COMBINED TEMPORAL AND SPECTRAL PROCESSING METHODS FOR SPEECH ENHANCEMENT

P. Krishnamoorthy

COMBINED TEMPORAL AND SPECTRAL PROCESSING METHODS FOR SPEECH ENHANCEMENT

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

P. KRISHNAMOORTHY

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

MAY 2009

Certificate

This is to certify that the thesis entitled "**COMBINED TEMPORAL AND SPECTRAL PROCESSING METHODS FOR SPEECH ENHANCEMENT**", submitted by **P. Krishnamoorthy** (04610206), a research scholar in the *Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

Dr. S. R. Mahadeva Prasanna
Associate Professor
Dept. of Electronics and Communication Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

My dear parents

K. Palanisamy and K. Sundarambal

for their love and support

&

My guide

Dr. S. R. M. Prasanna

for his guidance and inspiration

Acknowledgements

First and foremost, I feel it as a great privilege in expressing my deepest and most sincere gratitude to my supervisor Dr. S. R. M. Prasanna, for his excellent guidance throughout my study. His kindness, dedication, hard work and attention to detail have been a great inspiration to me. My heartfelt thanks to you sir for the unlimited support and patience shown to me. I would particularly like to thank him for all his help in patiently and carefully correcting all my manuscripts. I have no doubts that finishing my degree in a proper and timely manner was impossible without his helps, suggestions and advices.

I am also very thankful to my doctoral committee members Professor P. K. Bora, Dr. A. Mitra and Dr. P. K. Das for sparing their precious time to evaluate the progress of my work. I express my heartfelt thanks to Dr. Rohit Sinha for providing valuable suggestions on this thesis. I am also thankful to Professor S. Dandapat and Dr. A. Permal for the guidance and support given to me in this work.

I would also like to thank the Head of the Department and the other faculty members for their kind help in carrying out this work. I am also grateful to all the members of the research and technical staff of the department without whose help I could not have completed this thesis. My special thanks to L. N. Sharma sir for maintaining an excellent computing facility and various resources useful for the research work.

Thanks go out to all my friends at the Electro Medical and Speech Technology (EMST) Laboratory. They have always been around to provide useful suggestions, companionship and created a peaceful research environment. They all contributed directly or indirectly to this thesis, be it academic help, proofreading and volunteering to be a test subject.

I extremely thankful to Nirmala madam and Jayanna sir for the care shown and the help given to me during my stay at IITG.

I have no words to express my thanks to five most important persons namely, M. Sabarimali Manikandan, D. Senthil Kumar, S. Karthikeyan, C. Shyam Anand and P. Saravanan. My work in this remote place definitely would not be possible without their love and care that helped me to enjoy my new life in this IITG.

My special thanks to D. Govind for his timely help in all respects. Special thanks also go to Pathi sir, Genemala, Ahmad Ali, Sumitra and Padam Priyal for their help and the care shown during my stay.

I thank all my fellow research students and M. Tech students for their cooperation. During these four years at IITG I have had several friends that have helped me in several ways, I would like to say a big thank you to all of them for their friendship and support.

My deepest gratitude goes to my family for their continuous love and support throughout my studies. The opportunities that they have given me and their unlimited sacrifices are the reasons where I am and what I have accomplished so far.

Finally, I believe this research experience will greatly benefit my career in the future.

P. Krishnamoorthy

Abstract

This thesis proposes a combined temporal and spectral processing (TSP) approach for the enhancement of degraded speech. Three major sources of degradation, namely, background noise, reverberation and speech from the competing speakers are considered in this work. Temporal processing refers to the processing of excitation source information in the time domain. This involves identification and enhancement of speech-specific regions. Spectral processing involves estimation and removal of degrading components, and also identification and enhancement of speech-specific spectral components. The temporal processing is based on the fact that the significant excitation of the vocal tract takes place at the instants of glottal closure and onset of events like burst, frication and aspiration. Depending on the nature of degradation, the excitation source will have many other random peaks in addition to the original instants of significant excitation. Temporal processing method identifies the original instants of significant excitation and emphasizes the region around them in the excitation source signal to obtain the enhanced speech. The spectral processing is based on the fact that the spectral values of the degraded speech will have both speech and degrading components. The spectral components of degradation are therefore estimated and removed. Further, there are spectral peaks that are perceptually important which are identified and enhanced. The quality of the speech signal processed by the combined temporal and spectral processing is found to be enhanced better compared to the degraded speech as well as the signals that are processed by the individual temporal and spectral processing methods.

The major contributions of this thesis are as follows:

- Combined TSP method for the enhancement of noisy speech.
- Combined TSP method for the enhancement of reverberant speech.
- Combined TSP method for the enhancement of multi-speaker speech.
- Evaluation of these methods in the speaker recognition task under degraded conditions.

The other contributions of this thesis are as follows:

- A set of speech-specific features to identify the high signal to noise ratio regions of degraded speech.
- A new fine level processing method to identify the instants of significant excitation of noisy speech.
- A new fine level processing method to identify the instants of significant excitation of reverberant speech.
- A method to estimate the pitch of multi-speaker speech using time-delay estimation.

Keywords: Temporal processing, spectral processing, combined temporal and spectral processing, noisy speech, reverberant speech, multi-speaker speech, speaker recognition.

Note: The degraded speech signals and the corresponding enhanced speech signals by the proposed methods are included in the CD-ROM for listening. A detailed description about the sound files is given in the pdf file (readme.pdf) enclosed in the same.

Contents

List of Figures

List of Tables

List of Acronyms

ASA	Auditory Scene Analysis
AR	Auto Regressive
ARMA	Auto Regressive Moving Average
ANN	Artificial Neural Networks
AMT	Auditory Masking Threshold
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
DYPSA	Dynamic Programming Projected Phase-Slope Algorithm
EM	Expectation Maximization
EVD	Eigen Value Decomposition
FET	Frequency to Eigen domain Transformation
FFT	Fast Fourier Transform
FOD	First Order Difference
FOGD	First Order Gaussian Differentiator
FIR	Finite Impulse Response
GC	Glottal Closure
GMM	Gaussian Mixture Models
HE	Hilbert Envelope
HMM	Hidden Markov Models
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform

IFFT	Inverse Fast Fourier transform
KLT	Karhuen-Loeve Transform
LLR	Log-Likelihood Ratio
LP	Linear Prediction
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
LSA	Log Spectral Amplitude
LSD	Log Spectral Distance
MA	Moving Average
MAP	Maximum *a Posteriori* Adaptation
MCLT	Modulated Complex Lapped Transform
MFCC	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MS	Modulation Spectrum
MTF	Modulation Transfer Function
NSS	Nonlinear Spectral Subtraction
NMT	Noise Masking Threshold
OLA	Overlap Add
PDF	Probability Density Function
PESQ	Perceptual Evaluation of Speech Quality
PLP	Perceptual Linear Prediction
PSD	Power Spectral Density
PSLER	Peak-to-Sidelobe Energy Ratio
RCC	Real Cepstarl Coefficients
segSNR	Segmental Signal to Noise Ratio
SHE	Smoothed Hilbert Envelope
SNR	Signal to Noise Ratio
SP	Spectral Processing

SRR	Signal to Reverberant Ratio
SSA	Speech Specific Approaches
STFT	Short Time Fourier Transform
STSA	Short Time Spectral Amplitude
TP	Temporal Processing
TSP	Temporal and Spectral Processing
UBM	Universal Background Model
VQ	Vector Quantization
VAD	Voice Activity Detector
WSS	Weighted Spectral Slope

List of Symbols

a_k	Linear predictive coefficients
A_l	Time varying amplitude
α	Over subtraction factor
b	Pre-emphasis filter coefficient
$b(n)$	Zero mean Gaussian noise
β	Spectral floor factor
$c(n)$	Cepstral coefficients
C	Number of Mel cepstral coefficients
C_s	Similarity measure
C_{sig}	Signal distortion
C_{bak}	Background intrusiveness
C_{ovl}	Overall quality
$C_h(n)$	Cepstrum of the reverberant impulse response
$C_s(n)$	Cepstrum of the speech signal
χ	Scale factor
$d(n)$	Additive background noise
D	Down sampling factor
ΔSNR	SNR gain
$e(n)$	LP residual
$e_h(n)$	Hilbert transform of the LP residual
$E(k)$	DFT of the LP residual
$E(z)$	Z-transform of the glottal excitation signal
f	Real frequency
f_{mel}	Perceived frequency

