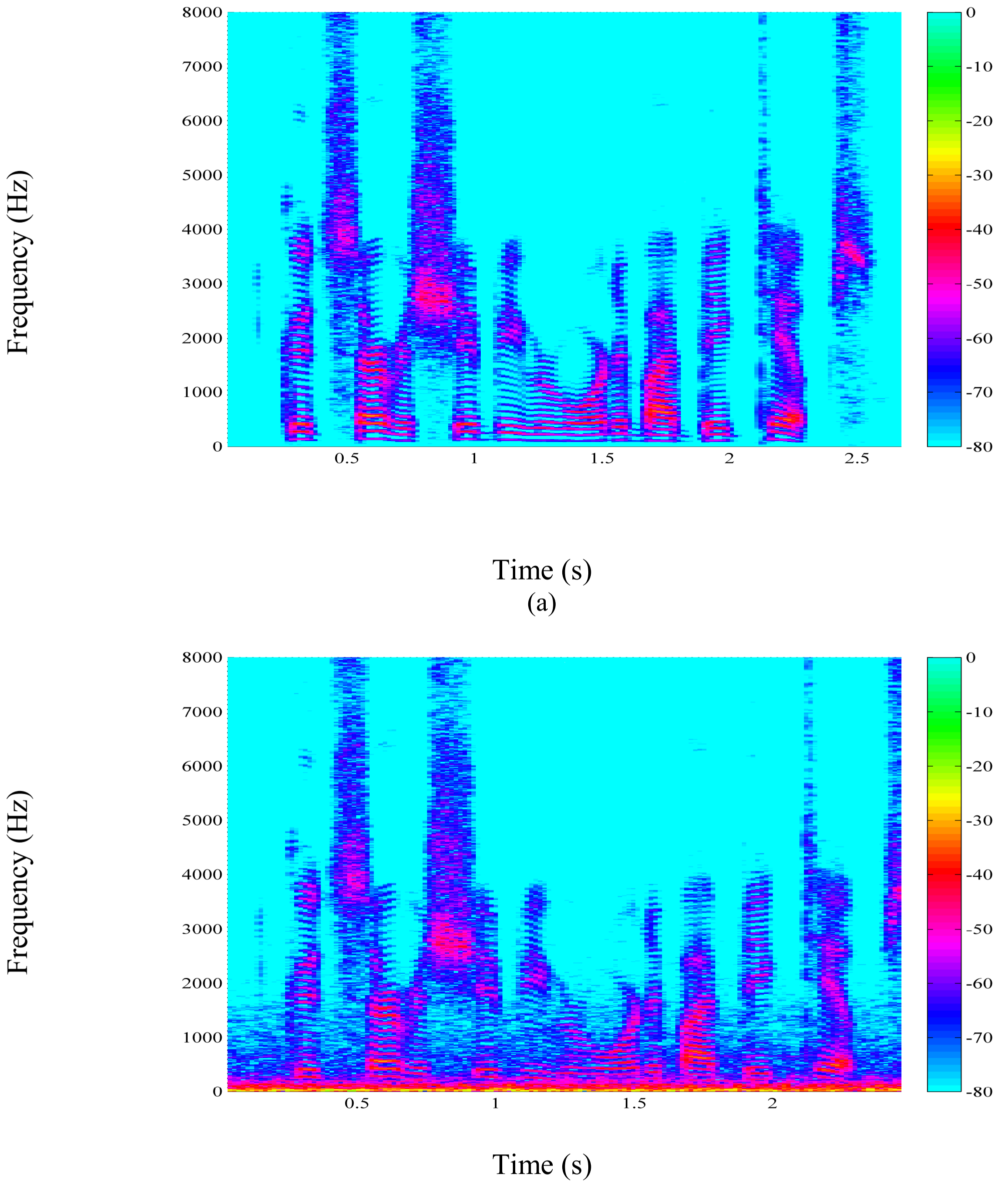
The performance of EMDF technique for speech enhancement was tested on four speech utterances (equal male and female) obtained from the TIMIT database. The clean speech signals were corrupted with the car interior noise used for evaluating IMCRA, as performed in (Cohen, 2003). This non-stationary background noise source was obtained from the Noisex-92 database. The EMDF system's performance was compared with the standard OMLSA/IMCRA algorithm at enhancing the noisy speech signal. A sampling frequency of 16 kHz was used. The EMD-based denoising stage by partial reconstruction using Equation (3.10) is applied to speech blocks of length 40,000 samples. The value for M used in the denoising varies with the SNR level and is selected using Equation (3.11) that was developed in Section 3.3.1. The denoised signal was split up into frames of length 512 samples and a window overlap factor of 50%.

The OMLSA speech estimator gain was used to perform enhancement of the noisy speech output using the noise estimate obtained from the basic IMCRA on the full-band signal, $x[n]$ and the noise estimate from the proposed system on the denoised signal, $x_D[n]$. In order to assess the relative performance of the speech enhancers, the objective measure of segmental SNR improvement for the enhanced speech signals using the EMDF system, when compared to the full-band OMLSA/IMCRA system is given in Table 3.3. These enhancement results were obtained from speech signals contaminated with car interior noise under various SNR levels. An improved quality of speech enhancement obtained by EMDF for speech enhancement is observed. The results show average improvements in segmental SNR of the order 8 dB in this non-stationary adverse noise condition.

Car interior noise environment	
Input SNR (dB)	**Segmental SNR (dB) improvement of EMDF over OMLSA/IMCRA**
10	6.2
8	6.6
6	7.2
4	7.6
2	7.8
0	8.7
-2	9
-4	9
-6	9.3
-8	9.5
-10	9.7

Table 3.3: Segmental SNR improvement using EMDF over OMLSA/IMCRA for speech enhancement under varying SNR conditions

The spectrogram for a clean male speech utterance is given in Figure 3.10(a). This speech signal was contaminated with car interior noise at -10 dB SNR and its spectrogram is shown in Figure 3.10(b). The noisy speech was enhanced using both techniques. The spectrograms for the enhanced speech using the OMLSA/IMCRA and the EMDF system are illustrated in **Figure 3.11**(a) and **Figure 3.11**(b) respectively. These plots demonstrate the improved noise suppression using EMDF. In Figure 3.11(a) and in Figure 3.11(b), the residual noise components during unvoiced speech activity and speech pauses are highlighted with open arrows on the spectrograms for speech enhanced by the OMLSA/IMCRA and the EMDF systems respectively. Comparison of these regions shows that these noise components are significantly attenuated using the EMDF technique. The areas highlighted with solid arrows in Figure 3.11(a) and Figure 3.11(b) show that EMDF retains more of the low frequency voiced speech components. In Figure 3.12(a) and Figure 3.12(b), the frequency axis is scaled to 1 kHz to allow closer examination of low frequencies of the spectrum of the enhanced speech signals. These plots demonstrate the improved noise suppression using EMDF.



(a)

(b)

Figure 3.10: Spectrograms for (a) Original clean speech (b) Speech contaminated by car interior noise at -10 dB

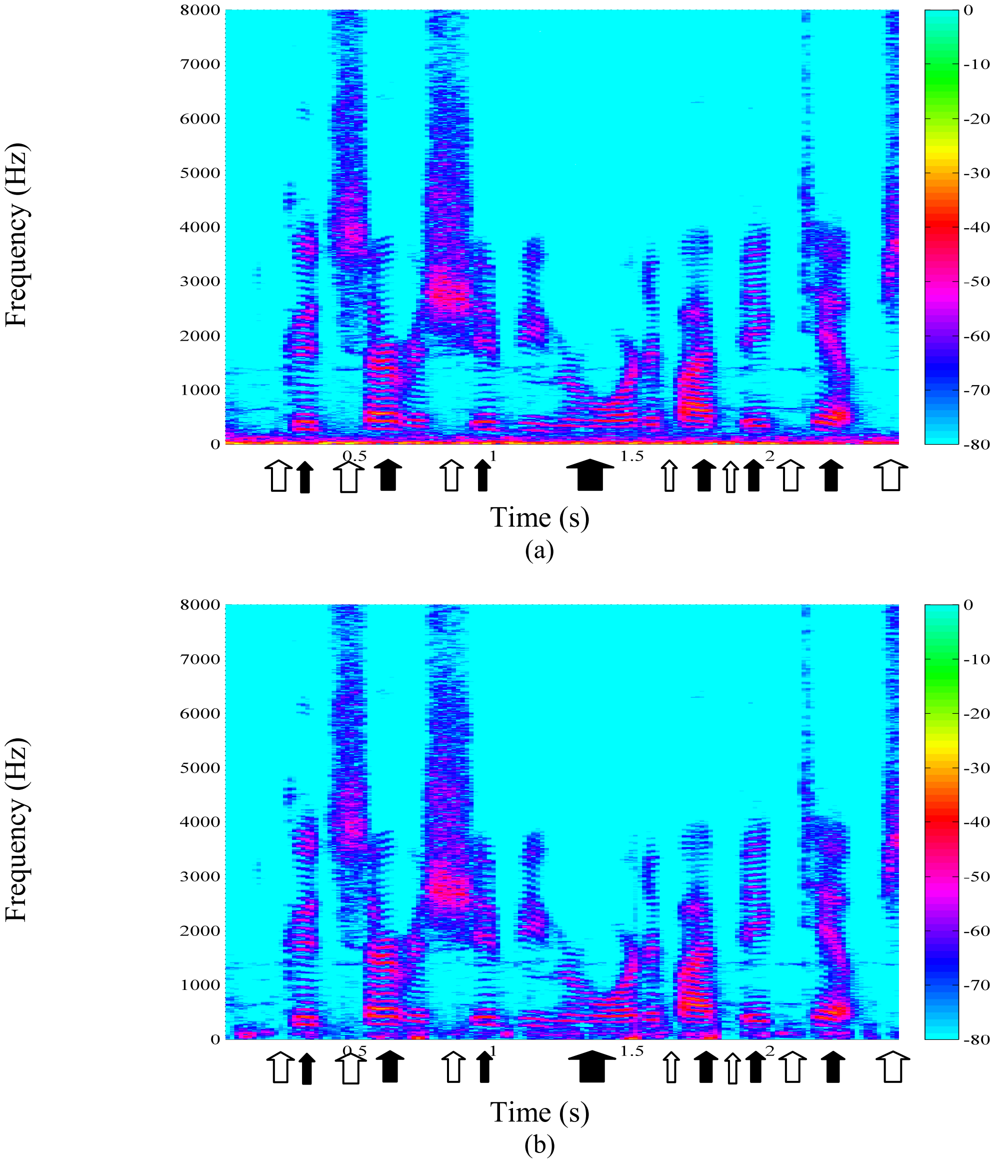
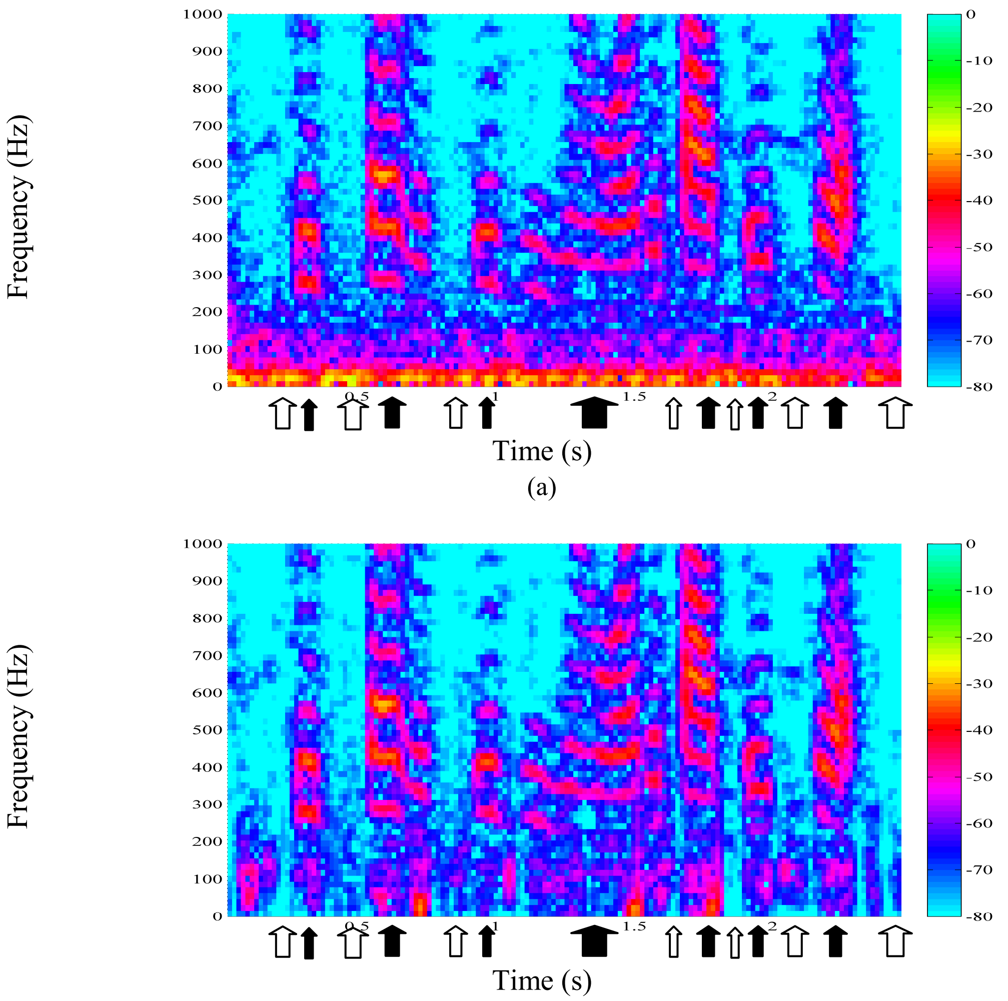


Figure 3.11: Comparison of the spectrograms for speech enhanced by both methods in car interior noise at -10 dB. (a) Speech enhanced by OMLSA/IMCRA (b) Speech enhanced by EMDF



(a)

(b)

Figure 3.12: Comparison of the spectrograms up to 1 kHz for speech enhanced by both methods in car interior noise at -10 dB. (a) Speech enhanced by OMLSA/IMCRA (b) Speech enhanced by EMDF

3.4 Generalized EMD based Filtering for Speech Enhancement

In Section 3.3, the EMDF system for speech enhancement in low frequency noise environments was presented and evaluated under car interior noise conditions. In this section, an updated and generalized EMDF system is proposed for use under other noisy conditions. This new system is shown in Figure 3.13. It can be seen that this new system first performs OMLSA/IMCRA enhancement stage to produce the speech estimate $\hat{s}[n]$. N IMFs are formed from the EMD decomposition of $\hat{s}[n]$. The

EMD based denoising of this speech estimate is then performed to reduce residual low frequency noise components *after* the OMLSA/IMCRA stage.

The method for selecting the IMF index M follows the previous steps shown in Section 3.3.1. However, two updates are made to this algorithm. First, Equation (3.11) is no longer applied, and values for M<4 are accepted. Examination of Table 3.1 shows that this has greater significance for denoising unvoiced speech segments. Secondly, if no peaks are identified, then all IMFs $I_j[n]$ are used in the partial reconstruction (i.e. M=N) of the denoised speech $\hat{s}_D[n]$ in Equation (3.16) below:

$$s_D[n] = \sum_{j=1}^{M} I_j[n] \tag{3.16}$$

This is performed to reduce speech distortion effects. This speech estimate $\hat{s}_D[n]$ will be used to compare the performance of speech enhancement of the EMDF system with that obtained from the full-band OMLSA/IMCRA system.

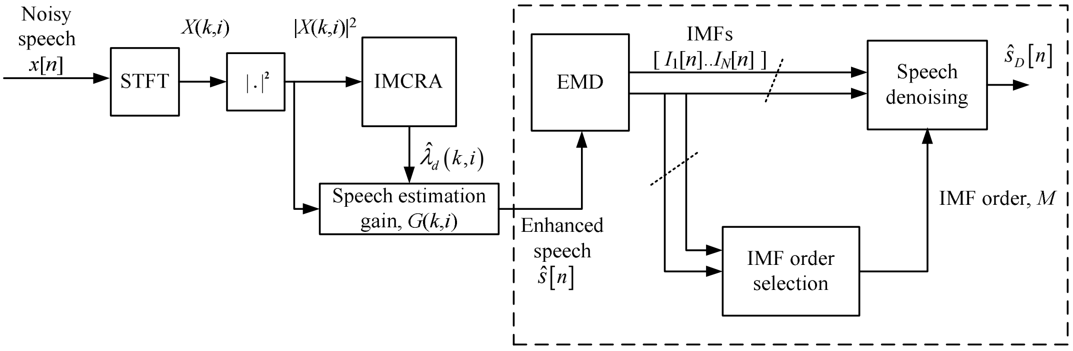


Figure 3.13: Block diagram of the generalized EMDF system for speech enhancement

3.4.1 Performance Evaluation

The performance of the generalized EMDF technique proposed in this section for speech enhancement was tested on 192 speech utterances from 24 different speakers (16 male and 8 female) obtained from the core test set of the TIMIT database. The clean speech signals were corrupted with car interior noise, babble noise and military vehicle noise used for evaluating the speech enhancement systems. These non-stationary background noise sources were obtained from the Noisex-92 database. The generalized EMDF system's performance was compared with the standard OMLSA/IMCRA algorithm at enhancing the noisy speech signals. A sampling frequency of 16 kHz was used. The signal was split up into frames of length 512 samples and a window overlap factor of 50%. The EMD-based denoising stage by partial reconstruction using Equation (3.16) is applied to speech blocks of length 512 samples.

In order to assess the relative performance of the speech enhancers, the objective measures of segmental SNR (segSNR) and Weighted Spectral Slope (WSS) improvements (Section 2.6) for the enhanced speech signals using the generalized EMDF system, when compared to the full-band OMLSA/IMCRA system is given in Table 3.4. It must be noted that negative values for the WSS improvement indicate better enhancement performance and reduced speech loss, whereas positive values for segSNR indicate better performance. These enhancement results were obtained under various SNR levels. An improved quality of speech enhancement is obtained by the generalized EMDF system as the results show improvements in segmental SNR and WSS under all noise conditions. It can be seen that the best overall improvements are obtained under car interior noisy conditions which is dominated by low frequency noise components.

Input SNR (dB)	Car interior noise		Babble noise		Military vehicle noise	
	segSNR (dB)	WSS	segSNR (dB)	WSS	segSNR (dB)	WSS
10	3.6	-17.6	0.3	-7.2	2.3	-21.1
8	4.8	-23.1	0.5	-9.5	2.7	-27
6	5.8	-28.7	0.6	-11.7	3.1	-32.9
4	6.9	-34.6	0.7	-14	3.4	-38.4
2	7.8	-39.9	0.8	-16.7	3.8	-43.6
0	8.5	-45.1	0.9	-19.3	4	-48.3
-2	9.2	-49.5	1	-22.1	4.3	-52.5
-4	9.7	-53.4	1	-24.5	4.5	-56.4
-6	10.1	-56.7	1	-26.6	4.6	-59.6
-8	10.5	-60	1	-28.5	4.7	-62.6
-10	10.7	-62.8	1	-30.4	4.7	-65.5

Table 3.4: Segmental SNR (dB) and WSS improvements obtained when comparing the generalized EMDF system to the full-band OMLSA/IMCRA for various noise types and SNR levels

3.5 Conclusion

In this chapter, novel EMD-based, dual-microphone and single-microphone techniques for enhancing noisy speech signals were presented. In the first part of this chapter, the SEAEMD technique for dual-microphone noisy speech scenarios was presented. When a noise reference is available, the SEAEMD technique can be used to effectively perform noise reduction from the noisy speech signal. The results illustrate that there is considerably more residual noise when ANC is used, compared to SEAEMD. Therefore, for a non-stationary speech signal, the SEAEMD technique proposed in this chapter can be used to provide improved levels of enhancement to the noisy speech, compared to conventional ANC techniques.

In the second part of this chapter, two EMDF techniques for speech enhancement were presented. The basic IMCRA technique is effective at updating the noise spectrum by applying recursive averaging. However, in noise environments with strong low frequency noise environments, IMCRA on the full-band signal does not update the noise power accurately. The first EMDF method for speech enhancement performs a denoising stage first on the IMFs formed from the EMD of the noisy signal. The speech components are then estimated from the denoised signal to give improved results compared to the full-band OMLSA/IMCRA system. The performance of this technique was evaluated using speech contaminated with car interior noise. The second EMDF method performed denoising of the residual low frequency noise components after the OMLSA/IMCRA system. This method was shown to give improved results under various noisy conditions.

4

Local Binary Patterns for 1-D Signal Processing

4.1 Introduction

Local Binary Patterns (LBP) have been extensively used in 2-D image processing (Ojala and Pietikainen, 1999) (He et al., 2009). LBP has been shown in (Ojala et al., 2002) to be a computationally simple, discriminative descriptor of texture. The motivation for the above applications is that an image can be described by a combination of texture patterns. In this chapter, the aim is to develop a 1-D LBP signal processing framework and demonstrate its applicability on a real problem. Real time systems such as hearing aids require fast processing of the input signal while maintaining low computational complexity. One common process in speech systems is Voice Activity Detection (VAD) which attempts to estimate periods of speech and non-speech. VAD decisions are made based on different features of the speech signal and some recent techniques for VAD were discussed in Section 2.7.1. VAD performance is affected by the SNR of the noisy speech and its performance depends on computational complexity and parameter tuning.

This chapter is organized as follows. In section 4.2, a novel 1-D LBP operator is developed and presented as a signal processing tool. A 1-D LBP code for a neighbourhood of sampled data is produced by thresholding the neighbouring samples against centre samples of a processing window. This procedure is iteratively done across the entire signal and a segment of the 1-D signal is alternatively described by a sparser occurrence histogram of LBP codes. This 1-D LBP method is a computationally efficient signal processing tool and in this chapter, the 1-D LBP will be applied to speech systems that have constraints with respect to complexity. In section 4.3, a LBP-based segmentation of a 1-D signal is used to illustrate the processing capability of the 1-D LBP. A computationally simple LBP-based VAD algorithm is designed in section 4.4. This method combines the 1-D LBP occurrence histogram of the underlying signal with a local power measure of the noise source, to identify the voiced, unvoiced and non-speech components. Different types of contaminating non-stationary noise sources must be considered, and therefore the noise statistics are updated, even during speech activity. In section 4.6, the performance of the new VAD technique is demonstrated on speech samples taken from the TIMIT database (John S. Garofolo et al., 1993) and contaminated with non-stationary noises from the Noisex-92 database (Varga and Steeneken, 1993). Finally, concluding remarks are presented in section 4.7.

4.2 1-D Local Binary Patterns

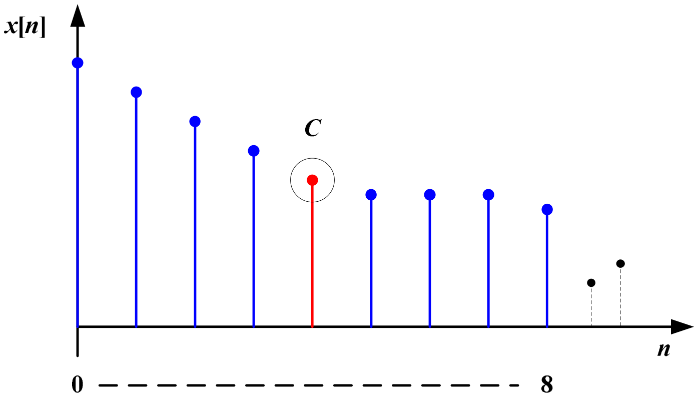
The 1-D LBP operator is adapted from the 2-D LBP (Ojala et al., 2002). It examines a neighbourhood of data samples from a signal $x[n]$ and assigns an LBP code to each centre sample after thresholding them against the neighbouring samples. The 1-D LBP operating on a sample value $x[n]$ is defined as:

$$LBP_P\left(x[n]\right) = \sum_{r=0}^{\frac{P}{2}-1} \left\{ S\left[x\left[n+r-\frac{P}{2} \right] - x[n] \right] 2^r + S\left[x[n+r+1] - x[n] \right] 2^{r+\frac{P}{2}} \right\} \tag{4.1}$$

where the Sign function $S[.]$ is given by:

$$S[x] = \begin{cases} 1 \text{ for x} \geq 0 \\ 0 \text{ for x} < 0 \end{cases} \tag{4.2}$$

and where the P neighbouring samples are thresholded around the centre sample from the neighbourhood of P+1 data samples from the signal $x[n]$ of length N for n=[P/2 : N-P/2]. The Sign function S[.] transforms the differences to a P-bit binary code. The binomial weight applied to each thresholding operation converts the binary code into a unique LBP code.