F_s	Sampling frequency
F_{cs}	Pitch frequency of the clean speech
F_{tp}	Pitch frequency of the temporally processed speech
F_r	Frequency deviation
$g(n)$	Inverse filter impulse response
$g_d(n)$	Discrete time first order Gaussian differentiator
γ_k	*a posteriori* SNR
G	Gain factor for the excitation signal
$h(n)$	Room impulse response
$h_d(n)$	Direct signal
$h_a(n)$	Early reflections
$h_l(n)$	Late reflections
$h_e(n)$	HE of the LP residual
$h_m(n)$	Mean subtracted HE of the LP residual
$h_{em}(n)$	Emphasized HE of the LP residual
$h_1(n)$	Normalized HE sequences of speech signals collected at mic-1
$h_2(n)$	Normalized HE sequences of speech signals collected at mic-2
$h_{s1}(n)$	Emphasized HE of speaker-1
$h_{s2}(n)$	Emphasized HE of speaker-2
$h_{12}(n)$	Separated HE of the LP residual
$H(l,k)$	Spectral gain function
$H(Z)$	Vocal-tract system filter transfer function
i	Gaussian mixture number
$I_0(.)$	Zeroth order modified Bessel function
$I_1(.)$	First order modified Bessel function
j	Frequency band
k	Frequency index (DFT bin)
l	Frame Index
L	Number of samples per frame
L_g	Gaussian window length

L_r	Frame rate in samples
L_p	Frequency index corresponding to the pitch frequency
λ	Slope parameter (Sigmoidal non-linear mapping function)
M	Total number of triangular Mel-weighting filters
μ_i	Mean vector
$\|m(\omega)\|$	Modulation spectrum
n	Discrete time signal time index
N	Number of points used for computing DFT
N_1	Threshold for early and late reverberation
N_c	Total number of correctly detected low SNR/SRR and high SNR/SRR frames
N_f	Total number of incorrectly classified low SNR/SRR frames
N_i	Length of impulse response
N_p	Total number of harmonics
N_t	Total number of frames
N_{tl}	Total number of low SNR/SRR frames
ω_l	Time varying frequency
p	Linear prediction order
P_{EL}	Percentage of energy loss
P_{NR}	Percentage of noise residue
P_a	Percentage of accuracy in determining the instants of significant excitation
P_c	Percentage of correct detection accuracy
P_f	False detection rate
$P(X\|\Omega)$	GMM likelihood
r	Fixed relevance factor for MAP adaptation
$R(l)$	Normalized autocorrelation sequence
ρ_{12}	Normalized cross-correlation sequence
$\mathbf{R_s}$	Autocorrelation matrix
$s(n)$	Clean speech signal
$S(k)$	Clean speech spectrum
$\hat{S}(k)$	Enhanced speech spectrum

SNR_i	Input signal to noise ratio
SNR_o	Output signal to noise ratio
ξ_k	*a priori* SNR
σ	Standard deviation
Σ_i	Covariance matrix
$S_{zz}(l,k)$	PSD of the reverberant speech
$S_{za}(l,k)$	PSD of the early reverberant components
$S_{zl}(l,k)$	PSD of the late reverberant components
$\hat{S}_{zl}(l,k)$	Smoothed and shifted version of PSD of the reverberant speech
T	Threshold of sigmoidal nonlinear mapping function
T_i	Impulse train duration
T_{60}	Reverberation time
θ_l	Time varying phase
τ	Correlation time index
$u(n)$	Unit step function
$w_f(n)$	Fine weight function
$w_d(k)$	Frequency domain window function for harmonic sampling
$w_g(n)$	Gross weight function
w_{gm}	Minimum value of gross weight function
$w_{\min}$	Minimum value of weight function
$w(l)$	Rayleigh smoothing function
w_i	Weights of Gaussian mixture
$x(n)$	Discrete time signal
$\tilde{x}(n)$	Analytic signal of $x(n)$
$\langle x(n) \rangle$	Average value of the signal $x(n)$
X	Feature vectors
$X_s(k)$	Sampled speech spectrum
$y(n)$	Noisy speech signal
$Y(k)$	Noisy speech spectrum
$z(n)$	Reverberant speech signal

1

Introduction

Contents

1.1 Objective of the Thesis

Speech signals collected over distant microphones or uncontrolled environments may be affected by background noise, reverberation and speech from other speakers to result in degraded speech. The degraded speech is uncomfortable to perceive and gives poor performance when features are extracted for automatic speech processing tasks. The degraded speech therefore needs to be processed to provide perceptual enhancement and also better features for further processing. The processing of the degraded speech for improving quality and intelligibility is termed as speech enhancement. Several methods have been developed for speech enhancement. Majority of them may be broadly grouped into spectral processing and temporal processing methods. In spectral processing methods, the degraded speech is processed in the frequency domain for enhancing the speech components. Alternatively, in temporal processing methods, the degraded speech is processed in the temporal domain for enhancing the speech components. Further, majority of the spectral processing methods involve estimation and removal of degradation component, whereas temporal processing methods involve identification and enhancement of speech specific features. The temporal and spectral processing methods may therefore be treated as complementary approaches for speech enhancement. Also, each of them have their own merits and demerits. These two approaches may be effectively combined by exploiting their merits and aiming to minimize the demerits. This may lead to speech enhancement methods which are more effective and robust compared to only spectral processing or temporal processing. Exploration of the same is the motivation for the research work reported in this thesis. The methods proposed in this thesis work are therefore termed as ***combined temporal and spectral processing (TSP) methods for speech enhancement.***

1.2 Issues in Speech Enhancement

Speech is one of the most desirable modes of communication among humans. It involves several stages, from the coding of thought or information in the talker's brain, to its successful decoding by the listener's brain [1]. In this chain of human communication, the acoustic signal at the output of the speech production system is the carrier of information. This acoustic signal travels through the medium to reach the speech perception apparatus of the listener, where decoding of the acoustic signal and message understanding is made. Several automatic speech processing systems have also found their way in everyday life through their use in mobile communication, speech and speaker recognition, aid

for the hearing impaired and numerous other applications. In all these speech communication systems the quality and intelligibility of speech is of utmost importance for ease and accuracy of information exchange. Here, the quality of speech refers how a speaker conveys an utterance and includes such attributes like naturalness and speaker recognizability. Intelligibility is concerned with what the speaker had said, that is, the meaning or information content behind the words [2]. Both human and automatic speech communications are effective in controlled environments. This is due to the high quality and intelligibility of speech. However, in many situations of practical interest, speech signals are affected by various types of degradations like background noise, reverberation and speech from other speakers. The degraded speech needs to be processed to enhance the speech components present in the signal. The main objective of speech enhancement is to improve the quality and intelligibility of the degraded speech [3]. The methods employed in practice take the nature of degradation into consideration for enhancing the speech components. This is because the signal characteristics will be different for each degradation. Since there are three major types of degradation namely, background noise, reverberation and speech from other speakers, we have the following cases:

(i) Enhancement of speech degraded by background noise (Noisy Speech)

(ii) Enhancement of speech degraded by reverberation (Reverberant Speech)

(iii) Enhancement of speech degraded by competing speakers (Multi-Speaker Speech).

Over the years, researchers and engineers have developed various methods to address the problem of speech enhancement. Yet, due to complexities involved, this area of research still poses a considerable challenge. In general, speech enhancement involves processing of degraded speech signals in temporal or spectral domains. Any such processing introduces its own distortion into the processed speech signal. Typically more the processing employed for reducing the degrading component, more will be the distortion introduced. Hence speech enhancement is a tradeoff between the actual reduction of degrading component and its own distortion. Therefore the performance of the speech enhancement methods is measured in terms of quality and intelligibility of the processed signal [4]. The two performance measures are not correlated. It is also well known fact that improving the quality of the noisy signal does not necessarily elevate its intelligibility. On the contrary, quality improvement is usually associated with loss of intelligibility relative to that of the degraded signal [5].

In terms of production and perception, the message which is formulated in the transmitter's brain by means of neurological process gets transformed into speech by means of a series of muscular movements of the vocal tract. The excitation to the vocal tract is provided by the puffs of air released from the lungs. Depending on the nature of excitation of the vocal tract, speech can be classified into two broad categories namely, voiced speech and unvoiced speech. The excitation of voiced speech is due to the quasi-periodic vibration of the vocal folds, whereas in case of unvoiced speech, the excitation is due to the burst or turbulence of air due to the constriction somewhere along the length of the vocal tract [6]. The signal energy for voiced regions is significantly higher compared to that for the unvoiced regions. Thus in case of degradation, voiced regions (high signal to noise ratio (SNR) regions) play a crucial role in perception [7]. Further in case of voiced speech, the regions around the instants of glottal closure are high SNR relative to the other portions and hence are perceptually significant [8,9].

The perceptual aspects of speech are considerably more complicated and less well understood [10]. However, there are a number of commonly accepted aspects of speech perception which play an important role in speech enhancement systems. Perceptual cues of highly degraded speech can be thought of two levels, namely, cognitive and acoustic levels [11]. At the cognitive level, perception of degraded speech is aided by knowledge of context of conversation, the syntax and semantics of the context and high level features like intonation and duration. At the acoustic level, sound perception in degraded conditions happens mostly by extrapolation of information from the high SNR regions to the low SNR regions in the temporal domain [8]. Furthermore, it is generally understood that the short-time spectrum also plays central importance in the perception of speech. Specifically, the formants in the short-time spectrum are more important than other details of the spectral envelope [10, 12].

The research work reported in this thesis exploits these two factors at the acoustic signal level for developing speech enhancement methods. The proposed methods employ temporal and spectral processing of degraded speech. The temporal processing involves identification and enhancement of high SNR regions in the time domain representation of the degraded speech signal. Spectral processing involves estimation and elimination of degradation component. Also identification and enhancement of speech specific spectral features in the frequency domain representation of degraded speech.

1.3 Temporal or Spectral Processing for Speech Enhancement

1.3.1 Enhancement of Noisy Speech

Speech degradation by the additive background noise often occurs due to sources such as air conditioning units, fans, cars, city streets, factory environments, helicopters and computer systems etc. The speech degraded by the additive background noise is commonly termed as noisy speech. The problem of enhancement of noisy speech has received considerable attention over the past three decades. The reasons being, its wide range of applications and limitations of the available methods. Many solutions have been developed to deal with the noisy speech enhancement problem. Generally, these solutions can be classified into two main areas: Temporal processing and spectral processing based speech enhancement techniques.

Among the available spectral based noisy speech enhancement techniques, the spectral subtraction [13] and minimum mean square error (MMSE) spectral amplitude estimation methods [14, 15] have been widely adopted for suppressing additive background noise. The standard spectral subtraction method estimates the magnitude spectrum of the underlying clean speech by subtracting an estimate of the noise spectrum from the noisy speech spectrum in the short-time Fourier transform (STFT) domain. The greatest asset of this approach lies in its simplicity, since all that is required is an estimate of the mean noise power. However, this approach introduces some artifacts referred as musical noise, due to spectral estimation problems. Several techniques to reduce the musical noise have also been proposed over the past two decades [16]. As widely agreed, the best algorithm from this perspective is the one proposed by Ephraim and Malah [14,15]. Ephraim and Malah [14] derived a MMSE short-time spectral amplitude (STSA) estimator for speech enhancement under the assumption that the Fourier expansion coefficients of the original signal and the noise may be modelled as independent, zero-mean, Gaussian random variables. The enhanced speech is obtained by minimizing the mean squared error between the STSA of the clean speech and the enhanced speech. This estimator gives very good results in practice, with a noticeable reduction in musical noise.

A class of temporal processing methods have been proposed by exploiting the excitation source characteristics of the speech signal for enhancement [17,18]. The basic principle of the excitation source information based temporal processing method is to identify the high SNR regions in the excitation source signal, and derive a weight function that emphasizes the high SNR regions relative to the low SNR regions. The excitation source signal of the noisy speech samples are multiplied with the weight

function, and the modified signal is used to excite the time-varying all-pole filter derived from the noisy speech to generate the enhanced speech. The main merit of these methods is that, they do not produce the type of distortion which the spectral subtraction produces. At the same time the amount of noise suppression is low as compared to spectral based methods.

1.3.2 Enhancement of Reverberant Speech

Reverberation is one of the most important phenomenons which affect the quality of speech communication, in which delayed copies of the speech acoustic waveform, called echoes, are added to the direct speech. The received signal over a distant microphone or uncontrolled environment generally consists of direct sound, reflections that arrive shortly after the direct sound (early reverberation), and reflections that arrive after the early reverberation (late reverberation). The combination of the direct sound and early reverberation is sometimes referred to as the early sound component [19]. The early reverberation components enhance both audibility and intelligibility of direct speech. Early reverberation also causes spectral distortion called coloration. In contrast, late reverberation impairs speech intelligibility [20]. It cannot be integrated with the direct sound or with the early components of reverberation [19].

Several reverberant speech enhancement methods have been proposed using single and multiple microphones. However, until now there are no practical and robust dereverberation techniques available mainly because the degradation is non-stationary, correlated with the signal and cannot easily be modeled. Recently, the spectral processing based methods, especially spectral subtraction based reverberant speech enhancement methods play an important role in the enhancement of reverberant speech. The spectral subtraction based enhancement methods aim at the suppression of late reverberation to improve speech intelligibility [19, 21]. There is another class of excitation source information based reverberant speech enhancement algorithms which primarily aim to emphasize the high signal to reverberant ratio (SRR) regions relative to the low SRR regions of the reverberant speech signal in the temporal domain [22, 23]. The basis for the temporal processing technique is that in case of reverberant environments, the excitation source signal of voiced speech segments contain the original impulses followed by several other peaks due to multi-path reflections. Consequently, dereverberation is achieved by attenuating the peaks in the excitation sequence due to multi-path reflections, and synthesizing the enhanced speech waveform using the modified excitation source signal and the time-varying all-pole filter with coefficients derived from the reverberant speech. The high SRR regions are

emphasized by deriving the weight function to modify the excitation source characteristics at fine and gross levels [22].

1.3.3 Enhancement of Multi-Speaker Speech

One of the challenging tasks in speech processing is the enhancement of speech of individual speaker from the speech collected over multi-speaker environment. In a multi-speaker environment, like meetings, discussions and cocktail parties, several speakers will be speaking simultaneously. The signal collected by a microphone has other speakers speech as degradation that needs to be minimized.

Several methods have been proposed in the literature for processing speech collected in a multi-speaker environment. Depending on the number of microphones used for collecting multi-speaker data, the methods can be divided into single and multi-channel cases. In a single channel case, speech signal is processed to emphasize speech of one of the speakers over the other and is more commonly termed as co-channel separation. In a multichannel case, the speech signal is processed to emphasize speech of each speaker over rest of the speakers. The enhancement of desired speech signal can be done effectively and relatively easily, if the speech signals are collected simultaneously over two or more spatially distributed microphones. In such a case one could exploit the delay in the speech signals produced by an individual at any two microphone locations. The delays obtained for different speakers are different as all the speakers cannot be at the same location simultaneously.

Similar to noisy speech and reverberant speech enhancement methods, in multi-speaker enhancement also many methods have been proposed using the spectral characteristics of the speech and also there exist some methods that use the excitation information of speech production. The methods that use the spectral characteristics rely on the estimation of pitch of the individual speakers and using this information, the desired speaker is enhanced by retaining only pitch and harmonic components and ignoring the remaining spectral components [24, 25]. Since speech energy of a particular speaker is concentrated at the pitch and harmonics, speech signal corresponding to the speaker is synthesized using amplitudes of short time spectrum at the frequencies of harmonics [26]. However, it is generally difficult to obtain the pitch of an individual speaker from the multi-speaker signal. Alternatively, the methods that use the excitation information of speech rely on the time-delay between the microphone signals and also the excitation characteristics of individual speakers for speech enhancement. The basis for this method is that the relative positions of these instants of significant excitation in the direct component of the speech signal remain unchanged at each of the microphones for a given speaker.

These sequences differ only by a fixed delay corresponding to the relative distances of the microphones from the speaker. By estimating time delays and using the knowledge of excitation source characteristics a weight function is derived for each speaker to identify the speech components of desired speaker relative to other speaker [9, 27]. The high values in the weight function indicate the temporal regions where the corresponding speaker speech is predominant.