Thresholding

1	1	1	1		0	0	0	0

X

Binomial weighting

1	2	4	8		16	32	64	128

LBP code = 1+2+4+8 = 15

Figure 4.1: Computation of 1-D local binary pattern (1-D LBP)

An illustration of the 1-D LBP operator is given in Figure 4.1, where P is set to 8 and the centre sample C is circled. As in Eq. (4.1), the 8 neighbouring samples are thresholded against C to produce a binary code of 0000_1111. This code is then multiplied by the binomial weights given to the corresponding samples and the obtained values are summed to give the resulting LBP code of 15. The LBP codes can locally describe the data using the difference between a sample and its neighbours. For a constant or slowly varying signal, these differences cluster near zero. At peaks and troughs, the difference will be relatively large, whereas at edges, the differences in some directions will be larger than those from other directions. The local patterns formed from $x[n]$ can be described by the distribution of the LBP codes:

$$H_b = \sum_{\frac{P}{2}\le n\le N-\frac{P}{2}} \delta\left(LBP_P\left(x\left[n\right]\right),b\right) \tag{4.3}$$

where b=1..B and B is the number of histogram bins and each bin corresponds to an LBP code. $\delta(i,j)$ is the Kronecker delta function.

The 1-D LBP operator has been developed and it was shown that it extracts LBP codes from the analyzed signal. These codes can be used as features and applied to signal processing problems, as shown in the following sections. The procedure for extracting a LBP code for a neighbourhood of samples can be summarized and reiterated in Figure 4.2. The procedure iterates over all signal segments and can be used to describe a signal segment by a sparser LBP occurrence histogram.

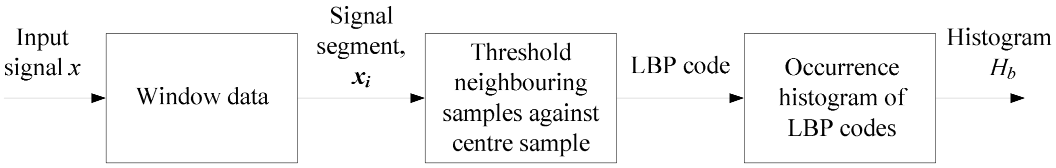

Input signal x
Window data
Signal segment, x_i
Threshold neighbouring samples against centre sample
LBP code
Occurrence histogram of LBP codes
Histogram H_b

Figure 4.2: Overview of 1-D LBP procedure on a neighbourhood of samples to extract a LBP code

The standard LBP_P operator produces 2^P different LBP codes. Extensions of the LBP_P are presented in (Ojala et al., 2002) for rotation invariant patterns LBP_P^r, uniform patterns LBP_P^u, and rotation invariant uniform patterns $LBP_P^{r,u}$. LBP_P^r is produced by circularly shifting the LBP code for the P neighbouring samples until its minimum value is found. In this way, LBP_P^r of the processed window produces the same code for all shifted versions of that code and it is therefore invariant to rotation. This transformation is given by:

$$LBP_P^r\left(x\left[n\right]\right) = \min\left(ROR\left(LBP_P\left(x\left[n\right]\right),i\right)\right) \text{ for } i = 0,1,..P\text{-}1 \tag{4.4}$$

where the function $ROR\left(LBP_P\left(x\left[n\right]\right),i\right)$ circularly shifts the P-bit binary code i times to the right and i<P.

A uniform pattern is defined by an LBP code which has at most two one-to-zero or zero-to-one transitions. A non-uniform pattern is defined by an LBP code which has more than two one-to-zero or zero-to-one transitions. Each uniform pattern is assigned to a separate bin and the remaining non-uniform patterns are assigned to a single histogram bin. The binary codes 1111_1111 (zero transitions) and 0000_0100 (two transitions) are examples of uniform codes. The binary code 0000_0101 (four transitions) is an example of a non-uniform code. LBP_P^{u} gives a histogram with $P(P-1)+3$ bins.

The rotation invariant, uniform $LBP_P^{r,u}$ shifts the uniform codes until they attain their minimum values and results in a histogram with $P+1$ bins for uniform patterns plus one bin for non-uniform patterns. In the previous paragraph, the binary code 0000_0100 was given as an example of a uniform binary code. Using Eq. (4.4), the rotation invariant, uniform code for this example is 0000_0001. The binary code of 0000_1111 which was evaluated earlier in the illustration in Figure 4.1 is an example of a LBP code that is uniform and is already rotation-invariant. The choice of which LBP to use depends on the need for either a more resolved representation or for a sparser histogram. In the work presented in this chapter, normalized histograms will be used for $LBP_P^{r,u}$ resulting in histograms with $P+2$ bins.

4.3 Unsupervised Signal Segmentation using 1-D LBP

The 1-D LBP operator is used to produce a histogram of LBP codes which can be used as an alternative representation of the signal. In signal segmentation, the histogram can be used as a non-parametric estimator of the empirical LBP feature histogram. Resistor Average Difference (RAD) (He et al., 2009) can be used for measuring the similarity of adjacent LBP histograms. RAD is derived from the non-symmetric Kullback-Leibler Distance (KLD) (He et al., 2009) which is used for measuring the difference between two histograms p and q. KLD is given by:

$$D_{KL}\left(p\|q\right)=\sum_{b=1}^{B}p(b)\left\{\lg\left(p(b)\right)-\lg\left(q(b)\right)\right\} \tag{4.5}$$

where B is the number of histogram bins and $p(b)$ and $q(b)$ are the number of occurrences in histograms p and q respectively at bin b. The RAD is defined as:

$$D_{RAD}\left(p,q\right)=\left[\left(D_{KL}\left(p\|q\right)\right)^{-1}+\left(D_{KL}\left(q\|p\right)\right)^{-1}\right]^{-1} \qquad \textbf{(4.6)}$$

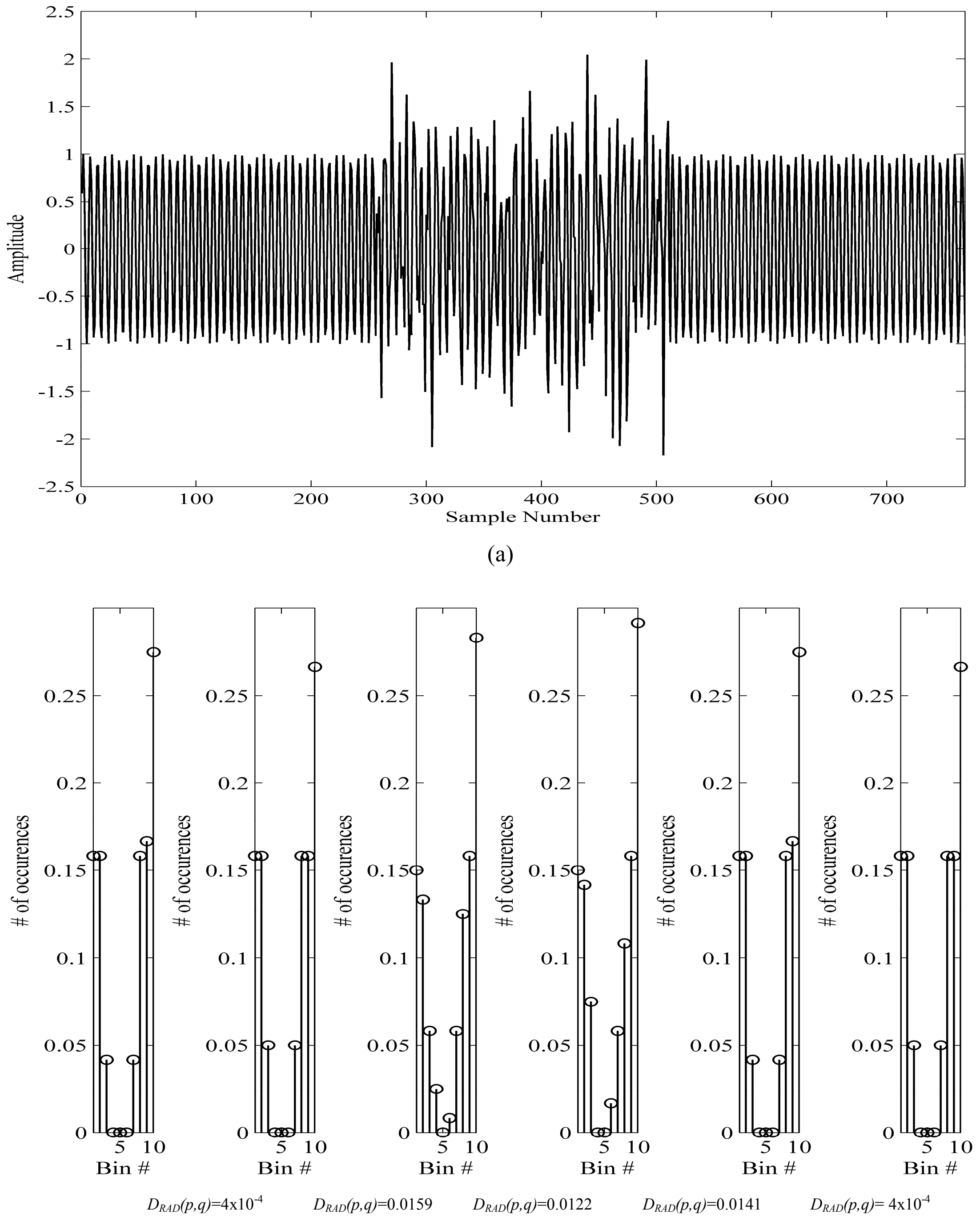
$D_{RAD}(p,q)$ between the two histograms p and q increases with dissimilarity and in contrast to KLD, RAD is symmetric (Johnson and Sinanovic, 2001).

4.3.1 Noise Onset Identification

In this example, the onset of noise is detected for a noise source switched on at some time τ. The signal $x[n]$ is first split into segments $x_i[j]$ of length W by applying a window $w[j]$ of length W as:

$$x_i\left[j\right]=x\left[iR+j\right]w\left[j\right] \text{ for } 0\leq j\leq W-1 \qquad \textbf{(4.7)}$$

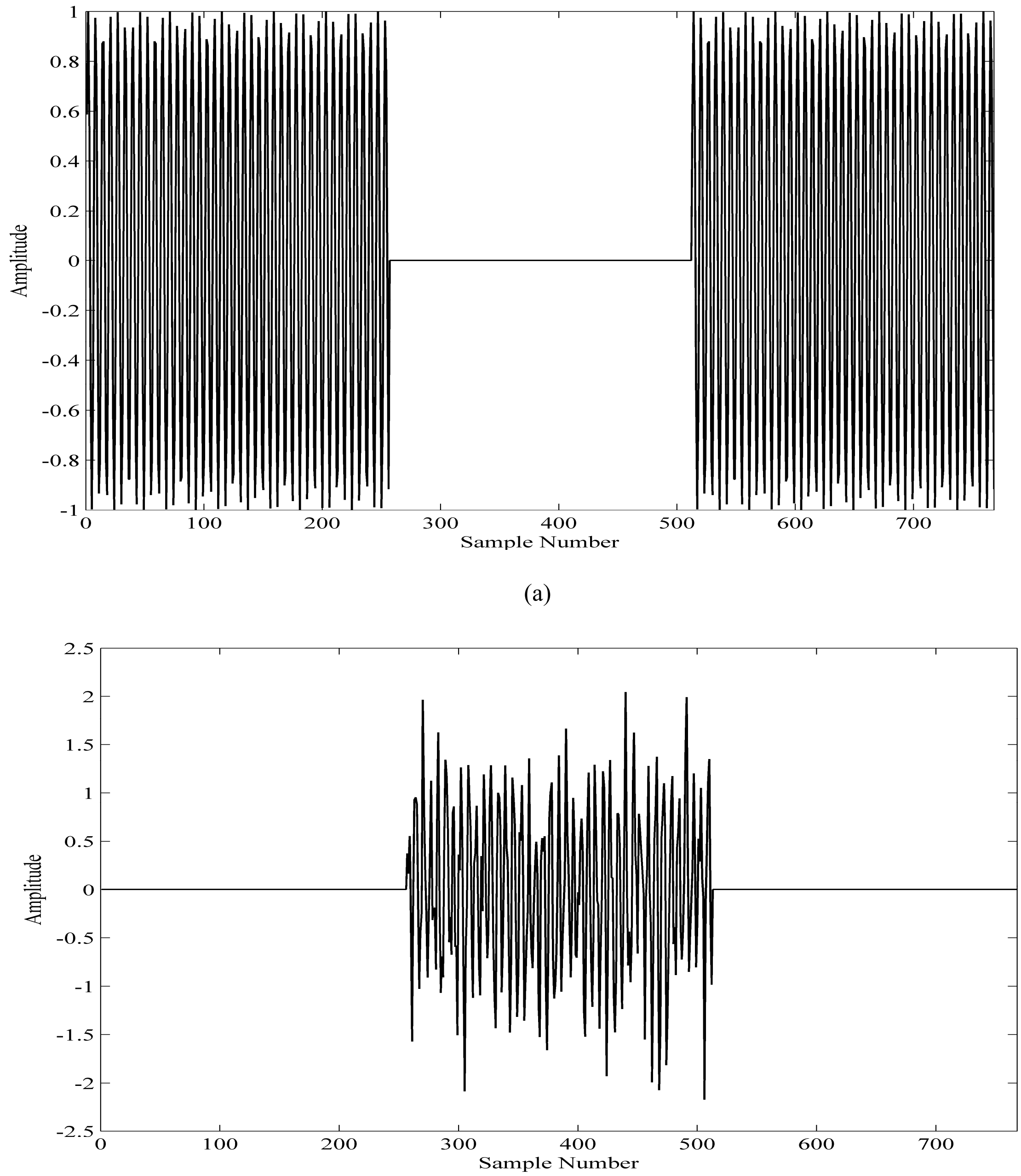
where i is the segment number, $R<W$ for overlapping segments and $R=W$ for contiguous segments. W is chosen to be small enough to capture transitions in the LBP feature histograms. $D_{RAD}(p,q)$ is measured for the segments of the adjacent histograms and similar segments are merged. When two adjacent segments are merged, their histograms are summed and normalized to produce the histogram of the new segment. This procedure continues until the segment does not expand and the previously merged segments are considered as a component of the signal with similar underlying LBP features.



(a)

(b)

Figure 4.3: Segmentation of a sinusoidal signal contaminated by AWGN (a) Original noisy signal (b) $LBP_8^{r,u}$ histograms of the 6 segments formed and $D_{RAD}(p,q)$ measure for adjacent histograms



(a)

(b)

Figure 4.4: Segmentation of a sinusoidal signal contaminated by AWGN (a) Segmented sinusoidal components (b) Noise affected segment

This procedure was performed for an artificially generated sinusoidal signal of length 768 samples which was contaminated by Additive White Gaussian Noise (AWGN) in the middle portion of the signal as shown in Figure 4.3(a). The signal was split up as in Eq. (4.7) with W=128 and R=128 and a rectangular window $w[j]$. The 1-D $LBP_8^{r,u}$ extension was used with P=8 to give a LBP histogram with P+2=10 bins as shown in Figure 4.3(b) for each segment. The D_{RAD} values for adjacent segments are shown for illustrative purposes. The results of the segmentation are shown in Figure 4.4(a) and Figure 4.4(b). It can be seen that the algorithm exactly separates the sinusoidal components from the noise affected portion based on the similarity of the underlying signal features. No overlap was used in this example, however, overlapping the segments will improve fidelity.

4.4 Voice Activity Detection using 1-D LBP

Traditional VAD detects speech activity in the presence of noise. VAD techniques do not usually distinguish between voiced and unvoiced components. Unvoiced speech contains high occurrences of non-uniform patterns and use of the uniform LBP extension, $LBP_P^{r,u}$, can distinguish between these two speech components. As an example, the clean speech utterance "*Good service should be rewarded by big tips*" was taken from the TIMIT database and is plotted in Figure 4.5(a). A sampling frequency of 16 kHz was used and the signal was segmented according to Eq. (4.7) with a rectangular window of length W=160 samples and no overlap. The $LBP_8^{r,u}$ for each segment was measured to give LBP histograms with 10 bins. Any non-uniform patterns are separated into a single bin. Figure 4.5(b) shows the plot for the non-uniform bin (bin 10) for each speech segment. This illustrates that the higher frequency unvoiced speech circled in Figure 4.5(a) and labelled "U" produce higher occurrences of non-uniform patterns. Non-uniform patterns occur in other portions of the signal. This is due to low-power recording noise from the speech sample used. This distinctive non-uniform marker can be used to identify unvoiced speech segments of the analyzed signal that have an increased number of occurrences in the non-uniform histogram bin.

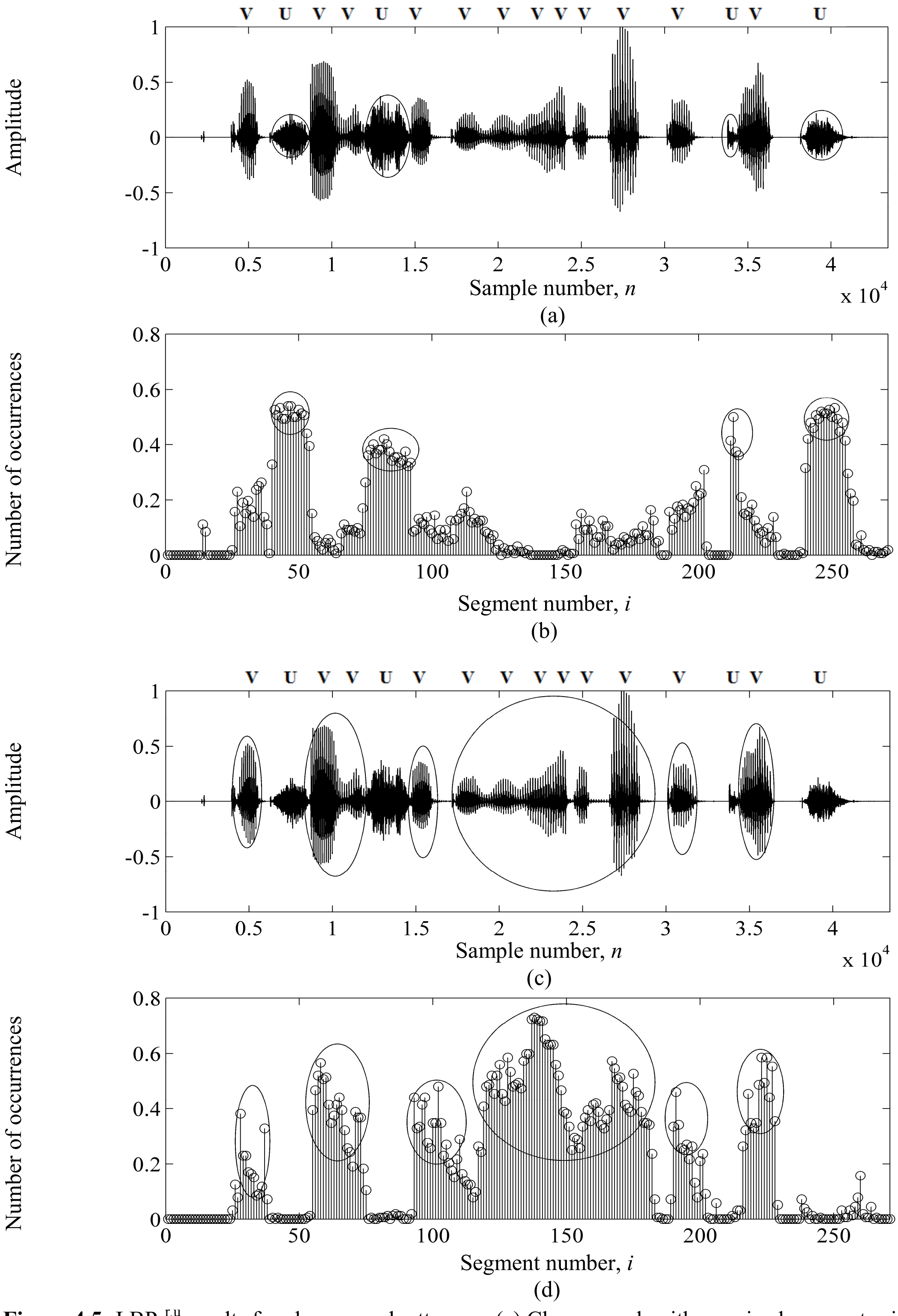


Figure 4.5: LBP$_8^{r,u}$ results for clean speech utterance (a) Clean speech with unvoiced segments circled (b) Occurrence results in non-uniform bin 10 from LBP feature histograms (c) Clean speech with voiced segments circled (d) Occurrence results in central uniform bin 5 from LBP feature histograms

The lower frequency voiced components are highlighted in circles and labelled "V" in Figure 4.5(c). These produce uniform patterns with the resulting plot shown in Figure 4.5(d). This shows the number of occurrences in the central uniform bin 5 for the segmented signal. The distribution of the patterns for speech signal shows peak activity in the uniform bin 5 at segments corresponding to voiced speech. This LBP feature relates to a particular rotation-invariant feature of the voiced components. It can be seen that during voiced speech activity there is significant activity in this central bin. Therefore, the occurrence histograms of these speech components can distinguish these two regions based on their extracted LBP features. Noise may contain non-uniform patterns and for noisy speech signals, the bin 5 features can also distinguish unvoiced speech components from weaker voiced speech components that have been more affected by the added noise. A higher resolved histogram such as LBP_P^u can be used if this criterion to distinguish unvoiced speech from noise or weak speech components affected by noise is required. LBP_P^u distributes the occurrences in the histogram over a larger number of bins and thus keeps activity low in any particular uniform bin for unvoiced speech.

Environmental sounds may contain low-frequency noise and periodic components whose spectra overlap with the voiced components of the speech signal. Therefore, discrimination of features that produce similar histograms from different sound sources is performed by incorporating a local power measure of the analyzed signal segment $x_i[j]$ to give the joint operator $LBP_P^{r,u}/VAR_{seg}$ where $VAR_{seg}(x_i[j])$ is given by:

$$VAR_{seg}\left(x_i\left[j\right]\right)=\frac{1}{W}\sum_{j=0}^{W-1}(x_i\left[j\right]-\overline{X}_i)^2 \text{ where } \overline{X}_i=\frac{1}{W}\sum_{j=0}^{W-1}(x_i\left[j\right]) \quad \textbf{(4.8)}$$

4.4.1 Algorithm

Consider the model described by:

$$x[n]=s[n]+d[n] \quad \textbf{(4.9)}$$

where $x[n]$ is the noisy speech signal, $s[n]$ is the original noise-free speech, and $d[n]$ is the noise source which is assumed to be independent of the speech.