1.4 Scope of the Present Work

As mentioned in the preceding section, most of the enhancement methods process degraded speech in either temporal or spectral domains for achieving enhancement. The scope of this work is to highlight and demonstrate the merits of combined TSP methods for processing degraded speech. The motivation for the same is justified as follows:

(i) In general, the focus of most of the spectral processing methods for speech enhancement is on the estimation (i.e., spectral characteristics of background noise, late reverberation, interfering speaker) and suppression of the degradation rather than enhancement of the characteristics of the speech signal. Information about the degradation needs to be continuously estimated, particularly, in non-stationary environments wherein degradation characteristics are constantly changing. Alternatively, the temporal processing methods that use the characteristics of excitation source information primarily aim at emphasizing the high SNR/SRR regions of degraded speech signal. Therefore no explicit knowledge of characteristics of degradation is required. The limitation of the temporal processing methods is that the level of removal of degradation achieved may not be significant as in the case of spectral based methods. The integration of these two approaches may lead to better suppression of degradation and also enhancement of high SNR/SRR speech regions. This may lead to improved performance compared to either temporal processing or spectral processing alone.

(ii) The region around the instants of significant excitation like instants of glottal closure and onset of events like burst, frication and aspiration in the temporal domain and formants and pitch and harmonics in the spectral domain are particularly important in the perception of speech. The degradations change the nature of the excitation signal by introducing random values. However, original locations of the instants of significant excitation remain unaltered. In spectral domain also degradation introduces the random spectral peaks into the original speech spectra.

However, the peak locations of the formants will remain unchanged. From the enhancement point of view, temporal processing methods identify and enhance the regions around original locations of the instants of significant excitation and spectral processing methods estimate and attenuate the degrading components. This leads to enhancement of perceptually significant spectral components. Thus the combination of these two approaches emphasizes both of these perceptual elements in the corresponding temporal and spectral domains.

(iii) From the speech production point of view, the temporal and spectral processing methods use independent information from the degraded speech. It will be therefore interesting to study whether they are exploiting complementary information for processing. If so, then they can be suitably combined to develop robust methods for the speech enhancement.

(iv) The temporal and spectral processing methods introduce their own distortion into the processed signal. The level of distortion may be kept minimum by processing to a moderate level in each domain than the usual high level (like over subtraction and very low weight function values).

Motivated by these observations, this work develops combined TSP methods for processing degraded speech. The primary objective is to show that the combined TSP gives better performance compared to the individual temporal or spectral processing methods. This is demonstrated by the following works:

(i) Combined TSP method for noisy speech enhancement

(ii) Combined TSP method for reverberant speech enhancement

(iii) Combined TSP method for two speaker separation, and

(iv) Combined TSP methods evaluation in speaker recognition under degraded condition.

1.5 Organization of the Thesis

The rest of this thesis is organized as follows:

Chapter 2 gives a review of the several existing methods for processing degraded speech. The review is mainly divided into three sections. Section 2.2 presents a review of the methods for processing speech degraded by background noise. Section 2.3 discuses the enhancement techniques for speech degraded by reverberation. Methods for enhancement of speech from multi-speaker environment are discussed

in Section 2.4. Summary of the review and the scope for the present work are given in Section 2.5. The organization of the work is given in Section 2.6.

Chapter 3 presents the proposed combined temporal and spectral processing method for enhancement of noisy speech. Various experimental studies and objective quality measures performed on the individual and the combined processing methods are described in this chapter.

In Chapter 4, a reverberant speech enhancement method based on combined temporal and spectral processing is developed. This method is based on the suppression of late reflections with the help of spectral processing and enhancement of high SRR regions with the help of temporal processing. Different experimental studies and objective quality measures performed on the individual and the combined processing methods are described in this chapter.

Chapter 5 describes the proposed two microphone based combined temporal and spectral processing method for the enhancement of multi-speaker (two speaker) speech. This chapter also discusses the several performance measures to assess the performance of the proposed two speaker speech enhancement method.

Chapter 6 provides the results of speaker recognition experiments performed in the presence of background noise, reverberation and interfering speaker speech by employing individual and the combined processing methods as a pre-processing stage.

Finally, Chapter 7 summarizes the work presented in this thesis, highlights the main contributions of the work and gives some directions for future research.

2

Methods for Processing Degraded Speech - A Review

Contents

2.1 Introduction

Speech signals in real world scenarios are often corrupted by various type of degradations. Most common degradations include background noise, reverberation and speech of competing speaker(s). Several enhancement approaches have been proposed to eliminate these degradations with minimal distortion to the speech signal. The approach to speech enhancement varies considerably depending on the type of degradation. For example, the type of processing suggested for enhancing speech degraded by background noise is different from that suggested for enhancing speech degraded by reverberation or for competing speaker(s). This chapter will provide an overview of enhancement techniques for speech degraded by background noise, reverberation and competing speaker(s). This chapter is organized as follows: Section 2.2 presents a review of the methods for processing speech degraded by background noise. Section 2.3 discuses the enhancement techniques for speech degraded by reverberation. Methods for enhancement of speech in multi-speaker environment are discussed in Section 2.4. Summary of the review and the scope for the present work is given in Section 2.5. The organization of the present work is given in Section 2.6.

2.2 Enhancement of Noisy Speech

The background noise is the most common factor degrading the quality and intelligibility of speech [28]. The term background noise refers to any unwanted signal that is added to the desired signal. Background noise can be stationary or non-stationary. Stationary noise, made by a computer fan or air conditioning, has a power spectral density that does not change over time. Non-stationary noise, caused by door slams, radio and television, has statistical properties that change over time. In practical applications, a speech signal captured with a close talking microphone has little background noise. However, if a distant microphone is used instead, a large amount of background noise is recorded along with the speech. Mathematically, speech degraded by the background noise can be expressed as the sum of clean speech and background noise [29]. That is,

$$y(n) = s(n) + d(n) \tag{2.1}$$

where $y(n)$, $s(n)$ and $d(n)$ denote the noisy speech, clean speech and background noise, respectively. In the frequency domain, it can be represented as

$$Y(k) = S(k) + D(k) \tag{2.2}$$

where k is the index of frequency bin.

The problem of enhancing speech degraded by the background noise received considerable attention in the literature for several decades and numerous methods have been proposed by the signal processing community. Majority of these methods may belong into one of these two categories: *Spectral processing methods* such as the spectral subtraction, minimum mean square error (MMSE) estimator and wavelet denoising methods and *temporal processing methods* such as linear prediction (LP) residual based methods. This section will briefly review the basic principles of these methods.

2.2.1 Spectral Processing Methods

The spectral enhancement methods are the most popular techniques for noise reduction, mainly because of their simplicity and effectiveness. Most of the spectral enhancement techniques rely on the basis that, the human speech perception is not sensitive to short-time phase [30, 31]. This is exploited in these enhancement methods, where only the spectral magnitude associated with the original signal is estimated. In case of noisy speech, most of the available spectral processing methods can be grouped into non-parametric and statistical model-based methods [16]. Methods from the first category usually remove an estimate of the distortion from noisy features, such as subtractive type algorithms and wavelet denoising. The statistical model based speech enhancement such as MMSE estimator uses the parametric model of the signal generation process [32].

2.2.1.1 Spectral Subtraction

Spectral subtraction is historically one of the first algorithms proposed in the field of background noise reduction that is still referenced today because of its minimal complexity and relative ease in implementation. Spectral subtraction is performed by subtracting the average magnitude of the noise spectrum from the spectrum of the noisy speech to estimate the magnitude of the enhanced speech spectrum [13]. The noise is assumed to be uncorrelated and additive to the speech signal. The noise estimation is obtained based on the assumption that the background noise is locally stationary so that the noise characteristics computed during the speech pauses are a good approximation to the noise

characteristics. Accordingly, the estimate of the enhanced speech spectrum is obtained as [13]

$$|\hat{S}(k)| = |Y(k)| - |\hat{D}(k)| \tag{2.3}$$

where $\hat{D}(k)$ is the average magnitude of the noise spectrum.

The enhanced spectrum obtained using the above relation may contain some negative values due to the errors in estimating the noise spectrum. The simplest solution is to half-wave rectifies these values to ensure a non-negative magnitude spectrum. This non-linear processing of negative values creates small, isolated peaks in the spectrum occurring at random frequency locations in each of the frames. Converted in the time domain, these peaks sound similar to the tones with frequencies that change randomly from frame to frame, that is, tones that are turned on and off at the analysis frame rate. This type of noise is commonly referred as musical noise [31, 33, 34]. The main factors contributing to the musical noise phenomenon include the large variance in the estimates of the noisy and noise signal spectra and the large variability in the suppression function [35]. The musical noise can be more annoying to the listeners than the original distortion caused by the background noise. Several modifications for the standard spectral subtraction method have been proposed to alleviate the speech distortion introduced by the spectral subtraction process [13, 33–62].

Boll [13] proposed few modifications such as magnitude averaging, residual noise reduction and additional signal attenuation during non speech activity to reduce the effect of musical noise. Berouti *et al.* [33] suggested a method to reduce the musical noise by subtracting an overestimate of the noise spectrum, while preventing the resultant spectral components from going below a preset minimum value. The proposed technique has the following form

$$|\hat{S}(k)| = \begin{cases} |Y(k)| - \alpha|\hat{D}(k)|, & |Y(k)| - \alpha|\hat{D}(k)| > \beta|\hat{D}(k)| \\ \beta|\hat{D}(k)|, & \text{otherwise} \end{cases} \tag{2.4}$$

where α is the over subtraction factor which is a function of the noisy signal to noise ratio (SNR) and calculated as

$$\alpha = \alpha_0 - \frac{3}{20}SNR, \quad -5dB \leq SNR \leq 20dB \tag{2.5}$$

where α_0 is the desired value of α at 0 dB SNR. Here, SNR is computed as the ratio of the noisy speech power to the estimated noise power. In general, higher the amount of over subtraction is, the stronger components with a low SNR get attenuated. This prevents musical noise. But, too strong

over subtraction will suppress too many components. Therefore, the value of α has to be carefully chosen in order to prevent both the musical noise and signal distortion [33]. The introduction of spectral floor β prevents the subtraction of spectral components of the enhanced speech spectrum falling below the predefined lower value.

A frequency adaptive subtraction factor based approach is proposed in [38, 39]. The motivation is based on the assumption that, in general noise may not affect the speech signal uniformly over the whole spectrum. Some frequencies are affected more severely than the others depending on the spectral characteristics of the noise. Accordingly, Lockwood and Boudy [38] proposed the non-linear spectral subtraction (NSS) method based. In NSS method, the over subtraction factor is frequency dependent in each frame of speech. Larger values are subtracted at frequencies with low SNR levels, and a smaller values are subtracted at frequencies with high SNR levels. Kamath and Loizou [39] extended this concept and developed a multi-band spectral subtraction method that divides the speech spectrum into N non-overlapping bands, and the over subtraction factor for each band is calculated independently. The individual frequency bands of the estimated noise spectrum are subtracted from the corresponding bands of the noisy speech spectrum. Several supplementary schemes such as spectral smoothing, formant intensification and comb filtering are proposed to improve the performance of the spectral subtraction [49]. In [44] a method is proposed to reduce the musical noise in silence and unvoiced region by dividing each silence and unvoiced frame of spectral subtracted speech into several sub-frames and randomizing the phases of each sub-frame over a uniform interval. Hasan *et al.* [50] proposed a self adaptive averaging factor to estimate the *a priori* SNR, which is applied to the conventional spectral subtraction algorithm. However, the performances of the above methods are not satisfactory in adverse environments, particularly when the SNR is very low. The reason is that in very low SNR conditions, it is still difficult to suppress noise without degrading intelligibility, and without introducing residual noise and speech distortion [32].

As reviewed above it is very difficult to minimize the musical noise without affecting the speech, and hence there is a trade off between the amount of noise reduction and speech distortion. Due to this fact, several perceptual-based approaches, wherein instead of completely eliminating the musical noise and introducing distortion, the noise is masked taking advantage of the masking properties of the auditory system [48, 51, 63–67]. The idea of the algorithms is based on the simultaneous masking properties of the human auditory system. The masking effect means that a stronger signal can make

a weaker signal occurring simultaneously inaudible. In other words, if the noise signal is weaker than the speech signal in the same frequency band, the noise signal is masked by the speech signal. Therefore, we can use less noise subtraction to avoid unnecessary distortion. Accordingly, instead of attempting to remove all noise from the signal, these algorithms attempt to attenuate the noise below the audible threshold [68]. Tsoukalas *et al.* [63] used a spectral subtraction technique based on the aspects of the auditory process. Their method considers an enhancement approach that uses the auditory masking threshold (AMT) [69] in spectral subtraction. Virag [64] presents a detailed analysis of the effect of variations in the subtraction parameters like the over-subtraction factor, the spectral flooring factor, and the exponent on the residual noise as well as the intelligibility of the enhanced speech. The methods that adopt the masking property of the human auditory system can reduce the effect of residual noise, but the drawback is the large computational effort associated with the subband decomposition and the additional fast Fourier transform (FFT) analyzer required for psychoacoustic modeling [32]. In summary, even though several improvements have been proposed, spectral subtraction approach is still a subject of many researchers to increase its performance in terms of minimizing the effect of musical noise and also making it suitable for non-stationary environments.

2.2.1.2 MMSE Estimator

In spectral subtraction based methods, there were no specific assumptions made about the distribution of the spectral components of either speech or noise. Ephraim and Malah [14,15] have proposed a system that utilizes the MMSE criteria using models for the distribution of the spectral components of the speech and noise signals. The MMSE - short time spectral amplitude (STSA) estimator for speech enhancement aims to minimize the mean square error between the short time spectral magnitude of the clean and enhanced speech signal. This method assumes that each of the Fourier expansion coefficients of the speech and of the noise process can be modeled as independent, zero-mean, Gaussian random variables. Derivation of the MMSE-STSA estimator is given in Appendix-A.

The MMSE log-spectral amplitude (MMSE-LSA) estimator for speech enhancement was proposed by Ephraim and Malah in 1985 [15]. In their previous work on MMSE estimation of the STSA, the aim was to enhance the speech by minimizing the error between the STSA of the clean speech and the enhanced speech. This optimality criterion gives very good results in practice, with a noticeable reduction in musical noise, but does not consider any of the non-linear characteristics observable in human perception [70]. To incorporate perceptually significant information into the algorithm, the

authors propose to minimize the mean square error between the logarithm of the STSA of the clean and enhanced speech. That is, the LSA estimator minimizes

$$E\left\{\left(\log_e A_k - \log_e \hat{A}_k\right)^2\right\} \tag{2.6}$$

where A_k denotes the spectral speech amplitude, and $\hat{A}_k$ is its optimal estimator. Ephraim and Malah did some subjective comparisons between MMSE-STSA and MMSE-LSA estimators to show that the enhanced speech using the MMSE-LSA estimator suffers from much less residual noise, while there is no perceptible difference in the enhanced quality of speech itself.

The Ephraim-Malah algorithm [14, 15] has received much attention by the scientific community. This is mainly due to its ability to achieve a highly satisfying overall quality of the enhanced speech and hence makes it suitable for practical implementations in applications like digital hearing aids [71]. In [72], using a statistical model similar to [14], and using the uncertainty of speech presence, a MMSE amplitude estimator is developed in the discrete cosine transform (DCT) domain. It has been shown through experiments that the DCT provides better energy compaction than the discrete Fourier transform (DFT).

A fundamental assumption made in the MMSE algorithms is that the real and imaginary parts of the clean DFT coefficients can be modeled by a Gaussian distribution. This Gaussian assumption, however, holds asymptotically for long duration analysis frames, for which the span of the correlation of the signal is much shorter than the DFT size. While this assumption might hold for the noise DFT coefficients, it does not hold for the speech DFT coefficients, which are typically estimated using relatively short (20-30 ms) duration windows [73]. This observation led researchers to derive a similar optimal MMSE-STSA estimator, but use of non-Gaussian distributions for modeling the real and imaginary parts of the speech DFT coefficients [73–78]. In particular, the Gamma [75] or the Laplacian [73] probability distributions are used to model the distributions of the real and imaginary parts of the DFT coefficients.