The algorithm presented below uses the 1-D LBP to separate noisy speech into voiced, unvoiced and non-speech components by the following steps:

1. Segment the input noisy speech signal $x[n]$ from Eq. (4.9) to give segments $x_i[j]$ using Eq. (4.7)
2. Perform $LBP_8^{r,u}$ for each segment $x_i[j]$ to obtain the normalized occurrence histogram for that segment
3. Separate all segments which have the normalized histogram bin $p(10)>NUthresh$ and label as unvoiced speech segments. $p(b)$ is the occurrence probability in histogram bin b and is given by:

$$p(b)=\frac{\text{Number of occurrences in histogram bin } b}{\text{Total number of occurrences in all bins}} \tag{4.10}$$

4. Measure $VAR_{seg}(x_i[j])$ for each segment and separate the LBP features with $VAR_{seg}(x_i[j])<thresh$. Label as non-voiced speech segments
5. Label remaining segments as voiced speech segments
6. Perform final grouping by assigning contiguous speech segments $T_{NV} < 30$ ms to non-voiced speech label

The value of *thresh* must be chosen to distinguish voiced speech from non-speech with similar LBP features. Values of 0.3 for *NUthresh* and 0.003 for *thresh* were selected empirically from experimental studies for speech contaminated with different noise types ranging down to 0dB SNR. This threshold value for *NUthresh* corresponds to the analysis of Figure 4.5(b) which shows the increased non-uniform bin activity corresponding to unvoiced speech components. The value of T_{NV} was chosen following (Hu and Wang, 2004) to remove the influence of noise intrusion.

4.5 Improved Voice Activity Detection using 1-D LBP

The 1-D LBP is able to distinguish the unvoiced and the voiced components of clean speech signals using the distinguishing features of higher activity in certain characteristic histogram bins. In Section 4.4, the activity in the non-uniform bin of the extracted histogram H_b of LBP codes was used as the distinguishing feature of unvoiced speech components in the presence of contaminating noise. Unvoiced components have relatively low energy and are usually affected by noise to a greater extent than higher energy voiced components. Therefore, this distinctive feature for unvoiced components improves fidelity by reducing misclassifications compared to the techniques presented in Section 2.7.1.1. A local variance measure of the analyzed segments was compared against a fixed threshold and used to separate the speech pauses. In the presence of non-stationary noises, this process can be improved by including a joint local variance measure as an estimate of the noise variance, which is updated during speech pauses as done in some of the energy-based VAD techniques discussed in Section 2.7.1.1.

In (Loizou, 2007a), a "silent" segment occurs when the spectral energy goes to zero or near the noise floor. Therefore, "silent" segments occur during speech activity during unvoiced speech at low frequencies (usually below 2 kHz) and during voiced speech at high frequencies (usually above 4 kHz). As a result, when speech is contaminated with low frequency noises such as car-interior noise, the unvoiced speech components are less affected than the voiced speech components. Correspondingly, when speech is contaminated with high frequency noise such as F16 cockpit noise, the voiced components are less affected than the unvoiced components of the speech. These observations were incorporated in the novel improved VAD based on 1-D LBP presented in this section. In this technique, the unvoiced speech components are first classified and a VAD decision rule is defined based on the ratio of the variance of the noisy speech to the estimate of the noise variance. This decision rule is used to classify the remaining voiced speech components and the speech pauses. The novel feature of this decision rule is that it is performed in the low frequency band below 2 kHz. Therefore, the noise variance estimate can be updated during the classified speech pauses, as well as during the detected unvoiced speech components which can be considered to be a "silent" segment in the evaluated frequency band. As a result, there will be more frequent updates to the noise variance estimate compared to techniques presented in Section 2.7.1.1, due to the additional variance tracking during speech activity.

The improved VAD using 1-D LBP is illustrated in Figure 4.6 and the algorithm is detailed in the following section.

4.5.1 Algorithm

This algorithm combines the discriminating features from the 1-D LBP and a locally updated estimate of the noise variance. The noisy speech signal $x[n]$ from Eq. (4.9) is segmented using Eq. (4.7) to give the noisy speech segments $x_i[j]$ in Figure 4.6. Each segment is then high-pass filtered with a lower cut-off frequency of 150 Hz to remove strong low-frequency noise components at low SNR values. As a result of this initial filtering, the LBP feature in the non-uniform bin will be less affected and also the noise variance estimate will be more reliable due to reduced fluctuations. The 1-D LBP operator, $LBP_8^{r,u}$ is applied to the high-pass filtered signal $x_{i,HP}[j]$ to obtain the normalized occurrence histogram H_b for each segment. Following section 4.4.1, all segments which have the normalized histogram bin $p(10)>NUthresh$ are separated and labelled as unvoiced speech segments. The remaining segments are evaluated using the decision rule defined in:

$$VAD(i) = \begin{array}{ll} 0 & \text{if } \dfrac{\lambda_{\tilde{x}}(i)}{\hat{\lambda}_d(i-1)} \leq \eta \\ 1 & \text{otherwise} \end{array} \quad \textbf{(4.11)}$$

where the value 1 denotes the presence of a voiced speech component and the value 0 denotes the presence a speech pause. $\lambda_{\tilde{x}}(i)$ is the variance of the low-frequency band-pass signal segment $x_{i,BP}[j]$ obtained using:

$$\lambda_{\tilde{x}}(i) = VAR_{seg}\left(x_{i,BP}[j]\right) \quad \textbf{(4.12)}$$

$\hat{\lambda}_d(i)$ is the estimate of the noise variance in the band-pass signal segment $x_{i,BP}[j]$ and is initialized during an assumed 120 ms phase of noise only at the beginning of the signal $x[n]$. If a voiced speech component is detected using Eq. (4.11), the noise variance estimate is not updated. However, as shown in Figure 4.6, during unvoiced speech components and speech pauses, the noise variance estimate $\hat{\lambda}_d(i)$ is updated using first order recursive averaging according to:

$$\hat{\lambda}_d(i) = \alpha_d \hat{\lambda}_d(i-1) + [1-\alpha_d]\lambda_{\tilde{x}}(i) \quad \textbf{(4.13)}$$

where the smoothing factor α_d ranges from [0, 1]. The compromise for the value of α_d relies on the stationarity of the contaminating noise source and a typical value of $\alpha_d = 0.96$ was used for this system.

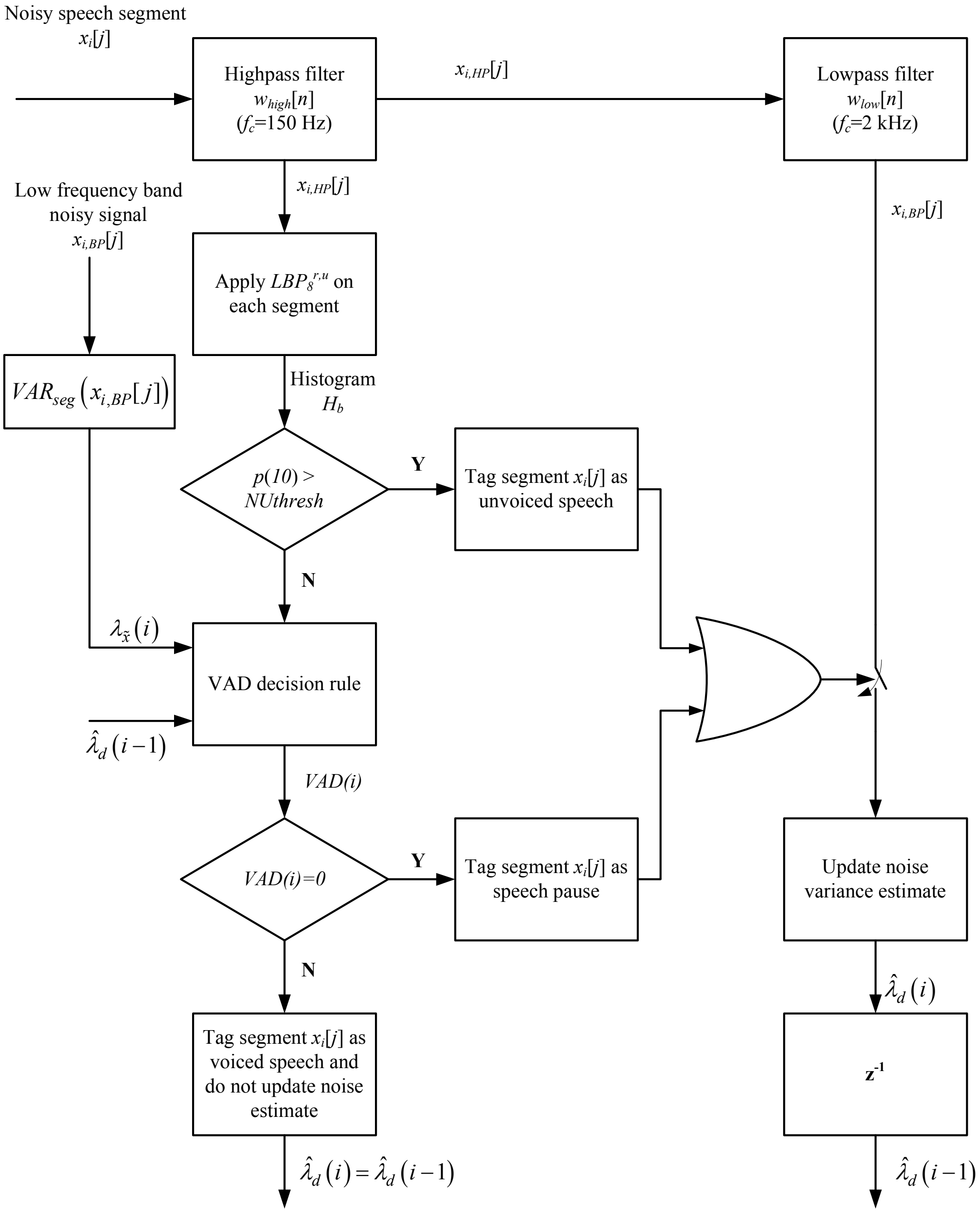


Figure 4.6: Flowchart of Improved Voice Activity Detection using 1-D LBP and a noise variance estimate

Following Section 4.4.1, grouping of the unvoiced speech components was performed by assigning contiguous speech segments T_{NV}< 30 ms to the speech pause label and *NUthresh* was set to 0.3 as before. This novel approach for VAD is flexible to change in parameter values which allow the modification of the system performance depending on the tolerance to varying the speech presence and the speech pause false alarm rates.

The ability of the proposed VAD algorithm at tracking the noise variance is demonstrated in Figure 4.7 by comparing the actual noise variance with the estimate $\hat{\lambda}_d(i)$ obtained. The 15 second test speech sequence used in Section 2.7.2 was used for this example, in F16 cockpit noise conditions at 5 dB SNR. The proposed VAD technique performs good tracking of the noise variance and is able to track the rising noise power even under this noisy condition.

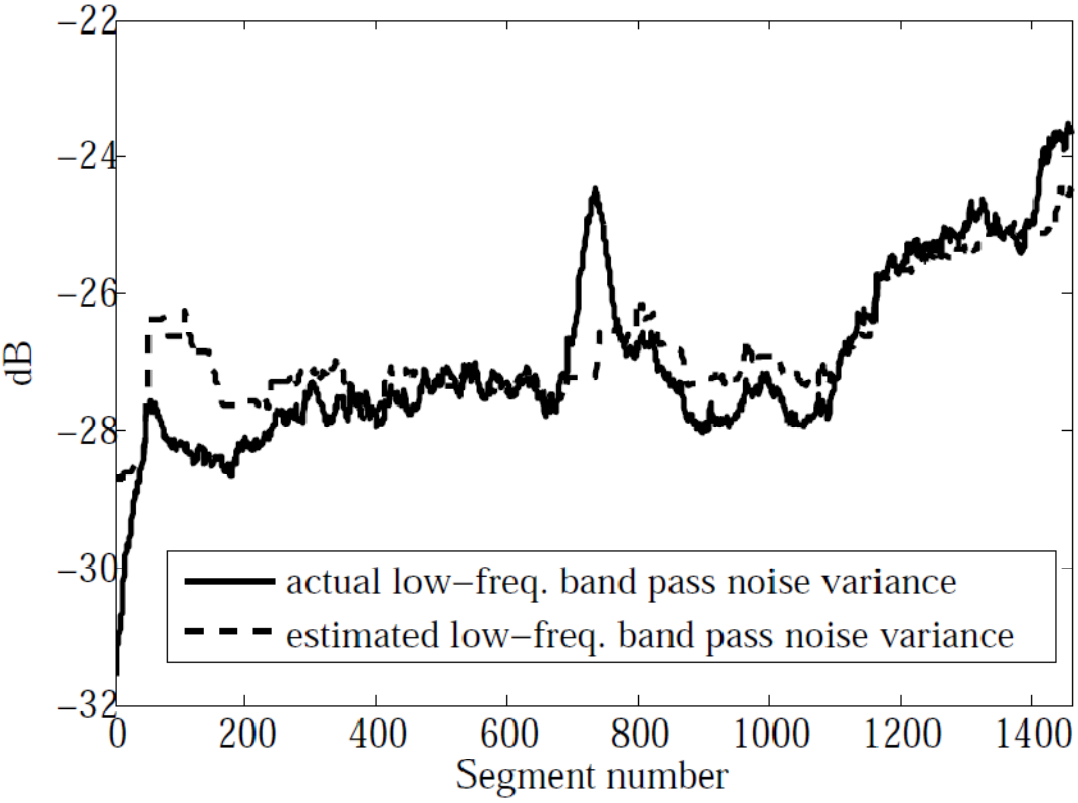


Figure 4.7: Estimation of the noise variance using the proposed VAD technique for speech contaminated with F16 cockpit noise at 5 dB SNR

4.5.2 VAD Decision Smoothing

VAD decision smoothing is performed at the end of the initial classification from the previous section to reduce misclassification errors. In this section, a novel VAD decision smoothing scheme is proposed. Phoneme duration analysis from (Vlaj et al., 2009) showed that the majority of vowels, diphtongs and semivowels have a duration of 80 ms. The majority of consonants were shown to have a duration of 110 ms.

Therefore, our decision smoothing scheme is based on the generalized assumption that the detected speech must have a minimum duration of 110 ms. The length of the detected speech burst is evaluated as τ_{bc}, and if its length is less than a predefined length τ_{hb}, $\tau_{hb} - \tau_{bc}$ frames are prefixed to the speech to satisfy this hangbefore criterion by modifying their VAD decision to speech presence classification. A constant hangover of τ_{ho} is adopted at the end of every speech burst by modifying the VAD decision for the subsequent frames to speech presence classification. In our work, the parameters were set with $\tau_{hb} = 80\text{ ms}$ and $\tau_{ho} = 30\text{ ms}$. This decision smoothing scheme is illustrated with an example of a detected speech burst of duration $< \tau_{hb}$ in Figure 4.8 below.

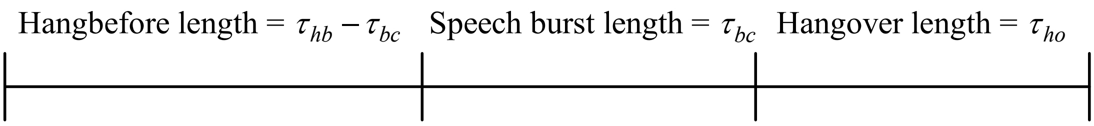
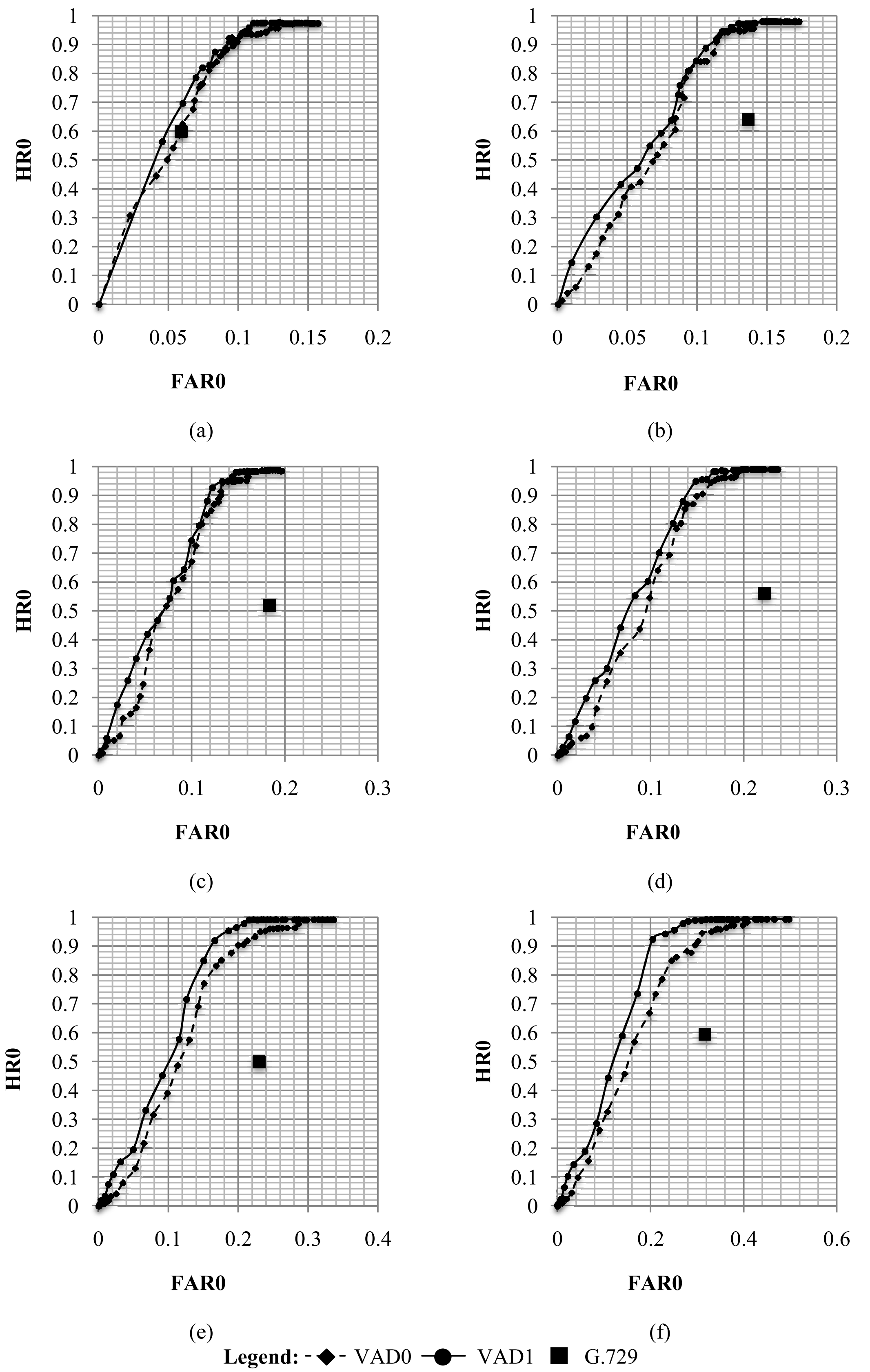

Hangbefore length = $\tau_{hb} - \tau_{bc}$
Speech burst length = τ_{bc}
Hangover length = τ_{ho}

Figure 4.8: Example of VAD decision smoothing scheme for a detected speech burst shorter than τ_{hb}

4.6 VAD Performance Evaluation

The improved VAD using 1-D LBP was tested on speech sequences which were obtained by artificially concatenating the sentences from three different speakers in the core test set from the TIMIT database (subset dialect region 7). The total length of the test set used was 84 seconds and the speech was mixed with car interior noise, F16 cockpit noise and babble noise obtained from the Noisex-92 database (Varga and Steeneken, 1993). The performance was evaluated at a range of SNRs from 20 dB to very noisy conditions at -10 dB. As discussed in Section 2.7.2, the speech pauses for the reference were labelled by using the G.729B VAD on the clean speech sequences.

The VAD algorithm presented in Section 4.5 will be referred to as “VAD0”. For this performance evaluation, an additional update is made to VAD0, by grouping the voiced speech components and assigning contiguous speech segments $T_V < 30$ ms to the speech pause label to reduce the effect of noise intrusions. This modified technique will be referred to as "VAD1". In this section, these novel techniques will be compared against the G.729B VAD which will be referred to as "G.729". VAD1 and VAD0 were performed on the noisy speech sequences with segments of length W=10 ms at a sampling frequency of 16 kHz. The telecommunication standard G.729 was performed on the noisy speech sequences with W=10 ms at a sampling frequency of 8 kHz. The speech pause hit rates $HR0$ from Equation (2.31) and the speech pause false alarm rate $FAR0$ from Equation (2.33) were evaluated for each noise condition. The performance of the VAD algorithm can be adjusted to the desired operating point by changing the threshold criteria in the algorithm's decision rule. The three methods were compared using the Receiver Operating Characteristics (ROC) by varying the VAD threshold η used in the decision from Eq. (4.11), while keeping the more robust decision threshold *NUthresh* fixed. The ROC are shown in Figure 4.9 using car interior noise, in Figure 4.10 using F16 cockpit noise and in Figure 4.11 using babble noise, with the legends shown at the bottom of the figures.