All MMSE-based methods need the estimate of the *a priori* SNR, the SNR of the k^{th} spectral component of the clean speech signal. Since the knowledge of clean signal is seldom available in practical systems, a decision-directed estimation and maximum likelihood (ML) estimation are the two approaches taken to compute *a priori* SNR [14]. Later several modifications are proposed in the decision-directed estimation. First, Cappe [79] provides a more detailed analysis on decision-directed

estimation approach and proposed a lower limit to the estimate of the *a priori* SNR in order to reduce the annoying musical tones. Later, Cohen introduced a causal and non-causal recursive estimators for the *a priori* SNR, which take into account the time-frequency correlation of speech signals [80–83]. The causal *a priori* SNR estimator is closely related to the decision-directed estimator of Ephraim and Malah. The non-causal *a priori* SNR estimator employs future spectral measurements to better predict the spectral variances of the clean speech. Experimental results show that the non-causal estimator yields a higher improvement in the segmental SNR and lower log-spectral distortion, than the decision-directed method and the causal estimator [81]. Even though several improvements have been made in MMSE estimator, the algorithms proposed by the Ephraim and Malah [14, 15] are still considered as the state of art algorithms for noisy speech enhancement.

2.2.1.3 Wavelet Denoising

Most of the single channel speech enhancement algorithms are applied in the frequency domain using short-time Fourier transform (STFT), which allows analyzing non-stationary speech signals. The speech signal is divided into short frames during which the signal is assumed to be stationary. STFT provides a compromise between time resolution and frequency resolution. But, once the frame length is chosen, the time resolution is same for all frequency components. Some of the speech enhancement algorithms are developed using wavelet transform [84]. Wavelet transform provides variable window size for different frequency components [32, 65, 84–98]. This allows the use of long time intervals to obtain low frequency components and short intervals for high frequency components and hence it provides a more flexible time-frequency representation of speech [84].

One popular technique for wavelet-based signal enhancement is the wavelet shrinkage algorithm which was proposed by Donoho [85]. Wavelet shrinkage is a simple denoising method based on the thresholding of the wavelet coefficients. The estimated threshold is supposed to define the limit between the wavelet coefficients of the noise and those of the target signal [87]. However, it is not always possible to separate the components corresponding to the target signal from those of noise by a simple thresholding. For noisy speech, energies of unvoiced segments are comparable to those of noise. Applying thresholding uniformly to all wavelet coefficients not only suppresses additional noise but also some speech components like unvoiced ones [84]. Consequently, the perceptive quality of the filtered speech will be greatly affected [87]. Therefore the wavelet transform combined with other signal processing tools like Wiener filtering in the wavelet domain and wavelet filter bank for

spectral subtraction have also been proposed for speech enhancement [99]. More recently a number of attempts have been made to use perceptually motivated wavelet decompositions coupled with various thresholding and estimation techniques to improve the performance [65, 84, 88–91, 97].

2.2.1.4 Noise Estimation Methods

The effectiveness of spectral processing approach relies on the consistency and accuracy of noise magnitude spectrum estimates. It refers to either estimating the noise spectra for spectral subtraction or estimating the *a priori* and *a posteriori* SNR in the case of MMSE algorithms. Two possible methods can be used for this purpose. The first method segments the noisy signal into two classes: Speech activity and speech pause regions. The noise is estimated during speech pause regions, where speech pause is usually detected by a traditional voice activity detection (VAD) algorithm [100–120]. An effective VAD algorithm is critical for achieving the better enhancement without degrading speech quality [121]. The second method performs a moving average of the short time spectra of the noisy signal, the time constant is selected to present a longer decay than the speech variation. The second method has the advantage that it does not require an explicit segmentation of speech activity regions.

In the first case, VADs have to be robust to adapt to the changes of the noise characteristics, which is a difficult task. In particular, unvoiced segments of the speech signal are more difficult to detect than voiced segments, because they are more similar to the noise and the SNR is generally lower in unvoiced than in voiced segments [115, 122]. Recently many techniques have been proposed in an attempt to overcome these limitations. Martin [41, 46, 123] proposed a method for estimating the noise spectrum based on tracking the minimum of the noisy speech over a finite window. This method is based on the observation that the power of the noisy speech signal in individual frequency bands often decays to the power level of the noise, even during speech activity [41, 46, 123]. Therefore this fact can be used to obtain an estimate of the noise level in individual frequency bands by tracking the minimum within a short window of the noisy speech spectrum. In Martin method, to search the minimum of the local energy, the following recursively smoothed periodogram is considered [16].

$$P_y(l,k) = \eta P_y(l-1,k) + (1-\eta)\,|Y(l,k)|^2 \tag{2.7}$$

where $P_y(l,k)$ is the smoothed power spectrum, $|Y(l,k)|^2$ is the short-time power spectrum of noisy speech and η is a smoothing constant which is generally chosen to be very close to 1 [46]. Then, a minimum frequency bin estimation is carried out by considering set of R previous smoothed periodogram

values. That is,

$$P_d(l,k) = \min\left(P_y(l-R,k) : P_y(l,k)\right) \tag{2.8}$$

where $P_d(l,k)$ is the estimate of the noise power spectrum.

To obtain reliable noise power estimates, the estimating window for minimum search must be large enough to cover any burst of speech activity, but it also has to be short enough to track the fast changes in the noise level. The typical time span of the segment may range from 400 ms to 1 sec [16]. In contrast to other methods the minimum statistics algorithm does not use any explicit threshold to distinguish between speech activity and speech pause. Recently several improvements have been made in the minimum statistics approach to improve the accuracy of the noise estimate [124–130].

2.2.2 Temporal Processing Methods

2.2.2.1 LP Residual Enhancement

Most of the studies on the speech enhancement discussed above focus on enhancement based on suppression of noise. These methods disturb the spectral balance in speech, resulting in unpleasant distortions in the enhanced speech. Yegnanarayana *et al.* proposed a noisy speech enhancement method by exploiting the characteristics of excitation source signal such as LP residual [17]. The basic approach for speech enhancement is to identify the high SNR portions in the noisy speech signal, and enhance those portions relative to the low SNR portions, without causing significant distortion in the enhanced speech. The residual signal samples are multiplied with the weight function, and the modified residual is used to excite the time-varying all-pole filter derived from the given noisy speech to generate the enhanced speech. In this method, enhancement is carried out by the following three steps: (i) Identification and enhancement of high SNR regions at the gross level, (ii) identification and enhancement of high SNR regions at the fine level and (iii) enhancement of spectral peaks over valleys.

At the gross level the regions corresponding to low and high SNR regions are identified from the characteristic of the LP residual. A weighting function for the residual signal samples is derived based on the smoothened inverse spectral flatness characteristics of the noisy speech signal. The spectral flatness characteristics are derived by comparing the energy in the residual signal with the energy in the noisy speech signal in each short interval of about 2 ms. At the fine level, for voiced segments, if the SNR is low in some short (1-3 ms) segments, then the residual signal in those regions can be given lower weight compared to the adjacent higher SNR segments. A weight function at the fine level

was derived from the residual energy plot to de-emphasize the segments corresponding to the valleys relative to the segments corresponding to the peaks. However for noisy speech, the residual signal is noisy and so the energy of the short segment of residual signal may not be reliable for deriving the weight function. Hence, the Frobenius norm of the Toeplitz prediction matrix constructed using the noisy speech samples in a frame of 2 ms duration is used to represent the short time energy of the corresponding frame of the LP residual signal.

In [131] a speech enhancement algorithm similar to [17] is proposed. It differs with the former residual weighting scheme in that the weights on the LP residuals are derived based on a constrained optimization criterion. Enhanced speech is obtained by exciting the time varying all pole synthesis filter with the enhanced residual. In [18] authors exploited the use of coherently added Hilbert envelope (HE) for LP residual reconstruction. The feature that the HE has large amplitude at the instant of strong excitation makes it a good indicator of glottal closure (GC), where an excitation pulse takes place. Therefore, applying the HE to LP residual as a weighting function has the effect of emphasizing the pulse train structure for voiced speech, which leads to an enhanced LP residual signal. Similarly, a multi channel speech enhancement method using the GC events of speech signal is also proposed using the HE of the LP residual and the enhancement is achieved by emphasizing the excitation around the GC events [9].

2.2.3 Signal Subspace Approach

Another class of noisy speech enhancement methods that has gained a lot of attention is the signal subspace approach [132–134]. The main principle of this approach is, each vector of noisy speech is composed of a signal plus noise subspace or simply signal subspace and the noise subspace. The noise subspace contains signal from noise only. Enhancement is achieved by removing the noise subspace and estimating the enhanced signal from the remaining signal subspace. The decomposition of the noisy signal into signal subspace and noise subspace can be done using either the singular value decomposition (SVD) [135, 136] or Karhunen-Loeve transform (KLT) [133, 137–139].

The SVD method proposed by Dendrinos *et al.* [135] is based on the idea that some of the eigenvectors and their corresponding eigenvalues of the observation data matrix contain the speech signal information, while other eigenvalues or eigenvectors represent only noise. The enhanced signal was reconstructed from the dominant eigenvalues along with their corresponding eigenvectors, while neglecting the small eigenvalues which typically carry only noise information. Jensen *et al.* [136] extended

this approach to colored noise using the quotient SVD (QSVD). A different formulation to subspace based methods for speech enhancement was provided by Ephraim and Van Trees in [133]. They also proposed two signal estimators: the spectral domain constraint (SDC) and the time domain constraint (TDC). The former attempted to spectrally shape the residual noise while the latter constrained residual noise energy [133]. The decomposition of the vector space of the noisy signal into signal and noise subspace can be obtained by applying the KLT to the noisy signal. The KLT components representing the signal subspace were modified by a gain function determined by the estimator, while the remaining KLT components representing the noise subspace were eliminated. The enhanced signal is obtained from the inverse KLT of the modified components. In [140], a frequency Eigen domain transformation (FET) is introduced to incorporate the masking properties into a signal subspace approach. The masking based subspace speech enhancement method is reported to yield an improved performance over conventional subspace approaches [139, 140]. However, a difficult task in subspace methods is to accurately determine the dimension of the subspaces in the presence of non-stationary noise and another main drawback of this approach is large computational load.

In addition to all these methods, several multi-microphone algorithms have been proposed for noisy speech enhancement. As already mentioned, multi-microphone noise reduction techniques can exploit the spatial information in the microphone signals when the speech and the noise sources are located at different positions, and hence are able to perform both spectral and spatial filtering. Because of this advantage, multi-microphone based methods generally give better results compared to single microphone based methods.

2.3 Enhancement of Reverberant Speech

Reverberation is caused by the superposition of an acoustic signal and its reflected signals of different delays and amplitudes [141]. When the speech signal is produced in a room, it follows multiple paths from source to microphone, some portion of the signal energy that reaches the microphone is transmitted directly through the air, while the remainder is reflected off one or more surfaces in the room prior to reception. Usually the earliest reflections arrive discretely, while later reflections arrive in rapid succession or concurrently as the number of paths the sound may take increases. These reflections result in signal attenuation and spectral distortion, called reverberation. Mathematically, this can be expressed as convolution of the speech signal with room impulse response [142]

$$z(n) = s(n) * h(n) \tag{2.9}$$

where $s(n)$ represents the original speech signal, $h(n)$ denotes the room impulse response, and * characterizes the linear convolution operation.

The room reverberation is completely characterized by the room impulse response. If the response can be accurately estimated, its effects can be reversed. However, the response $h(n)$ depends on geometric and acoustic characteristics of the room and also on the locations of source (speaker) and microphone. Generally the room impulse response is modelled as a zero-mean random sequence with a decaying exponential [21]. Thus,

$$h(n) = b(n)e^{-\overline{\delta}n}u(n) \tag{2.10}$$

where $b(n)$ represents a zero-mean white Gaussian noise, $u(n)$, the unit step function, and $\overline{\delta}$ is a damping constant related to the reverberation time T_{60}, obtained by [21]

$$\overline{\delta} = \frac{3\ln(10)}{T_{60}}. \tag{2.11}$$

Here, the reverberation time T_{60} is defined as the time needed for the sound energy to fall by 60 dB after the original sound source is turned off [143]. It is the main parameter used for characterizing the acoustics of an environment. Typical office rooms exhibit T_{60} between 0.2 and 0.6 sec while conference rooms present T_{60} of 0.8 to 1.2 sec [144]. Longer reverberation times mean that the sound energy stays in the room longer before being absorbed.

The speaker to receiver room impulse response can be divided into three segments: Direct sound, early reflections and late reflections [Fig. 2.1].

Direct Sound: The first sound that is received without reflection is the direct sound. In case the source is not in line of sight of the observer there is no direct sound. The delay between the initial excitation of the source and its observation is dependent on the distance and the velocity of the sound [19].

Early Reverberation: A little time later the sounds which were reflected off one or more surfaces (walls, ceiling, furniture, and floor of the room, etc.) will be received. These reflected sounds are separated in both time and direction from the direct sound. The reflected sounds form a sound component which is usually called early reverberation [19]. Early reflections, therefore, come from fixed directions surrounding the listener in a closed room [145, 146]. In a small room, the early reflections would arrive very close to the direct sound. In larger rooms, they would arrive later in time because of the longer propagation time for the sound waves across the room. However, early reverberation is not perceived as a separate sound to the direct sound so long as the delay of the reflections does not exceed a limit of approximately 80-100 ms with respect to the arrival time of the direct sound and is therefore considered useful with regard to speech intelligibility [19]. This is often referred to as the precedence effect [147]. The amplitudes of the early reflections also depend on the size of the room, as the sound level is inversely proportional to the distance traveled [145]. This early reverberation also causes a spectral distortion called colouration [19].

Late Reverberation: After the early reflections, the rate of the arriving reflections increases greatly. They are perceived either as separate echoes, or as reverberation, and impair speech intelligibility [19, 148, 149]. It affects temporal structure, spectral content, intensity and interaural differences among other parameters [145].

It is known that reverberation may be responsible for degrading the audible quality of speech recorded by distant microphones, thus causing problems in applications like hands-free telephony, teleconferencing, and hearing aids. Reverberation also degrades the speech characteristics used in automatic speech or speaker recognition [150]. Therefore, solving the reverberation problem is important for many speech processing applications. Various methods for improving the performance in reverberant environments have been proposed. Same as noisy speech enhancement methods, these methods also may be broadly grouped into two categories: *Temporal processing methods* and *spectral processing*

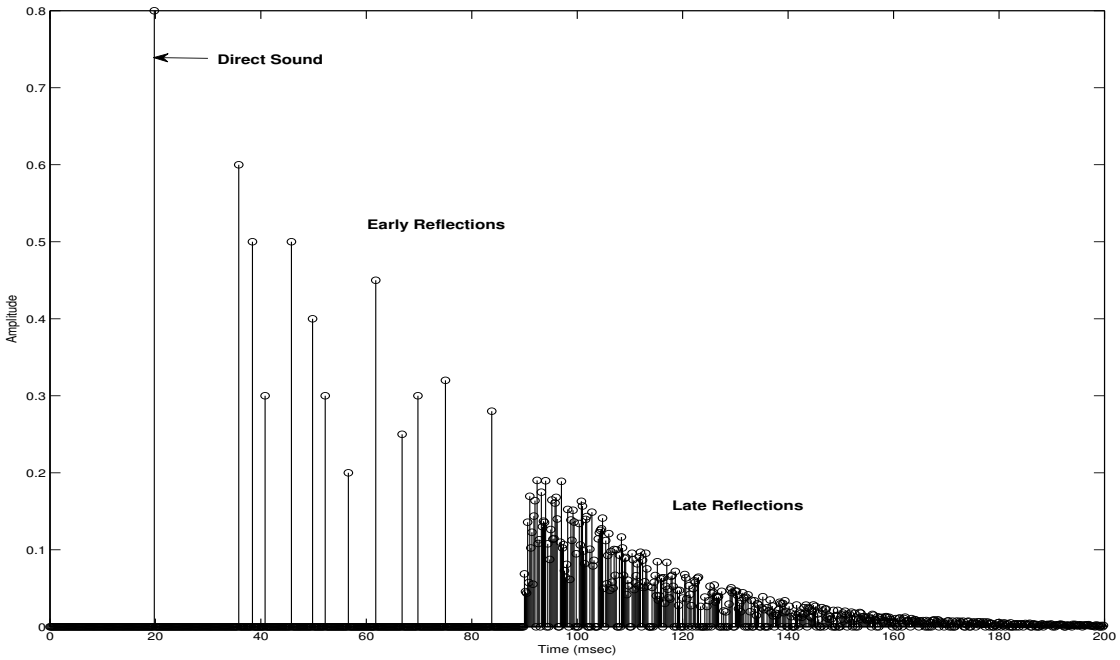


Figure 2.1: Schematic representation of an acoustic impulse response.

methods. The temporal processing methods obtain the enhancement by processing the reverberant speech signal in time or cepstral domain and spectral domain processing is accomplished in frequency domain. Besides these categories there are several *multi-stage algorithms* that have been proposed which process degraded signal in both time and frequency domains.