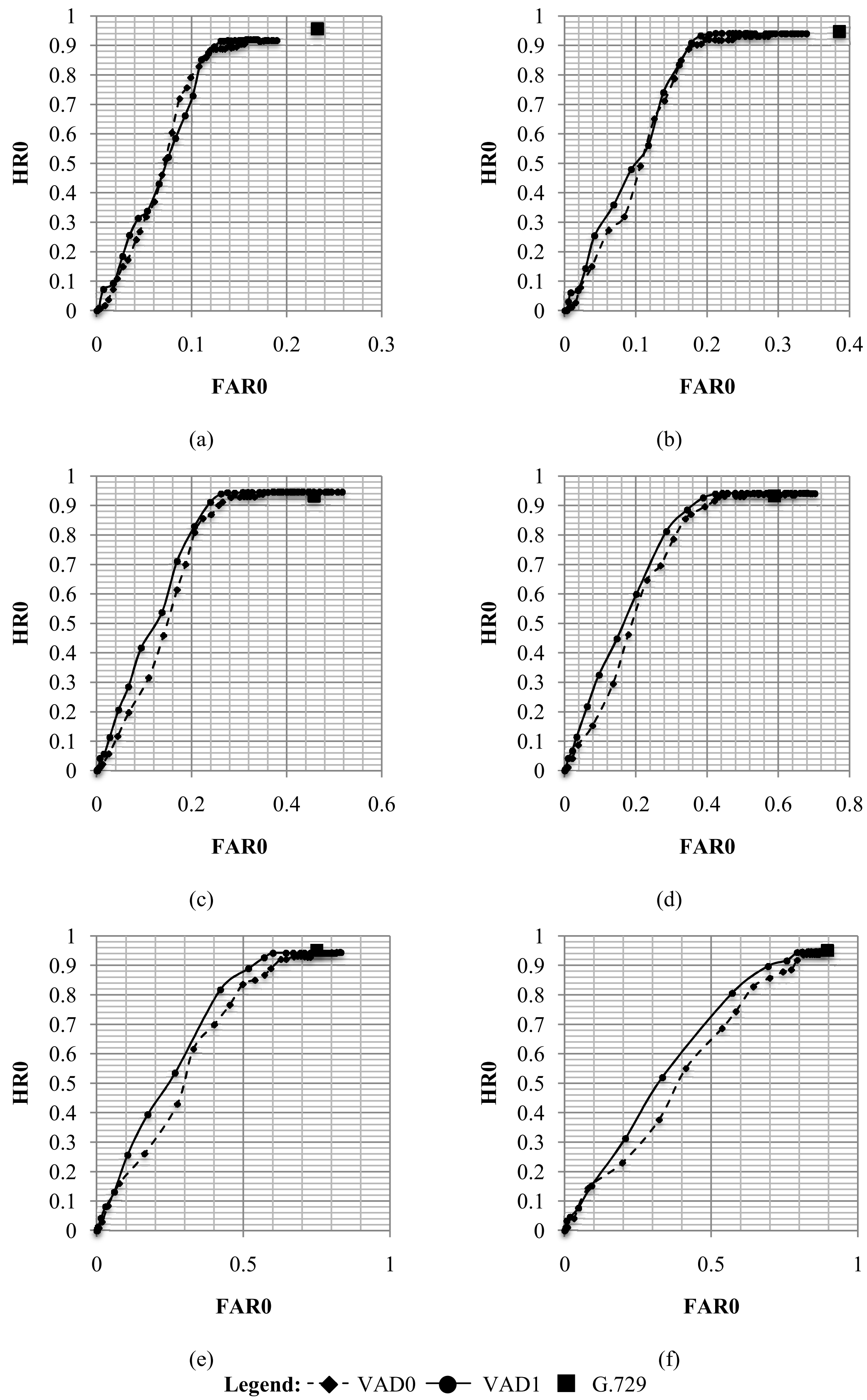


(a) (b)

(c) (d)

(e) (f)

Figure 4.9: ROC curves comparing the proposed VAD techniques against G.729 using car interior noise at different SNR values of (a) 20 dB (b) 10 dB (c) 5 dB (d) 0 dB (e) -5 dB (a) -10 dB



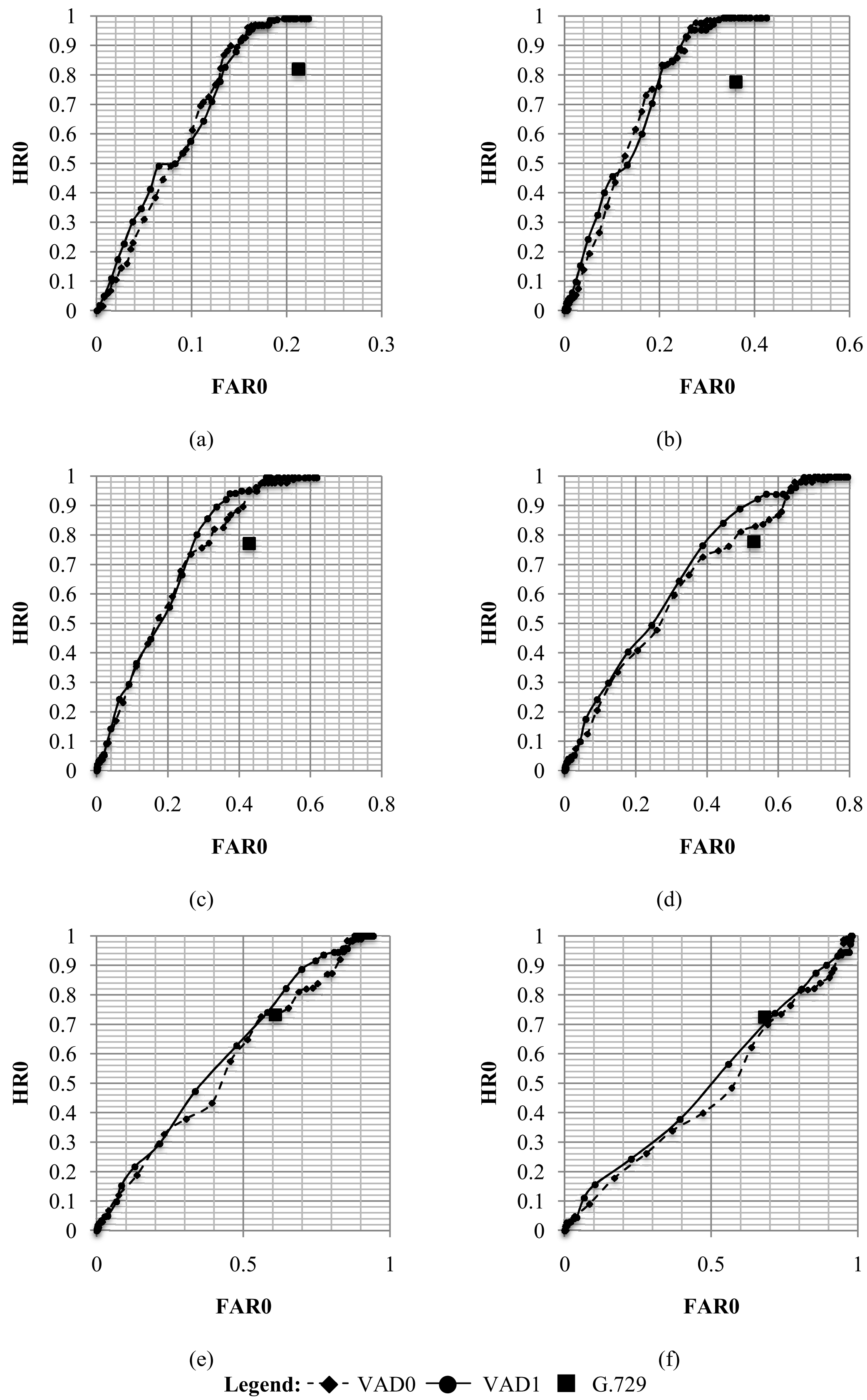
(a) (b)

(c) (d)

(e) (f)

Legend: -◆- VAD0 -●- VAD1 ■ G.729

Figure 4.10: ROC curves comparing the proposed VAD techniques against G.729 using F16 cockpit noise at different SNR values of (a) 20 dB (b) 10 dB (c) 5 dB (d) 0 dB (e) -5 dB (a) -10 dB



(a) (b)

(c) (d)

(e) (f)

Figure 4.11: ROC curves comparing the proposed VAD techniques against G.729 using babble noise at different SNR values of (a) 20 dB (b) 10 dB (c) 5 dB (d) 0 dB (e) -5 dB (a) -10 dB

Although the G.729 VAD was used on the clean speech to obtain the references for the speech pauses, the above ROC curves for the three VAD techniques in Figure 4.9, Figure 4.10 and Figure 4.11 show that the proposed algorithms VAD0 and VAD1 outperform G.729 over a range of SNRs and noise types. For a given speech hit rate *HR*0, the G.729 VAD on average has a higher false alarm rate for speech pause detection. It can also be seen that the modified approach VAD1 performs slightly better than VAD0 due to the rejection of shorter segments which are most likely present due noise intrusions. The VAD decision threshold η can be varied to decrease the *FAR*0 at the expense of reduced *HR*0. Therefore, depending on the application, a compromise must be made on the acceptable levels for misclassifying speech segments as speech pauses, in order to increase the amount of speech pauses detected. As expected, the best overall performance is obtained in the low-frequency noise car environment as shown in Figure 4.9. This is due to the improved tracking of the noise variance over time. The cockpit noise and babble noise environments are highly non-stationary and therefore, under these conditions, the noise level changes more rapidly. Also, speech reception is highly affected as the SNR decreases in these two noisy conditions which justify the significant decrease in performance at low SNRs below 0 dB. This effect is more prominent in babble noise scenarios as seen in Figure 4.11, where the speech intelligibility is known to decrease significantly with higher noise levels and the inability to discern speech at -10 dB SNR.

4.7 Conclusion

The histogram of the 1-D LBP codes of a signal gives a sparser, alternative signal representation. The LBP operation is fast and computationally inexpensive. It was shown to be a distinctive marker of certain features of the underlying signal. This property has been applied in preliminary work for simple signal segmentation and VAD. The 1-D LBP is able to distinguish the unvoiced components of speech signals using the distinguishing features of higher activity in certain characteristic histogram bins. The performance of the improved 1-D LBP VAD technique was shown to be superior to the G.729 VAD. Depending on the application, VAD techniques attempt to detect more speech activity or speech pauses, while keeping false alarm rates low. The improved 1-D LBP VAD was proposed as a computationally efficient technique, which aims at achieving this compromise. It also does not incorporate an independent noise estimation routine.

5

The Cocktail Party Effect: A Complex Auditory Scene

5.1 Introduction

The problem of attending to, selecting and understanding individual speakers in an environment with a mixture of speakers is known as the “cocktail-party problem” (Cherry, 1953, Arons). Therefore, a common task in signal processing in hearing devices is the separation of a target speaker from a mixture of speakers to improve speech intelligibility in these noisy environments. A target speaker separation system is summarized as shown in Figure 5.1, where s_i is the speech signal corresponding to the i^{th} source for i=1,2,...M . In this illustration, the target speaker can be arbitrarily defined. With more than one speaker (i>1) in Figure 5.1, the speech mixture signal is measured using the j^{th} microphones, x_j, for j=1,2...N. Target speaker separation can be performed using techniques which rely on acoustic cues of the source signals. Some examples for separating speech signals using cues such as harmonicity, onset times and spatial direction of the speakers are reviewed in (Darwin, 2008).

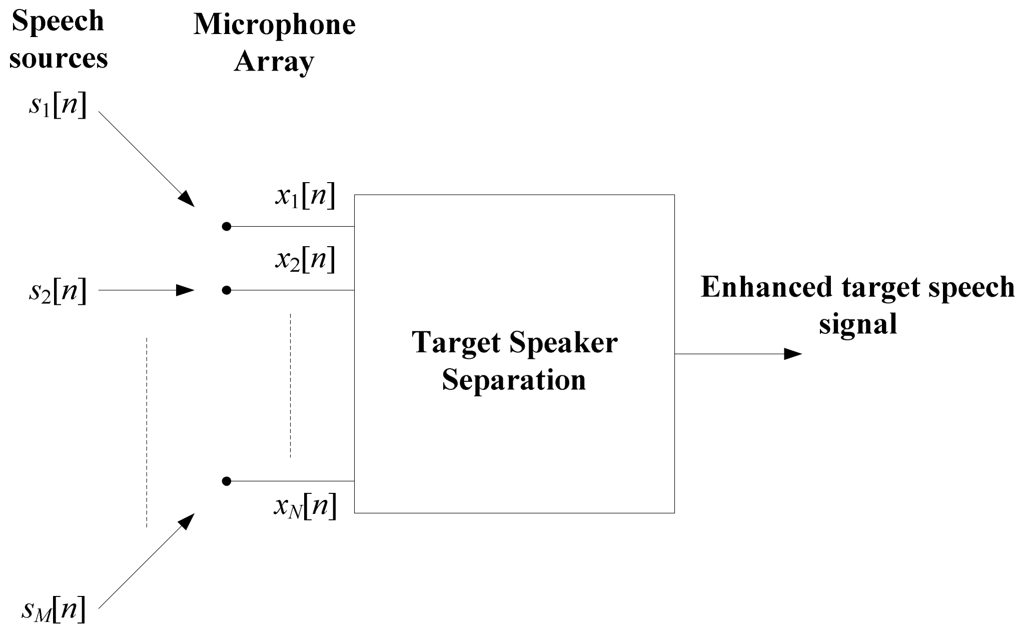


Figure 5.1: General system diagram illustrating enhancement of target speaker from a mixture of speakers

One of the research objectives in this thesis was to design a novel system for hearing aid users, which performs noise reduction and speech enhancement of signals received in environments with spatially localized interfering speakers. This system should steer the focus direction of the hearing aid to additional directions as well as 0/180 degrees. In this chapter, several recent target speaker separation techniques and speech demixing techniques will first be reviewed in section 5.2. In this thesis, binaural hearing aids that have a wireless transmission link between both hearing aid sides will be considered. Section 5.3 provides a brief background on typical hearing aid's signal processing functionality, followed by a review of recent techniques for achieving noise reduction in multi-talker scenarios, which are applicable to binaural hearing aids. Finally, conclusions will be made in Section 5.4.

5.2 Review of Target Speaker Separation

Blind source separation (BSS) (Cardoso, 1998) and Computational Auditory Scene Analysis (CASA) (Brown and Cooke, 1994) are common approaches for separating target sources from speech mixtures. BSS can be performed in the time and frequency domain whereas CASA is performed in the frequency domain by analyzing auditory scenes (Bregman, 1990). In this section, BSS and CASA techniques will reviewed and a detailed survey of recent approaches for separating speech mixtures is given in (Pedersen et al., 2007).

5.2.1 BSS of speech mixtures

Referring to the signals given in Figure 5.1, a generalization of the time-domain BSS approach is given below, where the speech mixture is first formulated as:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \qquad \textbf{(5.1)}$$

where $\mathbf{s} = \left[s_1[n]\ \mathrm{s}_2[n] ... s_M[n]\right]^T$ is the source matrix, $\mathbf{A}$ is the mixing matrix and $\mathbf{x} = \left[x_1[n]\ \mathrm{x}_2[n] ... x_N[n]\right]^T$ is the matrix observed of the mixed sources. The unmixing matrix $\mathbf{W}$ must then be estimated such that the sources may be identified as:

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{\Lambda}\mathbf{P}\mathbf{s} \qquad \textbf{(5.2)}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of amplitudes and $\mathbf{P}$ is a permutation matrix (Moon and Stirling, 1999). The main assumption behind this technique is the statistical independence of the sources $\mathbf{s}$ and of the outputs $\mathbf{y} = \left[\hat{s}_1[n], \hat{s}_2[n], ..., \hat{s}_M[n]\right]^T$. The frequency domain BSS generalization can be correspondingly derived as given in (Pedersen et al., 2007).

Many frequency domain BSS techniques perform ICA on each frequency bin to derive independent basis vectors corresponding to each source for mixture separation. However, these techniques are affected by the scaling and permutation ambiguity. In (Molla and Hirose, 2007), audio sources are separated from a single mixture (N=1) using EMD by performing subspace decomposition of the Hilbert spectrum (HS) (Huang et al., 1998) of the extracted IMFs. The results presented in their work were for mixtures which are separated by clustering the basis vectors extracted. The advantages of using the Hilbert spectrum as opposed to the STFT for separating the component sources are due to the higher resolution using the HS, the lower spectral overlap of component sources in the HS and the inverse transform of the HS in the resynthesis stage. The HS is derived from the analytic signal corresponding to the m^{th} IMF and the analytic signal z_m is defined as:

$$z_m = IMF_m + jh\left[IMF_m\right] \tag{5.3}$$

where $j=\sqrt{-1}$ and $h[.]$ is the Hilbert transform. From Equation (5.3), the real valued m^{th} IMF is unaffected, the inverse signal is obtained by filtering the imaginary part of the HS of the extracted IMFs. The results presented in (Molla and Hirose, 2007), only considered mixtures with two sources, and the performance improvements obtained were dependent on the amount of spectral overlap between the source signals. This technique is computationally intensive, since the separation is performed on all IMFs.

Recent approaches for multi-microphone (N>1) BSS proposed in (Nesta et al., 2008a, Nesta et al., 2008b) were shown to be robust to the permutation problem by estimating the Time Difference of Arrival (TDOA) corresponding to each source. This technique involves measuring coherence in the frequency domain between the received microphone signals, **x**. However, this technique was shown to be sensitive to the sensor spacing (25 cm. spacing required at a sampling frequency of 16 kHz) (Nesta et al., 2008b). It is also sensitive to the presence of large physical objects (e.g. human head in hearing aid users) between the microphones.

Other recent promising BSS techniques for speech have been proposed and are explained in (Pedersen et al., 2007, Makino et al., 2007a). However, one of the popular approaches, termed Degenerate Unmixing Estimation Technique (DUET), will be the focus of the following section.

5.2.1.1 DUET

In Figure 5.1, when the number of sources is more than the number of recorded microphone mixtures (M>N), BSS becomes an underdetermined problem and can be termed as *degenerate* since the mixing matrix $\boldsymbol{A}$ is not invertible. DUET was proposed in (Jourjine et al., 2000) to unmix sources from *anechoic* mixtures of attenuated and delayed sources. Two mixtures (*N*=2) of *M* speech sources which are measured by two of the sensors from Figure 5.1 can be given by:

$$x_1[n] = \sum_{i=1}^{M} s_i[n] \tag{5.4}$$

$$x_2[n] = \sum_{i=1}^{M} a_i s_i(n - \Delta_i) \tag{5.5}$$

where s_i is a source signal as defined before, Δ_i is the arrival delay between the two sensors due to the direction of arrival (DOA) of the sources and a_i is the relative attenuation of the source signals measured at the microphones. DUET is based on the assumption that the STFT of distinct sources s_i from the mixtures form W-Disjoint Orthogonal (WDO) sets as explained in (Jourjine et al., 2000), where WDO implies that only one source should be active at any time-frequency pair. WDO can alternatively be stated as:

$$S_1(\Omega,\tau)S_2(\Omega,\tau) = 0; \forall\Omega,\tau \tag{5.6}$$

where for simplicity, the assumption is made that only two sources are present and $S_1(\Omega,\tau)$ and $S_2(\Omega,\tau)$ are the STFT of the two sources, using a finite support window. It was later shown that the STFT of the distinct sources can be assumed to be approximate WDO sets as explained in (Yilmaz and Rickard, 2004, Makino et al., 2007b)

The mixture separation problem is solved by first estimating the relative attenuation and delay associated with each of the sources, and then generating a histogram to create a binary time-frequency mask that can be used to partition the sources. The DUET algorithm is summarized in Appendix A and was simulated in MATLAB to produce the histogram plot shown in Figure 5.2 to accurately estimate the unmixing parameters of delay and relative attenuation for a mixture of four source signals (*M*=4 from Equations (5.4)and (5.5)).

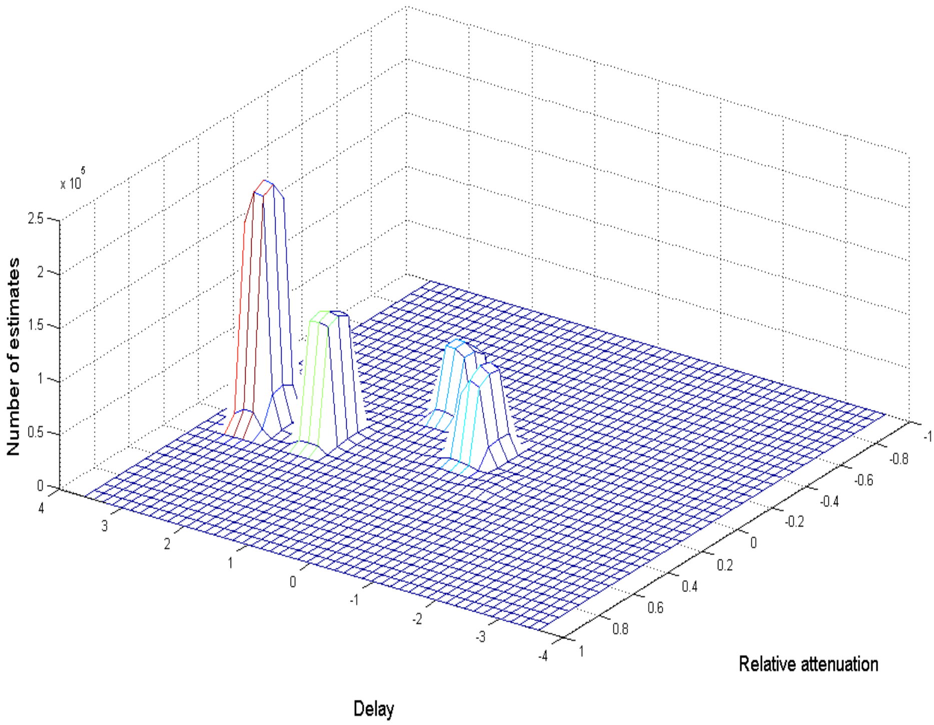


Figure 5.2: Histogram plot used to estimate the delay/attenuation parameters for the four source signals using the DUET algorithm

These estimates can then be used to unmix the four sources from the given mixtures. However, it can be seen that DUET requires prior knowledge of the number of sources present but this is not known in real situations. It was shown that DUET requires the estimation of the delay and attenuation parameters. Therefore, improved results are obtained when the STFT of the speech mixture is sparser, i.e. when there is reduced overlap of the sources in the time-frequency bins. DUET was also shown in (Jourjine et al., 2000) to be sensitive to the microphone spacing in the receiver array. A method was proposed in (Kim and Park, 2010) to deal with some of these problems by defining the approximate direction of the target speaker. Thus, only one set of parameters need to be estimated and this reduces the algorithm's convergence time.

An extension of the DUET algorithm known as DUET-ESPRIT (DESPRIT) has been proposed which combines the signal subspace technique of ESPRIT (Roy and Kailath, 1989) and the weighted histogram of DUET to separate sources when more than two mixtures are available and the detailed algorithm is given in (Rickard et al., 2005). The assumptions made by DUET and DESPRIT, as well as the robustness of these algorithms with *echoic* mixtures, are extensively covered in (Melia, 2007).

5.2.2 Computational Auditory Scene Analysis

CASA (Brown and Cooke, 1994) refers to the analysis of the auditory scene which can be assumed to comprise multiple speakers and background noise and is applicable to scenarios where one target source needs to be extracted from a noisy speech signal. CASA separates a target source from the noisy speech mixture by forming two auditory streams: a foreground speech stream with the target speaker and a background stream with the residual noise and interfering speech components. This procedure is summarized in Figure 5.3.

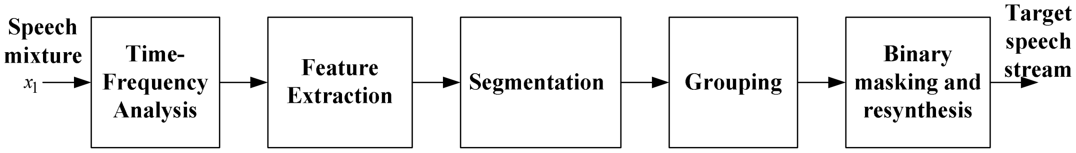
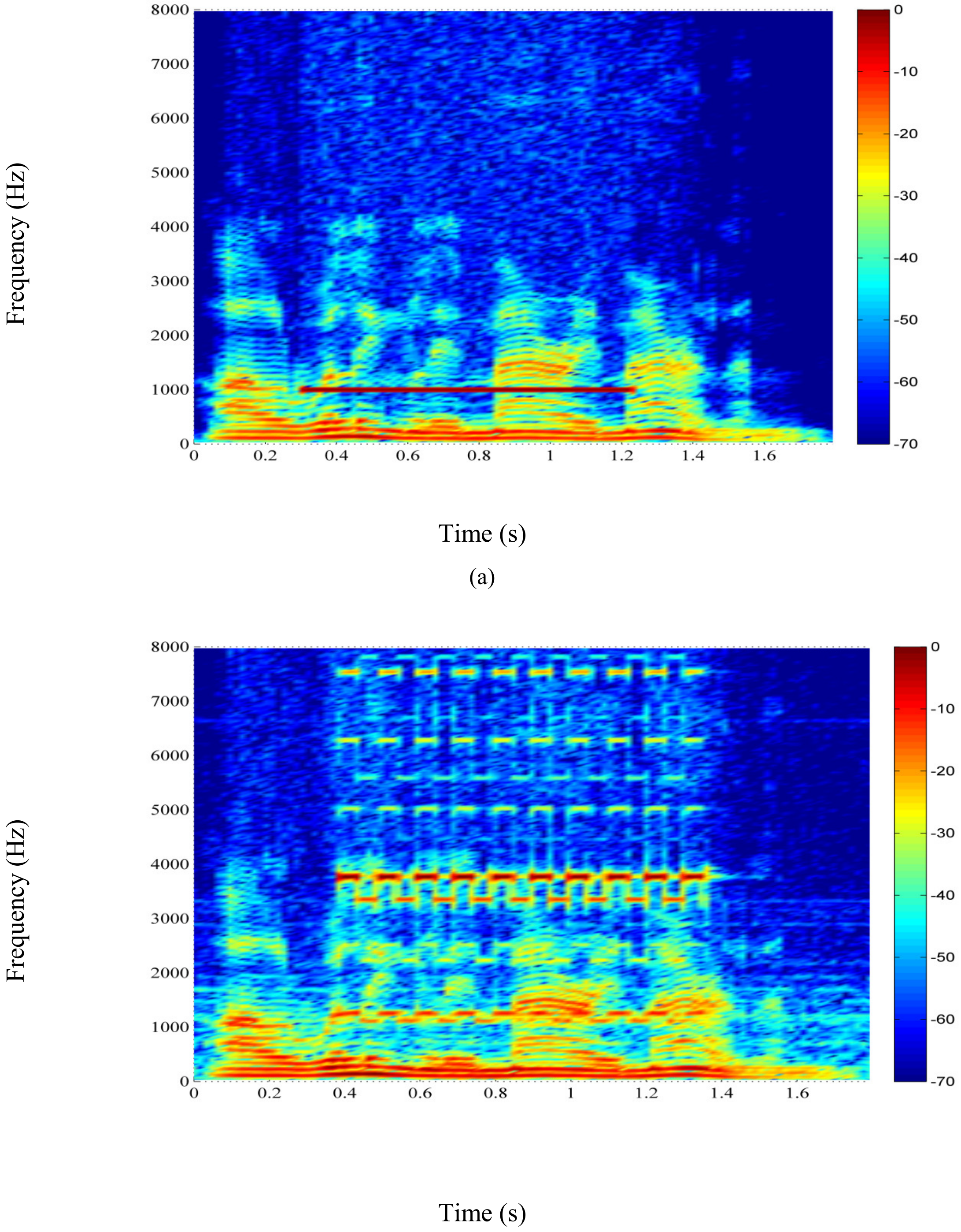


Figure 5.3: Summary diagram of CASA system

The motivation for CASA was provided by the work in (Bregman, 1990) on Auditory Scene Analysis (ASA). It was proposed that the auditory system groups sounds that share common acoustic cues to a specific source, whereas sounds which differ likely originated from different sources. As shown in Figure 5.3, time-frequency transformation of the measured noisy signal using spectral analysis techniques is first performed using techniques such as STFT. Next, auditory features are extracted from the voiced and the unvoiced speech components. For voiced components, features such as pitch and harmonicity, amplitude modulation and spatial direction may be used. Analysis of sounds with common onset and offset times may be used as a cue for both voiced and unvoiced speech components. This is covered in further detail in (Hu and Wang, 2006, Hu and Wang, 2007, Hu and Wang, 2004, Jin and Wang, 2009, Brown and Cooke, 1994, Shao et al., 2010). Following the feature extraction stage from Figure 5.3, local segments are formed by merging time-frequency units corresponding to common auditory cues. Then, grouping is performed on the extracted segments which belong to the target source. This results in the separation of the two speech streams and the derivation of a binary mask for the time-frequency units, where '1' corresponds to the foreground stream and '0' corresponds to the background stream.

This masking has been shown to improve speech intelligibility in (Li and Wang, 2009, Wang, 2005, Kjems et al., 2010) with the use of ideal binary masks (IBM) which was proposed to be the *computational objective* in CASA. An IBM was defined in their work as a binary mask which assigns '1' if the target energy of the time-frequency unit exceeds that of the interference energy. Otherwise, it is assigned a value of '0'. However, these methods for CASA are sensitive to the SNR of the measured signals, the spectral overlap of the sources in the time-frequency bins and the temporal continuity of the extracted features. This was demonstrated in (Hu and Wang, 2004) where the highest SNR improvements using CASA were obtained for noises such as a 1 kHz tonal noise and a telephone ring which both have low spectral overlap with the target speech source. The spectrograms of these two example mixtures used (Cooke, 1993) are shown in Figure 5.4(a) and Figure 5.4(b) below for the same speech utterance. It can be seen that the speech utterance also comprised minimal pauses in between the words which gave temporal continuity for improved extraction of pitch contours.



(a)

(b)

Figure 5.4: Spectrograms of two of the mixtures of a speech utterance and noise used for testing CASA (a) Speech mixed with 1kHz tonal noise (b) Speech mixed with telephone ring

5.3 Noise Reduction in Hearing Aids

A common signal processing problem encountered in hearing aids is the reduction of interfering noise sources in the presence of desired speech signals to improve the speech intelligibility. Adaptive signal processing techniques such as adaptive noise cancellation are commonly used when enhancing speech, with (SEAEMD from Section 3.2) or without (EMDF techniques from Section 3.3 and Section 3.4) a noise reference. Alternative approaches exist (Loizou et al., 2007), however in the case of non-stationary signals such as speech in complex hearing environments with multiple speakers, directional signal processing is required to improve speech intelligibility by enhancing the desired signal (Hamacher et al., 2005). CASA was reviewed in Section 5.2.2, and a study of its potential for use in hearing aid for noise reduction was performed in (Wang, 2008). In that study, it was concluded that CASA systems were not yet useful improving intelligibility in hearing aids due to factors such as speech distortion and residual noise resulting from the binary masking.

State-of-the-art hearing aids use two microphones in each device due to compactness and power consumptions constraints. These hearing aids utilize simple differential microphone arrays (DMAs) (Teutsch and Elko, 2001) to focus on targets in front or behind the user. In many hearing situations, the desired speaker azimuth varies from these predefined directions. Therefore, by using spatial information about the desired target's position, directional signal processing allows the beam to be steered to the desired focus direction in order to enhance the target source.

In directional signal processing algorithms, the sensor spacing is an important consideration. If the sensors spacing is too far apart with respect to the wavelength λ of the source signal, then spatial aliasing will occur. This results in incident sources from different locations producing the same array propagation vector, denoted as $d(\theta,\Omega)$. However, if the sensors are too close, then spatial discrimination is reduced since the aperture is smaller than required. Generally, the maximum distance allowed for the sensor spacing is half the wavelength corresponding to the maximum frequency component in the source signal.

5.3.1 Directional Signal Processing for Binaural Hearing Aids

Monaural hearing aids process the received microphone signals at each device independently. However, we will consider binaural hearing aids that have a wireless transmission link between both hearing aid sides. Binaural hearing aids are beneficial since they utilize and process the received microphone signals from both ears. Typical acoustic noise environments consist of mixtures of a diffuse noise field and directional interferers. In a diffuse noise field environment, techniques in (Kamkar-Parsi and Bouchard, 2009) use the received signal from both left and right hearing aid to estimate the noise spectrum which can then be used for noise reduction. The estimation of the noise PSD (Power Spectral Density) in one side of the hearing aid relies on the difference between the predicted noisy signal from the alternate hearing aid and the actual noisy signal transmitted. The technique was shown to work with speech in the presence of diffuse noise by exploiting the high degree of correlation of speech measured at both hearing aids at low frequencies. The results presented demonstrated its robustness in non-stationary noise and its ability to track the noise power spectrum even during speech activity.

For scenarios with directional interferers, recent approaches for binaural beamforming have been presented in (Widrow and Luo, 2003, Daoud et al., 2009, Lotter and Vary, 2006, Rohdenburg et al., 2007, Doclo et al., 2009, Roy and Vetterli, 2006, Srinivasan and den Brinker, 2009a, Srinivasan and den Brinker, 2009b). Beamforming is popular in receiver arrays where it performs spatial filtering, i.e. it filters received signals based on their spatial location, even when the individual signals overlap in the frequency domain. In (Daoud et al., 2009), a binaural delay-and-sum beamformer was implemented using the microphone signals from both sides of the hearing aid user’s head. This system steers to target speakers on either side of the user; however it only performs steering at low frequencies.

The binaural speech enhancement algorithm designed in (Lotter and Vary, 2006) comprises a cascaded superdirective beamformer (Bitzer and Simmer, 2001) with a post-filtering gain to enhance each microphone signal. The post-filter combines the beamformer output with the signals measured at both hearing aids to produce a gain which incorporates binaural information. In (Rohdenburg et al., 2007), a binaural beamformer was designed using a configuration with two 3-channel hearing aids. The beamformer constraints were set based on the desired look direction to achieve a steerable beam with the use of three microphones in each hearing aid. This was done using the minimum variance distortionless response (MVDR) beamformer (Veen and Buckley, 1998) which is given by:

$$\min_{\mathbf{w}} \mathbf{w}^{\mathbf{H}}\mathbf{R}_{\mathbf{xx}}\mathbf{w} \text{ subject to } \mathbf{C}^{\mathbf{H}}\mathbf{w} = \mathbf{1} \tag{5.7}$$

where $\mathbf{w}$ is the beamformer weight vector, $\mathbf{R_{xx}}$ is the covariance matrix of the microphone array signals. $\mathbf{C}$ is the constraint matrix containing the directional propagation vector for both the desired sources and the interferers from different directions. Essentially, a constraint is placed on the beamformer so that it passes the signal from the look direction with the defined gain and phase, while minimizing the output power due to interferers from other directions. The post-filter gain proposed in (Lotter and Vary, 2006) was also applied to produce the processed binaural output signals at each hearing aid. This design requires the use of three microphones in each hearing aid which is impractical and requires a high computational cost for current hearing aids. In addition, the system performances from (Lotter and Vary, 2006) and (Rohdenburg et al., 2007) were shown to be dependent on the propagation model used in formulating the steering vector.

Binaural multi-channel Wiener filtering (MWF) was used in (Doclo et al., 2009) to obtain a steerable beam by estimating the statistics of the speech signal in each hearing aid. MWF is computationally expensive and the results presented were achieved using a perfect voice activity detector (VAD) to estimate the noise while assuming the noise to be stationary during speech activity. Theoretical analysis and comparison of MWF schemes were presented in (Cornelis et al., 2010). In (Srinivasan and den Brinker, 2009b, Srinivasan and den Brinker, 2009a), a performance evaluation was presented between the relationship between data transmission rate of the wireless link in binaural hearing aids and the beamforming gain using a MWF scheme. For a rate constrained transmission link, an investigation was conducted on which signal should be transmitted between the hearing aids: an estimate of the target source, an estimate of the noise or the unprocessed signal. The decision was shown to be SNR dependent, however it was concluded that the unprocessed signal should be transmitted. This also results in reduced computational expenses which would arise in noise or signal estimation and also lower delays.

Various techniques for forming different directivity characteristics using microphone arrays have been reported in (Ihle, 2003), (Eichler and Lacroix, 2008) and(Ogawa et al., 2008). In (Ihle, 2003), one spatial null is achieved in a desired direction using DMAs with three microphones to estimate the ambient noise and then perform spectral subtraction to obtain the enhanced signal. In (Eichler and Lacroix, 2008) and (Ogawa et al., 2008), the directional response is obtained by combinations of the signals received at the microphone sensors configured as multi-pole arrays. However, these three methods were all shown to be sensitive to microphone array geometries which are not applicable to a typical hearing aid setup.

5.4 Conclusion

This chapter reviewed techniques for performing noise reduction for a desired speech signal, contaminated with interfering speakers. Hearing aids require computationally inexpensive signal processing approaches for this problem to keep power consumption and processing delay low. Therefore, this must be considered when designing a method to combat this noise reduction problem in a hearing aid.

6

Novel Direction Dependent Spatial Noise Reduction Techniques for Binaural Hearing Aids

6.1 Introduction

In this chapter, new techniques are presented for binaural systems in the hearing aid which can focus on additional directions as well as 0° and 180° using Differential Microphone Arrays (DMAs) and filtering techniques. A realistic size constraint is set such that each hearing aid uses two microphones. Due to the data transmission constraint in a binaural system, only one microphone signal is transmitted from each hearing aid to the other using a wireless data link. Recent and relevant techniques for noise reduction in binaural hearing aids were presented in Section 5.3.1, using directional signal processing. It was shown that the constraints imposed on real hearing aids as outlined above, were not satisfied by the previously outlined methods.

Hearing aids require computationally inexpensive algorithms with low latency. Therefore, most hearing aids utilize first order DMAs to perform its directional signal processing. Frequency domain beamforming involves transforming the data into the frequency domain and performing spatial-filtering in a narrow band. Filter banks are commonly found in practice and a major motivation of this approach is the decomposition of a full-band problem into a set of smaller sub-band problems. Therefore, the combination of filter banks with first order DMAs is a computationally efficient implementation of the hearing aid's directional signal processing.

In section 6.2, the binaural hearing aid setup is first presented along with the background necessary to understand DMA beamforming which is used for steering to 0° and 180°. In section 6.3, a "side-look" beamformer is developed which focuses its beam to either side of the head. The proposed technique decomposes the problem to process the low frequencies (< 1 kHz) and the high frequencies (≥ 1 kHz) independently. For the low frequencies, a binaural array is used and for the high frequencies, the head shadow effect is utilized to develop a system to achieve the side look. In section 6.4, a steerable binaural beamforming system is presented which can focus its beam to any desired source located at a range of given azimuths for the frequency range approximately up to 750 Hz. The proposed technique involves filtering of the noisy signal using an estimate of the desired source signal and an estimate of the noise signal. The noise estimate is obtained by combining the local microphone signals in each hearing aid with the single microphone signal wirelessly received from the alternate hearing aid.

Section 6.5 demonstrates the effectiveness of the systems using directivity plots from actual hearing aid signals in a real time environment. The Signal to Interference (SIR) gains obtained using the steerable beamformer system to attenuate multiple directional interferers, competing with a target speaker, are also presented. Finally, conclusions are made in section 6.6.

6.2 Binaural hearing aid set-up

The proposed scheme for the binaural hearing aid is illustrated in Figure 6.1, where the left and the right hearing aids are connected by a bidirectional wireless link. Size constraints impose that each hearing aid has two microphones separated by a distance of approximately 1cm. Due to practical rate constraints and minimization of power consumption, only one microphone signal is transmitted from each hearing aid to the other over a bidirectional audio link.

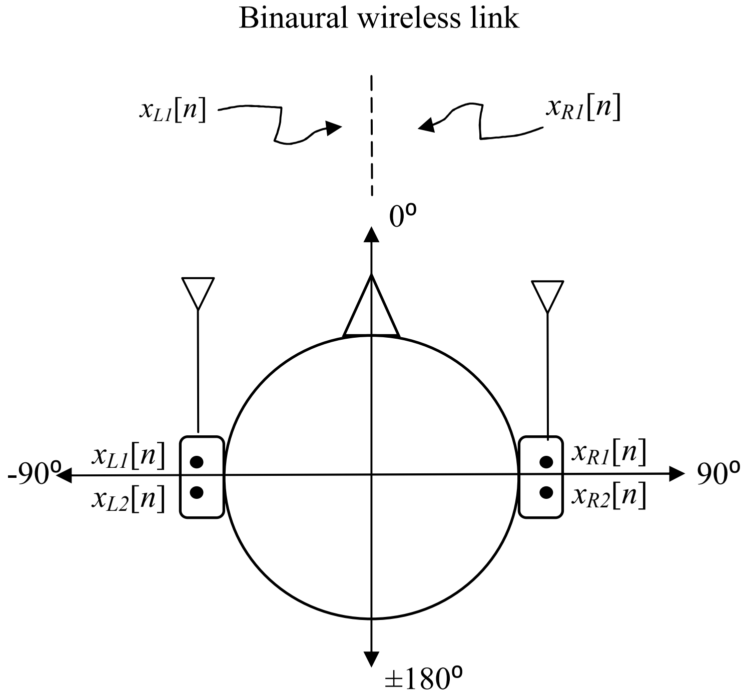


Figure 6.1: Binaural hearing aid configuration with wireless

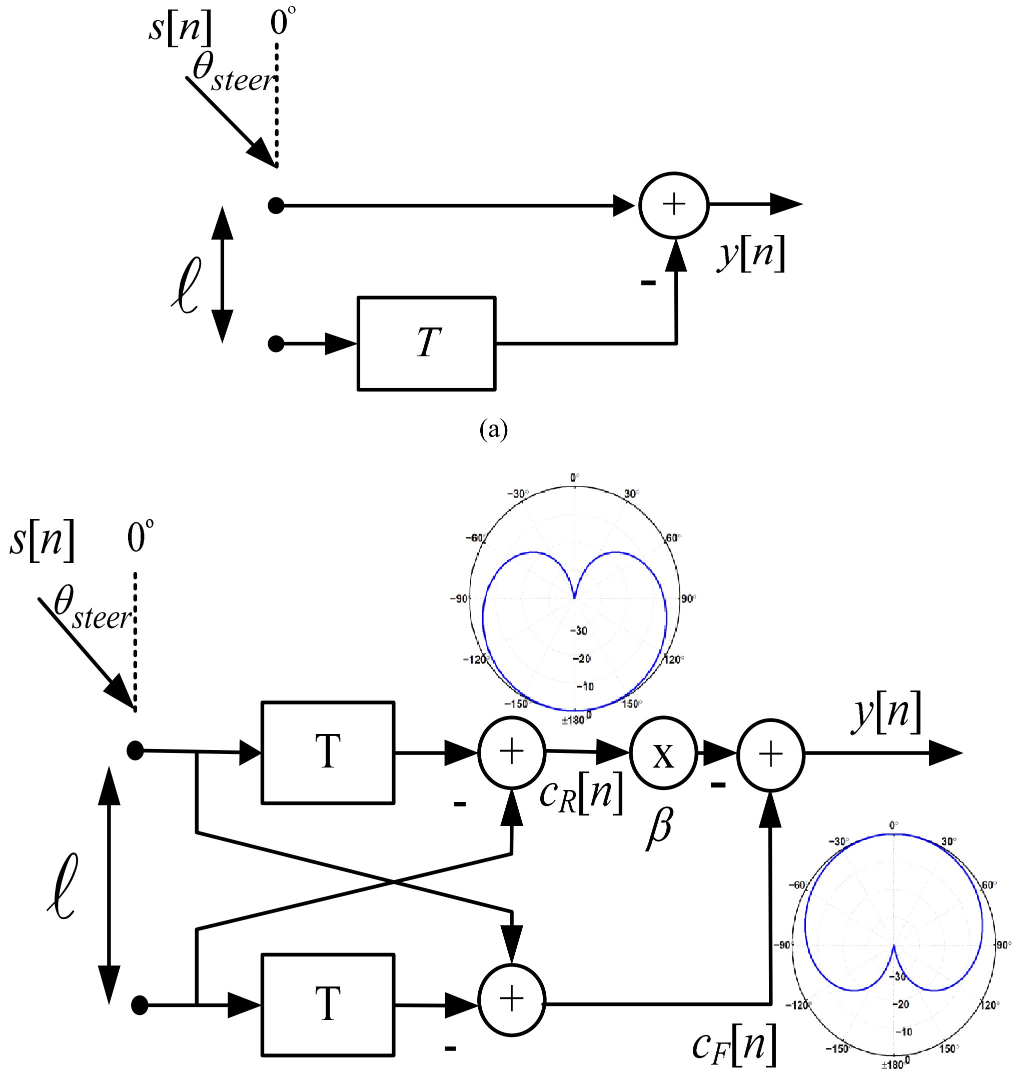
In Figure 6.1, the signals $x_{Li}[n]$ and $x_{Ri}[n]$ are the i^{th} microphone signals from the left and the right hearing aid respectively, where i=1,2. In Figure 6.1, $x_{L1}[n]$ corresponds to the signal being transmitted from the front microphone in the left hearing aid to the right hearing aid. Without loss of generality, $x_{R1}[n]$ corresponds to the right front microphone signal being transmitted to the left hearing aid. The main goal of the steerable beamforming work is the ability to steer the beam to specific directions Θ_{steer}, where Θ_{steer} is the set of 8 angles s.t.:

$$\Theta_{steer} = \pm 45 * a^{\circ} \ \forall \ a = 0,1..,4 \qquad \textbf{(6.1)}$$

and θ_{steer} will be used to denote an arbitrary angle from this set Θ_{steer}.

6.2.1 Differential Microphone Array (DMA)

Traditional monaural hearing aids (Section 5.3.1) use a first order DMA with two omni-directional microphones separated by a distance l (approx. 1 cm in an individual hearing aid) to generate a directional response. Its response is independent of frequency as long as the assumption of small spacing to acoustic wavelength, λ, holds. Beam steering to 0° and 180° is achieved using the basic first order DMA and adaptive DMA (ADMA), as described in this section.



(a)

(b)

Figure 6.2: (a)First order microphone array (b)Adaptive differential microphone array system

Consider the signal $s[n]$ impinging on the first order DMA at an angle θ_{steer} as illustrated in Figure 6.2(a). Under farfield conditions, the magnitude of the frequency (Ω) and angular (θ_{steer}) dependent response of the array is given by (Teutsch and Elko, 2001):

$$\left|H(\Omega,\theta_{steer})\right| = \left|1 - e^{-j\Omega(T+\frac{l}{v}\cos\theta_{steer})}\right| \qquad \textbf{(6.2)}$$

where v is the speed of sound. The delay T may be adjusted to cancel a signal from a certain direction to obtain the desired directivity response. A fractional delay may be implemented for small microphone spacings (Valimaki, 2000).

In conventional monaural hearing aids, the desired speaker is assumed to be in front of the hearing aid user and the microphone spacing l in an individual hearing aid may be denoted as l_{mon}, where l_{mon}=1 cm. The delay T is fixed to match the microphone spacing l_{mon}/v and sounds from the front first impinge on the front microphone. The desired directivity response is instead achieved using a back-to-back cardioid system as shown in the ADMA in Figure 6.2(b). From Figure 6.2(b), $c_F[n]$ is the cardioid beamformer output that steers the beam to the front (0°) and attenuates signals from the back direction. Correspondingly, $c_R[n]$ is the anti-cardioid beamformer output that steers the beam to the back (180°) and attenuates signals from the front direction. Using a steering parameter β, the array output $y[n]$ is given by:

$$y[n] = c_F[n] - \beta c_R[n] \tag{6.3}$$

Under this assumption of the location of the desired speaker, the output signal $y[n]$ in Equation (6.3) does not attenuate the signal from 0°. A single spatial notch is formed in the direction θ_{null} for a value of β given by (Teutsch and Elko, 2001):

$$\theta_{null} \simeq \arccos\frac{\beta-1}{\beta+1} \tag{6.4}$$

where the constraint for $0 \leq \beta \leq 1$ places the notch in the back-half plane for $90^{\circ} \leq \theta_{null} \leq -90^{\circ}$. It is known that the head shading effect results in phase and amplitude mismatch of the signals received at the microphones which varies with frequency (Puder, 2006). Therefore, in practical cases, the ideal value of the steering parameter β from Equation (6.4) varies slightly with frequency in order to place the notch in the desired direction.

In noisy environments, the parameter β can be adapted to steer the notch to direction θ_{null} of a noise source to optimize the directivity index. This is performed by minimizing the MSE of the output signal $y[n]$ from Equation (6.3). Using a gradient descent technique to follow the negative gradient of the MSE cost function, the parameter β is adapted by (Teutsch and Elko, 2001):

$$\beta[n+1]=\beta[n]-\mu\frac{\partial}{\partial\beta}E\left(y^{2}[n]\right) \quad \textbf{(6.5)}$$

where μ is the update step size and $E(.)$ is the expectation operator.

In environments such as car interiors, the desired speaker may be behind the hearing aid user at a direction of 180°. The ADMA system shown in this section is used for steering to this desired speaker, by simply swapping the received front and back microphone signals in the analysis presented. In this configuration, sounds from the back (180°) first impinge on the back microphone and this signal is delayed and subtracted as in the DMA technique already described. Therefore, from Equation (6.3), $c_F[n]$ now points to the 180° and $c_R[n]$ now steers to 0°, and the remaining processing steps are performed as before.

6.3 Binaural Side-Look Steering

It was mentioned previously that in certain hearing situations such as car interiors, the desired speaker may be behind the hearing aid user. In such listening environments, the scenario may arise where the desired speaker is on one side of the hearing aid user. Therefore, in this section, a system which performs side-look beam steering is realized using binaural hearing aids with a bidirectional audio link. It is known that at high frequencies, the Interaural Level Difference (ILD) between measured signals at both sides of the head is significant due to the head-shadowing effect. This is demonstrated where a binaural hearing aid system was set up as described in Section 6.2 with two “Behind the Ear” (BTE) hearing aids on each ear of a KEMAR dummy head. A zero-mean, white Gaussian noise, $w[n]$ was incident at an angle 90° and the two omni-directional signals were measured at the right hearing aid and the left hearing aid with sampling frequency f_s = 24 kHz. Figure 6.3 shows the power spectrum of the measured signals at the right and the left hearing aids. From the plot, it can be seen that ILD increases with frequency.

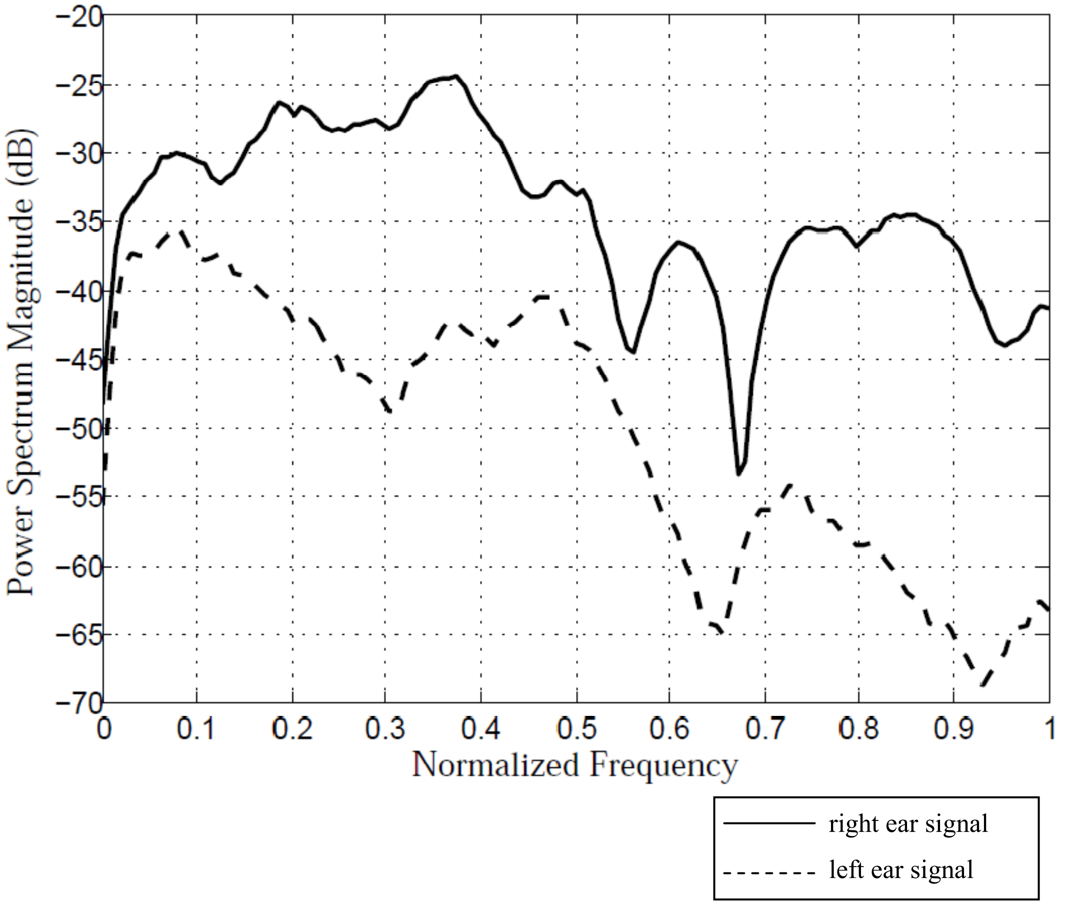


Figure 6.3: Power spectrum of left and right microphone signals illustrating ILD

This head-shadowing effect is exploited in the design of the binaural Wiener filter for the side-look steering system at high frequencies (≥ 1 kHz). At low frequencies (< 1 kHz), the acoustic wavelength λ is long with respect to the head diameter. Therefore, there is minimal change between the sound pressure levels at both sides of the head and the Interaural Time Difference (ITD) is the more significant acoustic cue. At low frequencies, a binaural first-order ADMA is designed to create the side-look. This side-look steering is decomposed into two smaller problems with a binaural ADMA for the low frequencies and a binaural Wiener filter approach for the high frequencies. The proposed system diagram is shown in Figure 6.4.

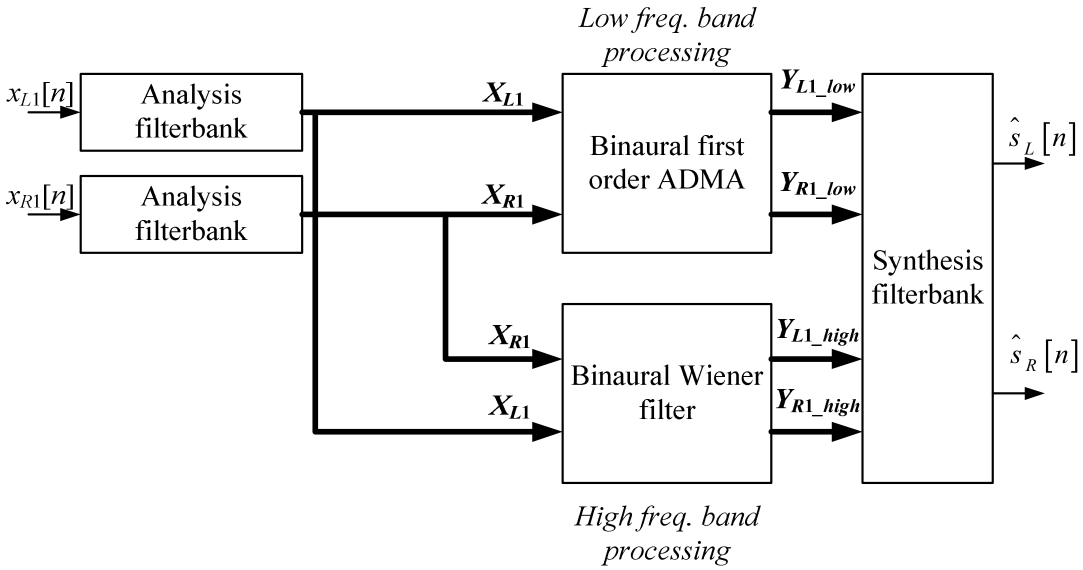


Figure 6.4: Block diagram of side-look steering system

The input noisy speech signals $x_{L1}[n]$ and $x_{R1}[n]$ comprise the desired speech signal $s[n]$ incident from one of 2 directions $\theta_{steer} = \pm 90°$, the noise signal $d[n]$ incident from direction θ_d, where correspondingly, $\theta_d = \mp 90°$. These input signals are decomposed into sub-bands by the analysis filterbank (Bauml and Soergel, 2008) to produce the frequency domain sub-band signals $X_{L1,k}$ and $X_{R1,k}$, where the index k=0,1..Q-1 refers to the k^{th} sub-band and Q is the number of sub-bands. The proposed directional processing systems presented in the next sections are then applied to yield the processed signals from the low frequency and the high frequency processing blocks. As shown in Figure 6.4, the outputs from the low frequency block are given as $\mathbf{Y_{L1\_low}}=[Y_{L1\_low,0}..Y_{L1\_low,P\text{-}1}]^T$ and $\mathbf{Y_{R1\_low}}=[Y_{R1\_low,0}..Y_{R1\_low,P\text{-}1}]^T$, where the first P sub-bands are included in the low frequency processing block (< 1 kHz). The outputs from the high frequency processing block are given by $\mathbf{Y_{L1\_high}}=[Y_{L1\_high,P}..Y_{L1\_high,Q\text{-}1}]^T$ and $\mathbf{Y_{R1\_high}}=[Y_{R1\_high,P}..Y_{R1\_high,Q\text{-}1}]^T$. These signals are then reconstructed using a synthesis filterbank (Bauml and Soergel, 2008). The signal from the side of the interferer is termed the interferer side and the signal on the side of the desired source is termed the focus side. A bidirectional audio link between the hearing aids is assumed.

6.3.1 High Frequency Side Look

The head-shadowing effect is exploited in the design of a binaural system to perform the side-look at high frequencies. The signal from the interferer side is attenuated across the head at these high frequencies. The analysis of the proposed system is given below.

6.3.1.1 System Model

Consider the scenario where a target speaker $s[n]$ is on the left side (θ_{steer} = -90°) of the hearing aid user and an interferer $d[n]$ is on the right side (θ_d = 90°). Figure 6.4 considers the left ear signal model $x_{L1}[n]$ recorded at the front left microphone and the right ear model $x_{R1}[n]$ recorded at the front right microphone given by:

$$x_{L1}[n] = s[n] + h_{L1}[n] * d[n] \tag{6.6}$$

$$x_{R1}[n] = h_{R1}[n] * s[n] + d[n] \tag{6.7}$$

where, $h_{L1}[n]$ is the transfer function from the front right microphone to the left front microphone, and $h_{R1}[n]$ is the transfer function from the front left microphone to the front right microphone. The related frequency domain subband signals are given by Equations (6.8) and (6.9) below:

$$X_{L1,k} = S_k + H_{L1,k}\ D_k \tag{6.8}$$

$$X_{R1,k} = H_{R1,k}\ S_k + D_k \tag{6.9}$$

Let the short-time spectral power of the subband signal X_k be denoted as Φ_k. The power of the signals from the focus side and the interferer side are given by $\Phi_{focus,k}$ and $\Phi_{\text{int},k}$ respectively. A classical Wiener filter W_k can be derived as:

$$W_k = \frac{\Phi_{focus,k}}{\Phi_{focus,k} + \Phi_{\text{int},k}} \tag{6.10}$$

For analysis purposes, it is assumed that $\Phi_{H_{L1,k}} = \Phi_{H_{R1,k}} = \alpha_k$ where α_k is the frequency dependent attenuation corresponding to the transfer function from one hearing aid to the other across the head. Therefore, Equation (6.10) can be simplified to:

$$W_k = \frac{\Phi_{S,k} + \alpha_k \Phi_{D,k}}{\left(1+\alpha_k\right)\left(\Phi_{S,k} + \Phi_{D,k}\right)} \tag{6.11}$$

As explained earlier in this section, at high frequencies the ILD attenuation $\alpha_k \to 0$ due to the head-shadowing effect and therefore Equation (6.11) tends to a traditional Wiener filter. At low frequencies, the attenuation $\alpha_k \to 1$ and the Wiener filter gain $W_k \to 0.5$. This limiting gain can be scaled to 1 for the low frequencies by using a multiplicative factor of 2 in W_k. The output filtered signal at each side of the head is obtained by applying the gain W_k to the omni-directional signal at the front microphones on both hearing aid sides.

Let $\mathbf{W}_{\text{high}}=[W_{P,}..W_{Q\text{-}1}]$, $\mathbf{X}_{\text{L1\_high}}=[X_{L1,P,}..X_{L1,Q\text{-}1}]$ and $\mathbf{X}_{\text{R1\_high}}=[X_{R1,P,}..X_{R1,Q\text{-}1}]$ and the outputs from both hearing aids are given by:

$$\mathbf{Y}_{\mathbf{L1\_high}} = \mathbf{W}_{\mathbf{high}}\mathbf{X}_{\mathbf{L1\_high}} \tag{6.12}$$

$$\mathbf{Y}_{\mathbf{R1\_high}} = \mathbf{W}_{\mathbf{high}}\mathbf{X}_{\mathbf{R1\_high}} \tag{6.13}$$

Therefore, the spatial impression cues from the focused and interferer sides are preserved since the gain is applied to the original microphone signals on either side of the head. When the target speaker is on the right side of the hearing aid user, this now becomes the focus side and the Wiener filter gain can be derived as in Equation (6.11) and is used to obtain to obtain the enhanced desired speech signal as in Equations (6.12) and (6.13).

6.3.2 Low Frequency Side Look

At low frequencies, the signal's wavelength is long compared to the distance l_{head} across the head between the two hearing aids. Therefore, the spatial aliasing effects which were discussed in Section 5.3 are not significant. Assuming l_{head}=17 cm, the maximum acoustic frequency $f_{spatial}$ to avoid spatial aliasing is approximately 1 kHz as given by:

$$f_{spatial} = \frac{v}{2l_{head}} \tag{6.14}$$

The proposed system for the low frequency side look is designed using the first-order ADMA (Figure 6.2(b)) across the head which is described below.

6.3.2.1 Binaural First Order ADMA

As before, the left side is assumed to be the focused side of the user and the right side is the interferer side. Therefore, a system is designed which performs directional signal processing to steer to the side of interest. Following section 6.2, consider the left ear signal $x_{L1}[n]$ and the right ear signal $x_{R1}[n]$. A binaural first order ADMA is implemented along the microphone sensor axis across the head pointing to θ_{steer} = -90°. Two back-to-back cardioids are thus resolved setting the delay T to l_{head}/v. Following Equation (6.3), the frequency domain subband array outputs are $Y_{L1\_low,k}=Y_{R1\_low,k}$, which are scalar combinations of a forward facing cardioid $C_{F,k}$ (pointing to -90°) and a backward facing cardioid $C_{B,k}$ (pointing to 90°). Correspondingly, when the focus side is the right side, the array outputs combine a forward facing cardioid $C_{F,k}$ (pointing to 90°) and a backward facing cardioid $C_{B,k}$ (pointing to -90°).

6.4 Steerable Binaural Beamformer

Section 6.2.1 discussed the method for steering the look direction to 0° and 180°. Section 6.3 covered the side-look steering system for focusing on target speakers at $\pm\ 90^{\circ}$. The main goal of the steerable system described in this section is to achieve the look direction θ_{steer} to one of the remaining four directions $\pm\ 45^{\circ}$ and $\pm\ 135^{\circ}$. First, the proposed parametric model used for achieving these desired look directions is presented. This model is then used to derive an estimate of the short-time power of the desired signal and an estimate of the short-time power of the interfering signal for enhancing the input noisy signal.

6.4.1 Parametric Steering Model

The input noisy speech signals $x_{L1}[n]$ and $x_{R1}[n]$ comprise the desired speech signal $s[n]$ from direction θ_{steer} and the noise signals $d[n]$ incident from different directions. The sub-band desired signal S_k with incident angle θ_{steer} and the interfering signal D_k are estimated in the frequency domain by a linear combination of directional signals, as explained in the next section. These directional signals used in this estimation are derived using the system illustrated in Figure 6.5.

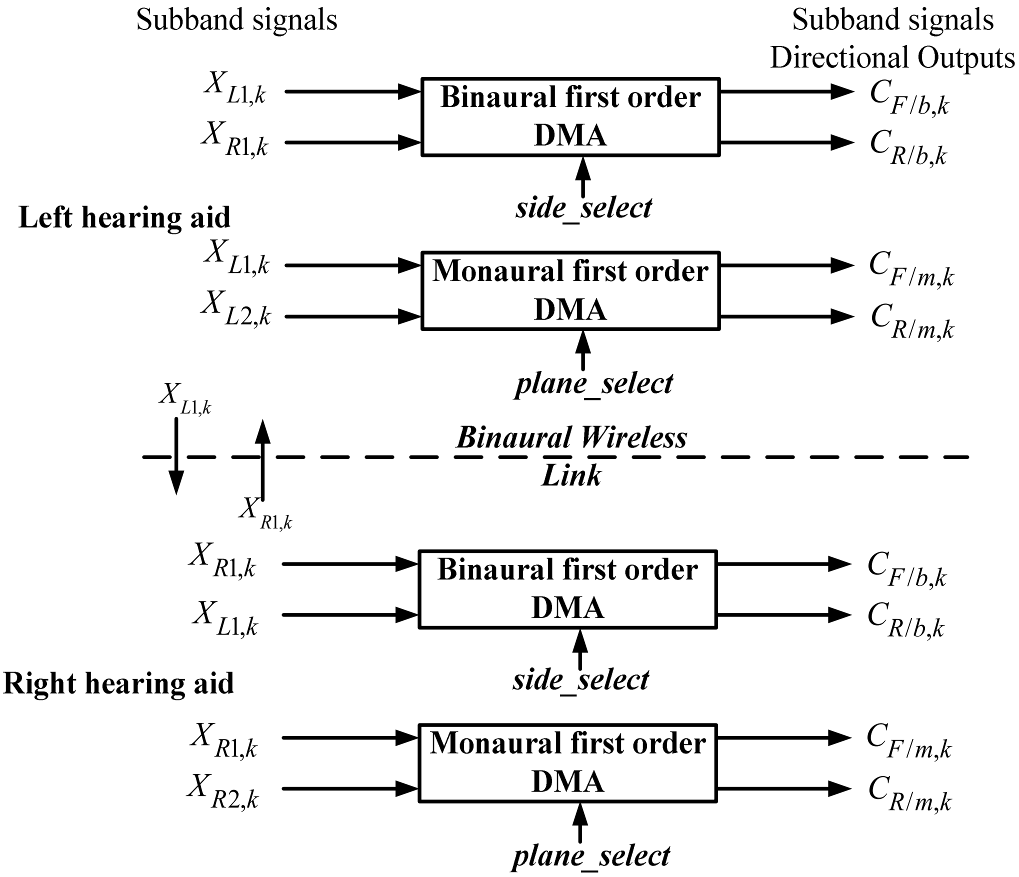


Figure 6.5: System diagram illustrating output directional signals for the signal and the noise estimation in the steerable binaural beamformer

In Figure 6.5, the outputs of the system $C_{F/b,k}$ and $C_{R/b,k}$ result from the binaural first order DMA and respectively denote the forward facing and backward facing cardioids. $C_{F/m,k}$ and $C_{R/m,k}$ result from the monaural first order DMA. These follow the same naming convention as in the binaural case. The parameters "*side*_select" and "*plane_select*" assume values in the range [0,1]. As indicated in Figure 6.5, the parameter "*side_select*" selects which microphone signal from the binaural array is delayed and subtracted, and therefore is used to select the direction to which $C_{F/b,k}$ and $C_{R/b,k}$ point. The parameter "*plane_select*" shown in Figure 6.5 is used to select which microphone signal from the monaural array is delayed and subtracted, and is used to select the direction to which $C_{F/m,k}$ and $C_{R/m,k}$ point. For both the monaural and the binaural cases, the directions to which these forward and backward facing cardioids point are summarized in Table 6.1, using different values for "*side_select*" and "*plane_select*".

side_select	$C_{F/b,k}$	$C_{R/b,k}$
1	Points to the right (90°)	Points to the left (-90°)
0	Points to the left (-90°)	Points to the right (90°)

(a)

plane_select	$C_{F/m,k}$	$C_{R/m,k}$
1	Points to the front (0°)	Points to the back (180°)
0	Points to the back (180°)	Points to the front (0°)

(b)

Table 6.1: Differential microphone array outputs for (a) binaural and (b) monaural cases

As detailed in Section 6.3.2, the binaural directional signals can be obtained for the low frequencies to avoid spatial aliasing effects. Thus, the directional processing system presented in this section is applicable for the sub-band signals corresponding to these low frequencies (<1 kHz). For conciseness, the sub-band index k will be omitted in the following analysis. It must be noted that in our approach in the following section, we will select fixed values for the steering factors presented.

6.4.2 Signal and Noise Estimation

As mentioned in the previous section, the signal and the noise estimates are derived from the combination of the directional signals using the parametric model. Table 6.2 indicates the parameter values for "*side_select*" and "*plane_select*" used to obtain these directional outputs needed to steer to a target at an angle θ_{steer}.

θ_{steer} (°)	***side_select***	***plane_select***
45	1	1
135	1	0
-135	0	0
-45	0	1

Table 6.2: Parametric values required for the steerable system to focus on targets from the desired angles θ_{steer}

As an example, let us consider a situation where the desired speaker $s[n]$ is assumed to be at azimuth $\theta_{steer} = 45°$. Since the direction of the desired signal θ_{steer} is known, an estimate of the desired signal power can be obtained from measuring the minimum of the power obtained from the directional outputs which mutually have maximum response in the direction of the signal. For this orientation, the parameters "*side_select*" and "*plane_select*" are both set to 1 from Table 6.2 to obtain the binaural and the monaural outputs as indicated in Table 6.1(a) and Table 6.1(b) respectively. From the frequency domain sub-band signals, hypercardioids (Gay and Benesty, 2000) Y_1 and Y_2 are obtained which have the maximum directivity index for a first order DMA and signals Y_3 and Y_4 create notches at 90°/-90° and 0°/180° respectively, given by:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} C_{F/m} \\ C_{F/b} \\ C_{F/m} \\ C_{F/b} \end{bmatrix} - \beta_{hyp} \begin{bmatrix} C_{R/m} \\ C_{R/b} \\ C_{R/m} / \beta_{hyp} \\ C_{R/b} / \beta_{hyp} \end{bmatrix} \tag{6.15}$$

where the steering parameter β_{hyp} ($\simeq 0.5$) is set to a value to create the desired hypercardioid which maximizes the directivity index in a diffuse noise field (Gay and Benesty, 2000). Equation (6.15) can be rewritten compactly in matrix form as:

$$\mathbf{Y} = \mathbf{C}_{\mathbf{F,1}} - \beta_{hyp}\, \mathbf{C}_{\mathbf{R,1}} \tag{6.16}$$

where $\mathbf{Y}=[Y_1\ Y_2\ Y_3\ Y_4]^T$, $\mathbf{C}_{\mathbf{F,1}}=[C_{F/m}\ C_{F/b}\ C_{F/m}\ C_{F/b}]^T$ and $\mathbf{C}_{\mathbf{R,1}}=[C_{R/m}\ C_{R/b}\ C_{R/m}/\beta_{hyp}\ C_{R/b}/\beta_{hyp}]^T$. An estimate of the short time desired signal power $\hat{\Phi}_S$ is obtained from measuring the minimum short time power of the four signal components in $\mathbf{Y}$ as:

$$\hat{\Phi}_S = \min\left(\Phi_Y\right) \tag{6.17}$$

The noise estimate is obtained by measuring the maximum power from two directional signals which mutually have a null placed in the direction θ_{steer} of the desired source. From Equation (6.4), the value θ_{null} is assumed to lie in the back half plane (i.e. >±90°). To obtain the noise estimate when the desired source is once again at θ_{steer} = 45°, we use parametric values of "*side_select*"=1 and "*plane_select*"=1 as before. Using the directional signals corresponding to these parameter values, the desired source is now in the relative back half plane. A null is placed in the target's direction by obtaining the value of β which places a null in the direction $\theta_{null} = 135^{\circ}$ using Equation (6.4). Therefore, this will null the desired source at the actual azimuth of 45° using both the monaural and binaural directional outputs. Let $\mathbf{C}_{\mathbf{R},\mathbf{2}}=[C_{R/m}\ C_{R/b}]^T$ and $\mathbf{C}_{\mathbf{F},\mathbf{2}}=[C_{F/m}\ C_{F/b}]^T$. These two signals are used to measure the signal $\mathbf{V}$ which is used for the noise power estimation as given by:

$$\mathbf{V} = \mathbf{C}_{\mathbf{R},2} - \beta_{steer}\ \mathbf{C}_{\mathbf{F},2} \qquad \textbf{(6.18)}$$

where $\mathbf{V}=[V_1\ V_2]^T$ and β_{steer} ($\simeq 0.1716$) is set using Equation (6.4).

An estimate of the short time noise power $\hat{\Phi}_D$ is obtained from the maximum of the short time power of the two noise components in $\mathbf{V}$ as given in:

$$\hat{\Phi}_D = \max\left(\Phi_{\mathbf{V}}\right) \qquad \textbf{(6.19)}$$

The corresponding Wiener filter gain W_s is obtained from:

$$W_{\mathrm{s}} = \frac{\hat{\Phi}_S}{\hat{\Phi}_S + \hat{\Phi}_D} \qquad \textbf{(6.20)}$$

The enhanced desired left and right signals, $\hat{S}_{L1}$ and $\hat{S}_{R1}$ respectively, are obtained by filtering the locally available omni-directional signal and is given by:

$$\hat{S}_{L1} = W_s X_{L1} \qquad \textbf{(6.21)}$$

$$\hat{S}_{R1} = W_s X_{R1} \qquad \textbf{(6.22)}$$

The enhanced signals are then reconstructed using a synthesis filter-bank. Steering to the other directions θ_{steer} of 135°, -135° or -45° is done by setting the parameter values of β_{steer} ($\simeq 0.1716$) using Equation (6.4) as before, and the values of "*side_select*" and *"plane_select"* as given in Table 6.2.

6.5 Performance Evaluation

In this section, the performance of the binaural side-look steering system from Section 6.3 and the steerable beamformer system from Section 6.4 are first evaluated by examining the output directivity patterns. A binaural hearing aid system was set up as described in Section 6.2 with two "Behind the Ear" (BTE) hearing aids on each ear, and only one signal being transmitted from one ear to the other. The monaural microphone spacing l_{mon}=1 cm and the binaural spacing l_{head}=17 cm were used as previously defined with a sampling frequency of 24 kHz. The measured microphone signals were recorded on a KEMAR dummy head in an anechoic room. The beampatterns were obtained by radiating a source signal along a circular path (with increments of 5°) at a constant distance of 1 metre from the dummy's head, as illustrated in Figure 6.6 below.

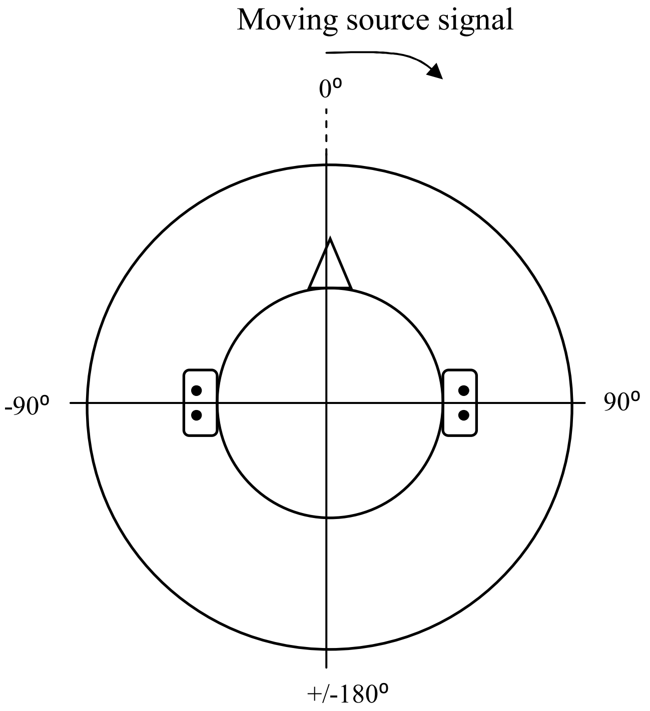


Figure 6.6: Set-up used for measuring directivity patterns using a KEMAR dummy head in an anechoic chamber

The steerable beamformer system was also evaluated in scenarios with a desired speaker at a fixed position and multiple interfering speakers located at different positions. The SIR_{gain} in each case is presented as a measure of the enhancement achieved using this technique, where SIR_{gain}(dB) is given by:

$$SIR_{\text{gain}} = SIR_{\text{out}} - SIR_{\text{in}} \tag{6.23}$$

and SIR_{in}(dB) is the SIR of the input noisy signal and SIR_{out}(dB) is the SIR of the output enhanced signal.

6.5.1 Binaural Side-Look Steering Beamformer

The binaural side-look steering beamformer was decomposed into two subsystems to independently process the low frequencies (<1 kHz) and the high frequencies (≥1 kHz). In this scenario, the desired source is located on the left side of the hearing aid user at -90° and the interferer on the right side of the user at 90°, as illustrated in Figure 6.7.

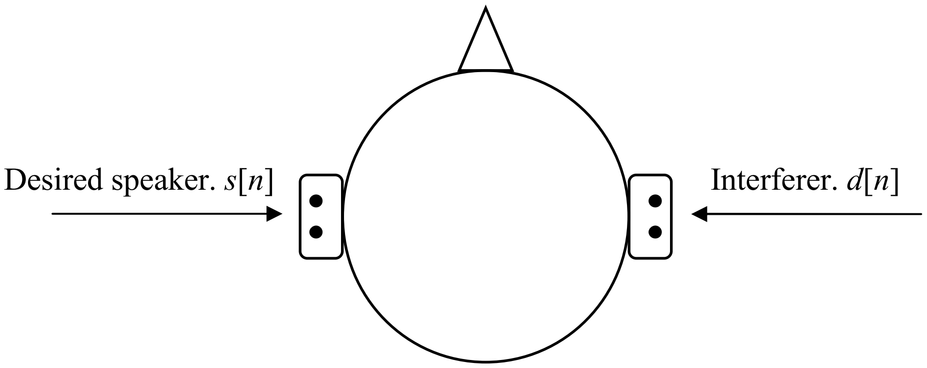
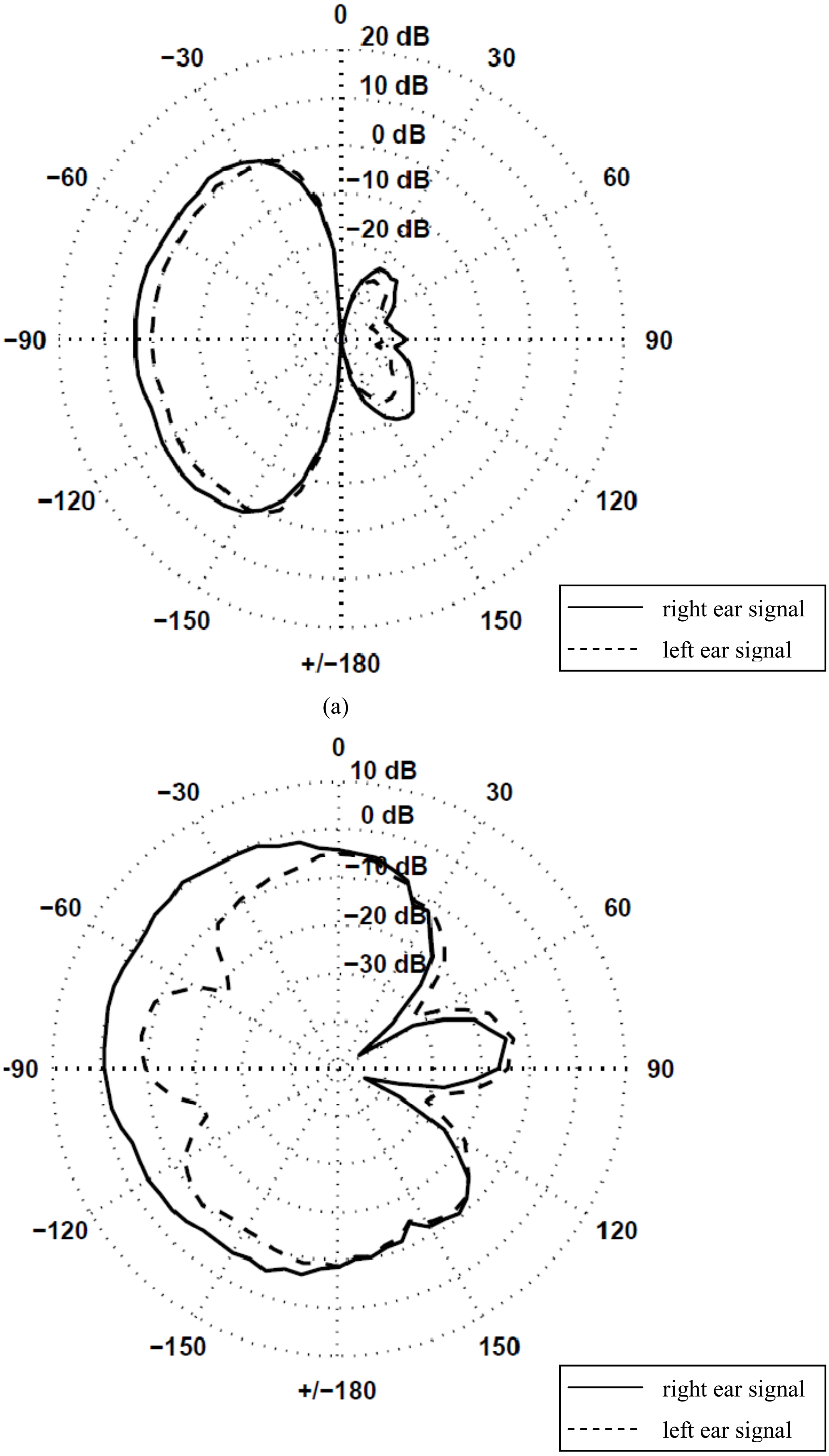


Figure 6.7: Illustration of hearing scenario with a desired speaker (-90°) and a localized interferer (90°)

The effectiveness of these two systems is demonstrated with representative directivity plots at 250 Hz (using the low frequency side-look system) in Figure 6.8(a) and at 2 kHz (using the high frequency side-look system) in Figure 6.8(b).



(a)

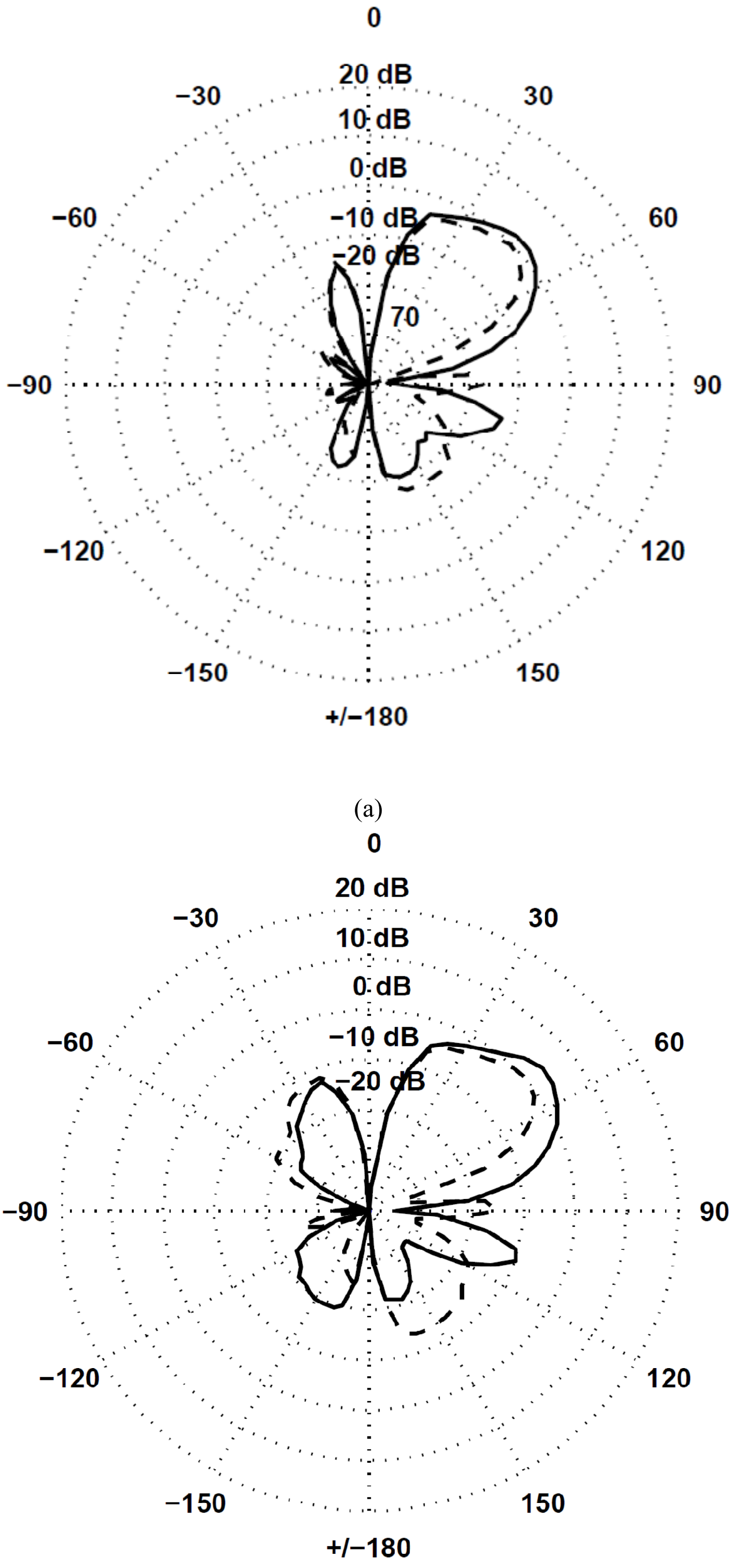
(b)

Figure 6.8: (a) Beam steered to left side at 250 Hz (b) Beam steered to left side at 2 kHz

In both plots, the responses from both ears are shown together to illustrate the desired preservation of the spatial cues. It can be seen that the attenuation is more significant on the interfering signal impinging on the right side of the hearing aid user. Similar frequency responses were obtained across all frequencies for focusing on desired signals located either at the left (-90°) or the right (90°) of the hearing aid user.

6.5.2 Steerable Beamformer: Example 45° steering

The performance of the steerable beamformer is demonstrated for the scenario described in Section 6.4.2 where the desired speaker $s[n]$ is at azimuth θ_{steer} = 45°. This steerable beamformer system operates up to approximately 750 Hz since the estimates were obtained from a combination of monaural and binaural array outputs (limited due to large spacing between hearing aids across the head). From Equations (6.17) and (6.19), estimates of the signal power $\hat{\Phi}_S$ and the noise power $\hat{\Phi}_D$ were obtained. A null is placed at the orientation of the desired speaker to estimate the noise from Equation (6.19). As described in Section 6.4.2, the corresponding value of $\beta \simeq 0.1716$ is calculated from Equation (6.4). The polar plot of the beampattern of the proposed steering system to 45° is shown from the left and right hearing aids at 250 Hz and 500 Hz in Figure 6.9(a) and Figure 6.9(b) respectively. As required, the maximum gain is in the direction of θ_{steer} = 45°.



(a)

(b)

Figure 6.9: (a) Beam steered to 45° at 250 Hz (b) Beam steered to 45° at 500 Hz

Three hearing scenarios were simulated with a desired male speaker at position θ_{steer} = 45° in the presence of localized speech interferers as illustrated in Figure 6.10 . The gender of the interfering speaker is also shown.

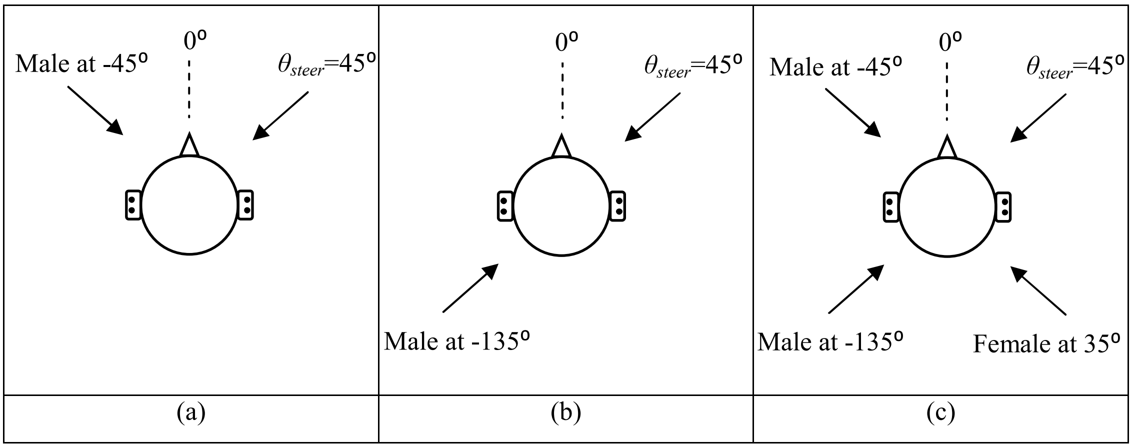


Figure 6.10: Hearing scenarios with desired speaker at θ_{steer} = 45° with localized speech interferers at (a) -45° (b) -135° (c) -45° and -135° and 135°

The SIR_{gain} obtained using this steerable beamformer for the hearing scenarios shown in Figure 6.10, with interferers positioned at directions θ_d, is presented in Table 6.3. In each noisy scenario, the SIR is measured at the right microphone signal and the SIR of the input noisy signal is set to 0 dB. The gender of the interfering speaker is denoted as "m" for male and "f" for female. The SIR_{gain} is measured over the operating frequency range of the beamformer (≤750 Hz). In the evaluated cases from Table 6.3, an SIR gain of approximately 8 dB is obtained.

Position of localized interferer, θ_d(°)	**SIR_{gain}(dB)**
-45 (m)	8.17
-135 (m)	7.94
-45 (m), -135(m), 135 (f)	8.05

Table 6.3: Performance of the steerable system focusing on a desired source at θ_{steer} = 45°, in the presence of localized interferers at azimuths θ_d

These simulations were performed using actual recorded signals. The steering of the beam can be adjusted to the direction θ_{steer} by fine-tuning the ideal values of the steering parameter β_{steer} that were derived in Section 6.4, using Equation (6.4), since they vary with frequency in real implementations.

6.6 Conclusions

State-of-the-art hearing aids are effective at steering the look direction of binaural hearing aids to directions of 0° and 180°. The side-look steering system and the steerable beamformer presented in this chapter were shown to be effective at steering to additional directions using beampatterns. The performance of the steerable beamformer was evaluated by also measuring SIR improvements in noisy speech environments. These presented methods are important as they provide techniques for reducing the noise level of interfering speakers when the desired speaker and the interferers are spatially separated. Although, the hearing aid's size and power consumption constraints impose that each device has two microphones, this result was achieved via the exchange of information between the hearing aids over a wireless link. The binaural side-look system comprises two sub-systems for processing low frequencies and high frequencies independently. For the steerable beamformer, the binaural directional signals were used in the estimation of the short-time power of the speech and the noise. Due to spatial aliasing effects, it was shown that the system operated approximately up to 750 Hz. Future work involves extending this system to operate at high frequencies.

An important consideration in real implementations is the balance between the level of attenuation of the noise sources and the resulting speech distortion in the enhanced signal. This trade-off can be satisfied by limiting the amount of noise reduction applied to the processed signal. The steering systems presented allow the hearing aid to focus on target speakers at the desired eight directions. This was tested in a real-time environment using hearing aids with multiple interferers from different directions and evaluated using informal listening tests to confirm the promising results and performance.

7

Conclusions and Future Work

7.1 Discussion

This thesis investigated methods for performing noise reduction for speech enhancement and Voice Activity Detection (VAD). Speech enhancement and VAD are fundamental to any system which processes speech and audio signals to improve speech intelligibility and quality. In this thesis, novel techniques were proposed to provide improved results compared to existing approaches in these domains.

Some examples of systems that can incorporate our novel solutions are mobile phones, binaural hearing aids, VoIP and voice recognition systems. These applications may employ noise reduction systems to denoise the noisy speech signal. However, low frequency noises may pose problems for noise reduction systems, as they may be unable to track this noise accurately. Our proposed EMDF systems perform improved suppression of this low frequency noise for speech enhancement applications. It was also shown that these EMDF systems are able to achieve higher speech quality output.

VAD techniques are used to identify speech activity and speech pauses, while keeping false alarm rates low. Noise estimation routines may use a VAD to update the noise spectrum estimate during speech pauses. Therefore, high speech pause hit rates are required from the VAD system, especially when tracking highly non-stationary noises. However, if speech pauses are incorrectly identified, then this results in speech leakage into the noise spectrum estimate. Voice transmission devices utilize VAD systems for identify and transmitting speech only segments. This results in efficient usage of the available bandwidth. However, incorrect speech pause identification from the VAD results in speech loss. The improved 1-D LBP system is a novel solution for the VAD problem, which is computationally efficient. This is essential for real-time applications which utilize VAD systems. The new VAD work improves on previously used standards in the communications arena. Therefore, this results in benefits such as improved usage of both the electromagnetic (EM) spectrum and the limited bandwidth available to many voice transmission devices.

In some applications, a remotely obtained noise reference is available and ANC techniques are commonly used to perform speech enhancement. The proposed SEAEMD technique was shown to provide improved speech quality with lower levels of residual noise.

Binaural hearing aids are increasing in popularity. Due to size and power constraints, two microphones are commonly used in each device. In this thesis, novel noise reduction techniques for binaural hearing aids have been presented for focusing on speakers located at other directions than 0° and 180°. This system was shown to be computationally efficient and was able to satisfy the constraints of two microphones in each device, as opposed to previously proposed techniques. This system was tested in a real-time environment using actual hearing aids and multiple interferers.

7.2 Future Work

It is obvious that the areas covered in this thesis have a great deal of potential for future work. The algorithm presented on the SEAEMD is not limited to speech and the principles may be applied and integrated into a wide range of solutions. It also presents exciting possibilities for 2-D signals.

Implementations of ANC systems may use a VAD, since the noise reference may contain a portion of the desired speech signal. Therefore, the 1-D LBP VAD system may be incorporated in the dual-microphone noise reduction systems, so that the adaptive filters are only updated during speech pauses. This prevents cancellation of the speech when leakage into the noise reference occurs. It also allows the filter to only adapt to the noise.

CASA has been reviewed as a method for separating a desired speech signal from a noisy signal, where the noise source may be due to interfering speakers or background noises. It must be recalled that it was proposed that the auditory system groups sounds that share common acoustic cues to a specific source, whereas sounds which differ likely originated from different sources. Next, auditory features are extracted from the voiced and the unvoiced speech components. Analysis of sounds with common onset and offset times may be used as a cue for both voiced and unvoiced speech components. In (Akbari and Soraghan, 2003), a fuzzy-based multi-scale edge detection was proposed to perform optimal edge detection. Edge-detection methods such as this present possibilities for adapting this technique to the onset/offset identification required in CASA systems, for extracting auditory cues.

Fourier transform based noise estimation techniques for single-microphone noise reduction for speech enhancement have been covered. These noise spectrum estimates can be used to obtain the speech estimate magnitude. However, during signal reconstruction, the noisy signal's phase is observable, and therefore is used in the inverse transform. Signal reconstruction without phase or a phase estimate has been covered in work such as (Balan et al., 2006), and this is an exciting area for future work in speech enhancement methods.

Appendix A

Degenerate Unmixing Estimation Technique Algorithm

The Degenerate Unmixing Estimation Technique (DUET) for BSS is summarized below for separating *M* sources from *N*=2 mixtures:

1. Perform the STFT of the two mixtures from Equations (5.4) and (5.5) to obtain $X_1(\Omega,\tau)$ and $X_2(\Omega,\tau)$:

$$X_1(\Omega,\tau)=\sum_{i=1}^{M}S_i(\Omega,\tau) \qquad \textbf{(A.1)}$$

$$X_2(\Omega,\tau)=\sum_{i=1}^{M}a_i e^{-j\Omega\Delta_i}S_i(\Omega,\tau) \qquad \textbf{(A.2)}$$

   where *j*=√-1

2. Calculate the local symmetric attenuation $\tilde{\alpha}(\Omega,\tau)$ by:

$$\tilde{\alpha}(\Omega,\tau):=\left|\frac{X_2(\Omega,\tau)}{X_1(\Omega,\tau)}\right|-\left|\frac{X_1(\Omega,\tau)}{X_2(\Omega,\tau)}\right| \qquad \textbf{(A.3)}$$

3. Calculate the local arrival delay $\tilde{\Delta}(\Omega,\tau)$ by:

$$\tilde{\Delta}(\Omega,\tau):=-\frac{1}{\Omega}imag\left(\ln\left(\frac{X_2(\Omega,\tau)}{X_1(\Omega,\tau)}\right)\right) \qquad \textbf{(A.4)}$$

4. Construct the two-dimensional histogram using the values of attenuation $\tilde{\alpha}(\Omega,\tau)$ and delay $\tilde{\Delta}(\Omega,\tau)$ (Equations (40) and (42) from (Yilmaz and Rickard, 2004))

5. Identify the *M* peaks from the histogram to approximate the mixing parameters $\hat{a}_i$ and $\hat{\Delta}_i$ corresponding to the i^{th} source
6. Construct the binary time-frequency masks, $B_i(\Omega,\tau)$ for the *M* sources, using the time-frequency pairs ($\hat{a}_i , \hat{\Delta}_i$) (Equations (43) and (44) from (Yilmaz and Rickard, 2004)). This binary mask $B_i(\Omega,\tau)$=1 when the i^{th} source is active and is set to 0 otherwise
7. Perform demixing to obtain the source signals, $\hat{S}_i(\Omega,\tau)$ by applying the time-frequency mask $B_i(\Omega,\tau)$ to the mixture

Author Publications

- European patent on a system for a direction dependent steerable binaural beamformer for noise reduction in hearing aids - Reference numbers 2010E02368 DE and EP10154098.7

- N. Chatlani, Eghart Fischer, J. J. Soraghan, “Direction Dependent Spatial Noise Filtering in Binaural Hearing Aids” (under review for Signal Processing)

- N. Chatlani, J. J. Soraghan, “EMD based Filtering (EMDF) for Speech Enhancement” (under review for IEEE Transactions on Audio, Speech and Language Processing)

- N. Chatlani, J. J. Soraghan, “Local Binary Patterns for 1-D Signal Processing”, in 18th European Signal Proc. Conference (EUSIPCO), Aug 2010, Aalborg, Denmark

- N. Chatlani, Eghart Fischer, J. J. Soraghan, “Spatial Noise Reduction in Binaural Hearing Aids”, in 18th European Signal Proc. Conference (EUSIPCO), Aug 2010, Aalborg, Denmark

- N. Chatlani, J. J. Soraghan, “EMD-based Noise Estimation and Tracking (ENET) with application to speech enhancement”, in 17th European Signal Proc. Conference (EUSIPCO), Aug 2009, Glasgow, UK

- N. Chatlani, J. J. Soraghan, “Advanced Signal Processing for Hearing and Hearing Defects”, PhD Workshop for Hearing Scientists 2009, Nottingham, UK

- N. Chatlani, J. J. Soraghan, “Speech Enhancement using Adaptive Empirical Mode Decomposition”, in 16th Int’l Conference on Digital Sig. Proc (DSP 2009), July 2009, Santorini, Greece

- N. Chatlani, J. J. Soraghan, "Adaptive Empirical Mode Decomposition for Signal Enhancement with application to speech," 15th Int'l Conference on Systems, Signals and Image Processing 2008, pp. 101-104, June 2008.

References

AKBARI, A. S. & SORAGHAN, J. J. Fuzzy-based multiscale edge detection. *IET Electronic Letters,* vol. 39:1**,** 30-32, Jan. 2003.

ARONS, B. *A Review of the Cocktail Party Effect* [Online]. Available: http://www.media.mit.edu/speech/papers/1992/arons_AVIOSJ92_cocktail_party_effect.pdf [Accessed October 2010].

BALAN, R., CASAZZA, P. & EDIDIN, D. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis,* vol. 20:3**,** 345-356, 2006.

BAUML, R. & SOERGEL, W. Uniform polyphase filter banks for use in hearing aids: Design and Constraints. *In:* 16th European Signal Processing Conference (EUSIPCO). Aug. 2008.

BENESTY, J., MAKINO, S. & CHEN, J. 2005. *Speech Enhancement*, Springer.

BEROUTI, M., SCHWARTZ, M. & MAKHOUL, J. Enhancement of Speech corrupted by acoustic noise. *In:* IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP). 208-211, 1979.

BHUNJUN, V., BROOKES, M. & WEN, J. Y. C. Eigendomain-based Noise Estimation with the Minimum Statistics Approach. *In:* International Workshop on Acoustic and Echo Noise Control (IWAENC), Paris. Sep. 12-14 2006.

BITZER, J. & SIMMER, K. U. 2001. Superdirective microphone arrays. *In:* BRANDSTEIN, M. S. & WARD, D. B. (eds.) *Microphone Arrays: Signal Processing Techniques and Applications.* Springer, Berlin, Germany.

BREGMAN, A. S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*, Cambridge, Mass.: Bradford Books, MIT Press.

BROWN, G. J. & COOKE, M. Computational auditory scene analysis. *Computer Speech & Language,* vol. 8**,** 297-336, 1994.

CARDOSO, J.-F. Blind Signal Separation: Statistical Principles. *Proc. of the IEEE,* vol. 86:10**,** 2009-2025, Oct. 1998 1998.

CHERRY, E. C. Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustic Society of America,* vol. 25**,** 975-979, 1953.

COHEN, I. Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging. *IEEE Transactions on Speech and Audio Processing,* vol. 11:5, Sep. 2003.

COHEN, I. & BERDUGO, B. Speech enhancement for non-stationary noise environments. *Signal Processing, Elsevier,* vol. 81**,** 2403-2418, Nov. 2001.

COOKE, M. P. 1993. *Modeling Auditory Processing and Organization*, Cambridge, UK: Cambridge Univ. Press.

CORNELIS, B., DOCLO, S., BOGAERT, T. V. D., MOONEN, M. & WOUTERS, J. Theoretical Analysis of Binaural Multi-Microphone Noise Reduction techniques. *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18:2**,** 342-355, Feb 2010.

DAOUD, D., KALLEL, F., GHORBEL, M. & HAMIDA, B. Spatial Filtering based Speech Enhancement for Binaural Hearing Aid. *In:* 6th International Multi-Conference on Systems, Signals and Devices. 2009.

DARWIN, C. J. Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 363:1493**,** 1011-1021, 2008.

DOBLINGER, G. Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands. *In:* Proceedings Eurospeech 2. 1513-1516, 1995.

DOCLO, S., MOONEN, M., BOGAERT, T. V. D. & WOUTERS, J. Reduced-Bandwidth and Distributed MWF-Based Noise Reduction Algorithms for Binaural Hearing Aids. *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17**,** 38-51, Jan 2009.

EICHLER, M. & LACROIX, A. Broadband Superdirective Beamforming using Multipole Superposition. *In:* 16th European Signal Processing Conference (EUSIPCO). Aug. 2008.

EPHRAIM, Y. & COHEN, I. 2006. Recent Advances on Speech Enhancement. *The Electrical Engineering Handbook.* CRC Press.

EPHRAIM, Y. & MALAH, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 32:6**,** 1109 - 1121 Dec. 1984.

EPHRAIM, Y. & MALAH, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 33:2**,** 443-445, Apr. 1985.

ETSI 1996. Digital cellular telecommunications system;Voice Activity Detection (VAD) for EnhancedFull Rate (EFR) speech traffic channels (GSM 06.82). *GSM: Global System for Mobile Communications.*

FLANDRIN, P., GONCALVES, P. & RILLING, G. Detrending and Denoising with Empirical Mode Decompositions. *In:* European Signal Processing Conference (EUSIPCO). 1581-1584, 2004.

FLANDRIN, P. & RILLING, G. Empirical Mode Decomposition as a Filter Bank. *IEEE Signal Processing Letters,* vol. 11**,** 112-114, Feb. 2004.

FUKUDA, T., ICHIKAWA, O. & NISHIMURA, M. Improved voice activity detection using static harmonic features. *In:* IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, TX. 4482 - 4485 March 2010.

GAY, S. L. & BENESTY, J. 2000. *Acoustic Signal Processing For Telecommunication*.

GAZOR, S. & ZHANG, W. A soft voice activity detector based on a Laplacian-Gaussian model. *IEEE Transactions on Speech and Audio Processing,* vol. 11:5**,** 498 - 505 Sep. 2003.

GERKMANN, T., BREITHAUPT, C. & MARTIN, R. Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio With Fixed Priors. *IEEE Transactions on Audio, Speech and Language Processing,* vol. 16:5**,** 910-919, 2008.

HAMACHER, V., CHALUPPER, J., EGGERS, J., FISCHER, E., KORNAGEL, U., PUDER, H. & RASS, U. Signal Processing in High End Hearing Aids: State of the Art, Challenges and Future Trends. *EURASIP Journal on Applied Signal Processing,* vol. 18**,** 2915-2929, 2005.

HASAN, M. K., SALAHUDDIN, S. & KHAN, M. R. A modified a priori SNR for speech enhancement using spectral subtraction rules. *IEEE Signal Processing Letters,* vol. 11:4, 450 - 453 2004.

HASAN, T. & HASAN, M. K. Suppression of Residual Noise From Speech Signals Using Empirical Mode Decomposition. *IEEE Signal Processing Letters,* vol. 16:1, 2-5, Jan. 2009.

HE, S., SORAGHAN, J. J., O'REILLY, B. F. & XING, D. Quantitative Analysis of Facial Paralysis Using Local Binary Patterns in Biomedical Videos. *IEEE TRansactions on Bioengineering,* vol. 57:7, July 2009.

HENDRIKS, R. C., JENSEN, J. & HEUSDENS, R. Noise Tracking using DFT Domain Subspace Decompositions. *IEEE Transactions on Audio, Speech and Language Processing,* vol. 16:3, 541-553, Mar. 2008.

HU, G. & WANG, D. L. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks,* vol. 15:5, 1135-1150, Sep. 2004.

HU, G. & WANG, D. L. 2006. An auditory scene analysis approach to monaural speech segregation. *Selected methods for acousitic echo and noise control.* Springer.

HU, G. & WANG, D. L. Auditory Segmentation Based on onset and offset analysis. *IEEE Transactions on Audio, Speech and Language Processing,* vol. 15:2, 396-405, Feb. 2007.

HU, Y. & LOIZOU, P. C. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Transactions on Audio, Speech and Language Processing,* vol. 16:1, Jan. 2008.

HUANG, N. E., SHEN, Z., LONG, S., WU, M. C., SHIH, H., ZHENG, Q., YEN, N., TUNG, C. & LIU, H. The Empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A,* vol. 454:1971, 903-995, 1998.

IHLE, M. Differential Microphone Arrays for Spectral Subtraction. *In:* Int'l Workshop on Acoustic Echo and Noise Control (IWAENC 2003). Sep. 2003.

ITU-T 1996. A Silence Compression Scheme for G.729 Optimized for Terminals conforming to ITU-T V.70 Annex B.

JIN, Z. & WANG, D. L. A supervised learning approach to monaural segregation of reverberant speech. *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 17:4**,** 625-638, May 2009.

JOHN S. GAROFOLO, LAMEL, L. F., FISHER, W. M. & FISCUS, J. G. 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia.

JOHNSON, D. H. & SINANOVIC, S. 2001. Symmetrizing the Kullback–Leibler Distance. *Technical Report.* Rice University.

JOURJINE, A., RICKARD, S. & YıLMAZ, O. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. *In:* Proc. IEEE International Conference on Acoustical, Speech, and Signal Processing, Turkey. 2985–2988, 2000.

JUNQUA, J. C., REAVES, B. & MAK, B. A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognizers. *In:* Eurospeech. 1371-1374, 1991.

KAMKAR-PARSI, A. H. & BOUCHARD, M. Improved Noise Power Spectrum Density Estimation for Binaural Hearing Aids Operating in a Diffuse Noise Field Environment. *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17:4**,** 521-533, May 2009.

KHALDI, K., BOUDRAA, A. O., BOUCHIKHI, A. & ALOUANE, M. T.-H. Speech Enhancement via EMD. *EURASIP Journal on Advances in Signal Processing,* vol. 2008**,** 8, 2008.

KIM, J.-S. & PARK, H.-M. Target speech enhancement based on degenerate unmixing and estimation technique for real-world applications. *Electronics Letters* vol. 46:3**,** 259-260, Feb. 2010.

KINGSBURY, N. G. Complex wavelets for shift invariant analysis and filtering of signals *Journal of Applied and Computational Harmonic Analysis,* vol. 10:3**,** 234-253, May 2001.

KITAOKA, N., YAMAMOTO, K., KUSAMIZU, T., NAKAGAWA, S., YAMADA, T., TSUGE, S., MIYAJIMA, C., NISHIURA, T., NAKAYAMA, M., DENDA, Y., FUJIMOTO, M., TAKIGUCHI, T., TAMURA, S., KUROIWA, S., TAKEDA, K. & NAKAMURA, S. Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance. *In:* IEEE Workshop on Automatic Speech Recognition & Understanding, 2007 (ASRU), Kyoto Dec. 2007.

KJEMS, U., PEDERSEN, M. S., BOLDT, J., LUNNER, T. & WANG, D. Speech intelligibility of ideal binary masked mixtures. *In:* European Signal Processing Conference (EUSIPCO), Aalborg, Denmark. 2010.

KLATT, D. Prediction in perceived phonetic distance from critical band spectra. *In:* IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP). 1982.

KOPSINIS, Y. & MCLAUGHLIN, S. Development of EMD-Based Denoising Methods inspired by Wavelet Thresholding. *IEEE Transactions on Signal Processing,* vol. 57:4, 2009.

LEI, S. & TUNG, Y. Wavelet-Based Speech Enhancement using Time-Adapted Noise Estimation. *IECE Transactions on Fundamentals of Electronics, Communications and Computer Sciences,* vol. 9**,** 2555-2563, Sep. 2008.

LI, K., SWAMY, M. N. S. & AHMAD, M. O. An Improved Voice Activity Detection Using Higher Order Statistics. *IEEE Transactions on Speech and Audio Processing,* vol. 13:5**,** 965-974, Sep. 2005.

LI, Y. & WANG, D. On the optimality of ideal binary time–frequency masks. *Speech Communication,* vol. 51:3**,** 230-239, 2009.

LOIZOU, P. 2007a. Noise Estimation Algorithms. *In:* PRESS, C. (ed.) *Speech Enhancement: Theory and Practice.* FL.

LOIZOU, P. 2007b. Noise Estimation Algorithms. *In:* PRESS, C. (ed.) *Speech Enhancement: Theory and Practice.* FL.

LOIZOU, P. 2007c. *Speech Enhancement: Theory and Practice,* FL.

LOIZOU, P. 2007d. Subspace Algorithms. *In:* PRESS, C. (ed.) *Speech Enhancement: Theory and Practice.* FL.

LOIZOU, P., COHEN, I., GANNOT, S. & PALIWAL, K. Special issue on Speech Enhancement. *Speech Communication,* vol. 49:7-8**,** 527-529, 2007.

LOTTER, T. & VARY, P. Dual-Channel Speech Enhancement by Superdirective Beamforming. *EURASIP Journal on Applied Signal Processing,* vol. 2006**,** 1-14, 2006.

MAKINO, S., LEE, T. W. & SAWADA, H. 2007a. *Blind Speech Separation, 2007*, Springer.

MAKINO, S., LEE, T. W. & SAWADA, H. 2007b. K-Means Based Underdetermined Blind Speech Separation. *Signals and Communications Technology.* Springer.

MARTIN, R. Noise PSD Estimation based on Optimal Smoothing and Minimum Statistics. *IEEE Transactions on Speech and Audio Processing,* vol. 9:5, Jul. 2001.

MARZINZIK, M. & KOLLMEIER, B. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing,* vol. 10:2**,** 109-118, February 2002.

MELIA, T. *Underdetermined Blind Source Separation in Echoic Environments Using Linear Arrays and Sparse Representations (PhD thesis).* University College Dublin, Ireland, 2007.

MOLLA, M. K. I. & HIROSE, K. Single-Mixture Audio Source Separation by Subspace Decomposition of Hilbert Spectrum. *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15:3**,** 893-900, 2007.

MOON, T. K. & STIRLING, W. C. 1999. *Mathematical Methods and Algorithms for Signal Processing.*

NEMER, E., GOUBRAN, R. & MAHMOUD, S. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing,* vol. 9:3**,** 217-231, Mar. 2001.

NESTA, F., SVAIZER, P. & OMOLOGO, M. 2008a. A BSS Method for Short Utterances by a recursive solution to the Permutation Problem. *5th IEEE Sensor Array and Multichannel Signal Processing Workshop* Darmstadt.

NESTA, F., SVAIZER, P. & OMOLOGO, M. 2008b. Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS. *IEEE Workshop on Machine Learning for Signal Processing.*

OGAWA, T., HOSOYA, K. & KOBAYASHI, T. Ears of the Robot: Noise Reduction using Four-Line Ultra-Micro Omni-Directional Microphones mounted on a Robot Head. *In:* 16th European Signal Processing Conference (EUSIPCO). Aug. 2008.

OJALA, T. & PIETIKAINEN, M. Unsupervised texture segmentation using feature distributions. *Pattern Recognition,* vol. 32**,** 477-486, 1999.

OJALA, T., PIETIKAINEN, M. & MAENPAA, T. Multiresolution Gray Scale and Rotation Invariant Texture Analysis with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24:7**,** 971-987, 2002.

OPPENHEIM, A. V. & SCHAFER, R. W. 1989. *Discrete Time Signal Processing* Prentice Hall.

PEDERSEN, M. S., LARSEN, J., KJEMS, U. & PARRA, L. C. 2007. *A Survey of Convolutive Blind Source Separation Methods*, Springer Handbook on Speech Processing and Speech Communication.

PUDER, H. Adaptive signal processing for interference cancellation in hearing aids. *Signal Processing,* vol. 86:6**,** 1239-1253, 2006.

RAMÍREZ, J., GÓRRIZ, J. M. & SEGURA, J. C. 2007. Voice Activity Detection. Fundamentals and Speech Recognition System Robustnessness. *In:* GRIMM, M. & KROSCHEL, K. (eds.) *Robust Speech Recognition and Understanding* Vienna, Austria: I-Tech Education and Publishing.

RAMÍREZ, J., SEGURA, J. C., BENÍTEZ, C., TORRE, Á. D. L. & RUBIO, A. An Effective Subband OSF-Based VAD With Noise Reduction for speech recognition. *IEEE Transactions on Speech and Audio Processing,* vol. 13:6**,** 1119-1129, Nov. 2005.

RANGACHARI, S. & LOIZOU, P. A noise-estimation algorithm for highly non-stationary environments. *Speech Communication,* vol. 48:2**,** 220-231, Feb. 2006.

RICKARD, S., MELIA, T. & FEARON, C. DESPRIT – Histogram based blind source separation method of more sources than sensors using subspace methods. *In:* IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Oct. 2005.

RILLING, G., FLANDRIN, P. & GONCALVES, P. On Empirical Mode Decomposition and its Algorithms. *IEEE-EURASIP Workshop NSIP,* vol., Jun. 8-11 2003.

RIX, A., BEERENDS, J., HOLLIER, M. & HEKSTRA, A. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. *In:* IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP). 749-752, May 2001.

ROHDENBURG, T., HOHMANN, V. & KOLLMEIER, B. Robustness Analysis of Binaural Hearing Aid Beamformer Algorithms by Means of Objective Perceptual Quality Measures. *In:* 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 315-318, Oct. 2007.

ROY, O. & VETTERLI, M. Rate-Constrained Beamforming for Collaborating Hearing Aids. *In:* Proc. of IEEE International Symposium of Information Theory (ISIT). 2809-2813, July 2006.

ROY, R. & KAILATH, T. ESPRIT - Estimation of Signal Parameters via Rotational Invariance Techniques. *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 37**,** 984-995, July 1989.

SELESNICK, I. W., BARANIUK, R. G. & KINGSBURY, N. G. The Dual-Tree Complex Wavelet Transform. *IEEE Signal Processing Magazine.* vol. 22:6**,** 123-151, Nov. 2005.

SHAO, Y., SRINIVASAN, S., JIN, Z. & WANG, D. A computational auditory scene analysis system for speech segregation and robust speech recognition. *Computer Speech & Language,* vol. 24:1**,** 77-93, 2010.

SMITH, J. O. 2008. *Spectral Audio Signal Processing* [Online]. Available: http://ccrma.stanford.edu/~jos/sasp [Accessed Dec. 2010 2010].

SOHN, J., KIM, N. S. & SUNG, W. Statistical Model Based Voice Activity Detection. *IEEE Signal Processing Letters,* vol. 6:1**,** 1-3, January 1999.

SOHN, J. & SUNG, W. A Voice Activity Detector employing Soft Decision Based Noise Spectrum Adaptation. *In:* IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), USA. 365-368, May 1998.

SOON, I. Y. & KOH, S. N. Low Distortion Speech Enhancement. *IEE Proceedings Vision, Image and Signal Processing,* vol. 147:3**,** 247 - 253 2000.

SRINIVASAN, S. & DEN BRINKER, A. C. Analyzing Rate-Constrained beamforming schemes in wireless binaural hearing aids. *In:* 17th European Signal Processing Conference (EUSIPCO). Aug 2009a.

SRINIVASAN, S. & DEN BRINKER, A. C. Rate-Constrained Beamforming in Binaural Hearing Aids. *EURASIP Journal on Advances in Signal Processing,* vol. 2009**,** 1-10, 2009b.

SRIPATHI, D. 2003. *The Discrete Wavelet Transform* [Online]. Available: http://etd.lib.fsu.edu/theses/available/etd-11242003-185039/unrestricted/09_ds_chapter2.pdf [Accessed Dec. 2010 2010].

TAHMASBI, R. & REZAEI, S. A Soft Voice Activity Detection Using GARCH Filter and Variance Gamma Distribution. *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 15:4**,** 1129-1134, May 2007.

TANYER, S. G. & ÖZER, H. Voice Activity Detection in Nonstationary Noise. *IEEE Transactions on Speech and Audio Processing,* vol. 8:4**,** 478-482, July 2000.

TEUTSCH, H. & ELKO, G. First and Second Order Differential Microphone Arrays. *In:* 7th International Workshop on Acoustic Echo and Noise Control (IWAENC 2001). 35-38, Sep. 2001.

THERRIEN, C. W. 1992. *Discrete Random Signals and Statistical Signal Processing*, Prentice Hall.

TUCKER, R. Voice activity detection using a periodicity measure. *In:* Communications, Speech and Vision, IEE Proceedings I 377-380, Aug. 1992.

VALIMAKI, V. Simple Design of Fractional Delay Allpass Filters. *In:* European Signal Processing Conference (EUSIPCO). Sep 2000.

VARGA, A. & STEENEKEN, H. J. M. 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems.

VEEN, B. D. V. & BUCKLEY, K. M. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*. vol. 2:5**,** 4-24, April 1998.

VLAJ, D., KOS, M., GRASIC, M. & KACIC, Z. Influence of Hangover and Hangbefore Criteria on Automatic Speech Recognition. *In:* 16th International Conference on Systems, Signals and Image Processing (IWSSIP), Chalkida June 2009.

WANG, D. L. On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis. *In:* Speech Separation by Humans and Machines, Boston, MA. 181–197, 2005.

WANG, D. L. Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in Amplification,* vol. 12:4, 332-353, Dec. 2008.

WEXLER, J. & RAZ, S. Discrete Gabor expansions. *Signal Processing 21,* vol. 3, 207-220, November 1990.

WIDROW, B. & LUO, F.-L. Microphone Arrays for Hearing Aids: An Overview. *Speech Communication,* vol. 39, 139-146, 2003.

WIDROW, B. & STEARNS, S. D. 1985. *Adaptive Signal Processing*, Prentice Hall.

WU, Z. & HUANG, N. E. A study of the characteristics of white noise using the empirical mode decomposition method. *Proceedings Royal Society London A,* vol. 460, 1597-1611, Jun. 2004.

YILMAZ, O. & RICKARD, S. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing,* vol. 52:7, July 2004.

YU, T. & HANSEN, J. An Efficient Microphone Array based Voice Activity Detector for driver's speech in noise and music rich in-vehicle environments. *In:* IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, TX. Mar. 2010.

ZOU, X., LI, X. & ZHANG, R. Speech Enhancement Based on Hilbert-Huang Transform Theory. *In:* IEEE CS Proceeding of the First International Multi-Symposium of Computer and Computational Sciences (IMSCCS'06). 208-213, Jun. 20-24 2006.