2.3.1 Spectral Processing Methods

As mentioned before, the spectral enhancement methods achieve dereverberation by modifying the short time magnitude spectrum of the reverberant speech. Initially, Flanagan *et al.* [151] proposed a spectral based two microphone approach for processing reverberant speech. The speech signal from each microphone is separated into several subbands. In each frequency band, the spectral amplitudes of the two signals are compared and the maximum amplitude was selected as the contribution for the reconstructed speech. This method exploits the periodic nature of the spectral distortion of speech caused by simple echo. Because the two microphones spaced at different locations have echoes of different delays and nulls appear at regular but different intervals in the spectra of the microphone outputs. An algorithm proposed by Allen *et al.* [152] first filters the two individual microphone signal

into frequency band, then the filtered outputs are compensated for delay differences and added. For each band the correlation between the two microphone signals are computed, and used as a gain factor for that band to suppress the spectral bands with low correlations in order to remove the reverberation effects. This is done based on the assumption that bands with high levels of coherence contains strong direct component whereas bands with low levels of coherence mainly contains reverberation. Another form of spectral processing method was proposed in [153] using the harmonic structure of speech based on pitch estimation. In the proposed method, the harmonic part of the speech is extracted by adaptive filtering. Averaging the ratio of DFT of the harmonic part of speech and that of the reverberant part, a dereverberation filter is calculated which reduces reverberation in both voiced and unvoiced speech segments. The technique is suitable when sufficient number of training utterances is available and the room impulse response does not change significantly.

Recently, spectral subtraction based spectral processing technique has been developed to suppress late reverberation effect [19,21,154–158]. Lebart *et al.* [21] introduced a single channel speech dereverberation method based on spectral subtraction to reduce this effect. In this study, the energy of the late reverberations is estimated using an exponential decaying function. That is, this method first divides the observed signal and the estimates the late reverberations into short frames, apply short-term Fourier transform (STFT) to calculate the power spectrum, and then subtract the power spectrum of the estimated late reverberations from that of the observed signal. In [21], the authors assumed that reverberation time was frequency independent and the energy related to the direct sound could be ignored. Hence, the authors assume that the signal to reverberation ratio (SRR) of the observed signal is smaller than 0 dB which limits the use of the proposed solution to situations in which the source-microphone distance is larger than the critical distance. Critical distance is the distance at which the reverberant sound field is equal in level to the direct sound from a sound source [143]. This issue has been addressed in [19], where the room impulse response model is generalized by considering the direct component and the reflections separately. Habets derived a novel estimator which is advantageous over the late reverberant power spectral density estimator proposed by Lebart in case the source-microphone distance is smaller than the critical distance [19]. The method estimates the power spectrum of the reverberation based on a statistical model of late reverberation and then subtracts it from power spectrum of the reverberant speech. One of the main problems in spectral subtraction is the nonlinear processing distortion, for example the musical noise caused by over-subtraction of the

reverberation. Since reverberation cannot be well-represented solely with such a simple model, i.e., an exponential decay model. This distortion degrades the quality of the processed speech. However, there are some well-known methods like spectral floor factor, *a priori* SRR estimation techniques available to reduce this non-linear distortion to a certain level. The main advantage of this method is that it is computationally simple and relatively robust to noise [159].

2.3.2 Temporal Processing Methods

2.3.2.1 Inverse Filtering

Reducing reverberation through inverse filtering is one of the most common approaches. The basic idea is to pass the reverberant signal through a second filter that inverts the reverberation process and recovers the original signal [142, 160–165]. This can be written as

$$s(n) * h(n) * g(n) = \hat{s}(n) \tag{2.12}$$

where $g(n)$ is the inverse filter impulse response and $\hat{s}(n)$ is the delayed replica of $s(n)$. There are several well known inverse filtering methods to dereverberate the original signal. For example, Neely and Allen proposed a method that used a single microphone to remove a minimum phase component from the room effect [142]. Miyoshi and Kaneda proposed another method that used a microphone array and constraining non-overlaps of zeros in all pairs of the impulse responses between the sources and the microphones [163]. Wang and Itakura proposed a method of acoustic inverse filtering through multi-microphone subband processing that selects the best invertible microphone in each subband and reconstructs the fullband signal by summing up the inverse filtered subband signals of the best microphones [166].

Experiments performed by Gillespie *et al.* [162] showed that the kurtosis of the LP residual is a reasonable measure of reverberation. The LP residual signal becomes more Gaussian due to reverberation; consequently, the kurtosis becomes smaller. Using this principle, in [162] the authors proposed a kurtosis maximization adaptive filtering algorithm using a modulated complex lapped transform (MCLT) subband filter to estimate the impulse response for the dereverberation of speech. A least mean squares (LMS) gradient descent approach is chosen to maximize kurtosis in an adaptive manner. A significant reduction for perceived reverberation was reported. However the authors found that use of the kurtosis maximization required a large amount of data and often resulted in unstable adaptation that was highly dependent on the room impulse response.

The challenge in the inverse filtering method is to find the inverse impulse response $g(n)$. The perfect reconstruction of the original signal exists only if the room impulse response function is a minimum phase filter, whose poles and zeros are all inside the unit circle. But in practical case, most room transfer functions are non-minimum phase due to late energy in the room impulse response [142]. Bees *et al.* [161] described a cepstral processing approach for estimation of impulse response from the reverberant speech signal only, by exponentially windowing speech segments to force a minimum phase characteristic and averaging of a number of such frames. They designed a least squared error inverse filter to remove the estimated impulse response from the reverberant speech. This appears to be the most successful estimation approach to date, but its practical use is limited severely by the requirement for a minimum phase characteristic after windowing, framing effects, cepstral overlap of speech and late reverberation [167].

Some researchers have proposed using a subspace method for estimating the impulse responses [159,168,169]. The room impulse responses are obtained from the null space of the covariance matrix of the observed signals. However, these subspace methods are highly dependent on a prior knowledge of channel orders, and are sensitive to errors in channel order estimates. Recently in [170,171] the authors proposed a single-microphone dereverberation method, named Harmonicity based dEReverBeration (HERB). HERB estimates the inverse filter for an unknown room transfer function by utilizing an essential feature of speech, namely harmonic structure. However, in general the methods based on inverse filtering are known to pose a sensitivity problem in that background noise or a small change in the transfer function results in severe performance degradation [159, 172].

2.3.2.2 Cepstral Filtering

Oppenheim *et al.* [173] proposed a single microphone dereverberation approach using cepstral filtering technique in which speech is considered as slowly varying in the cepstral domain with its cepstral components concentrated around the cepstral origin whereas the acoustic impulse response is characterized by pulses with rapid ripples concentrated far away from the cepstral origin. That is, speech articulation has a much higher variation rate than any speaker to receiver impulse response, for which changes can be considered as very slowly varying [173]. The latter depends mostly on the room characteristics. The complex cepstrum of reverberant speech $z(n)$ can be represented as

$$C_z(n) = C_s(n) + C_h(n) \tag{2.13}$$

where $C_h(n)$ is the complex cepstrum of the reverberant impulse response and $C_s(n)$ is the complex cepstrum of the speech signal. Therefore the dereverberation can be achieved by removing the cepstral components corresponding to the impulse response $C_h(n)$ by applying low-time lifter in cepstral domain. An alternative approach also discussed in [173] where a cepstral liftering procedure using a comb filter is considered for reducing the reverberation effect.

The cepstral filtering has been successfully applied to the enhancement of speech degraded by simple echoes [173]. Its use for the enhancement of speech affected by room reverberation poses several practical problems. Typically, frame based processing is used to calculate the cepstrum of a signal. Since reverberation effects are generally much longer than typical frame lengths, the current frame does not contain all the reverberation effects of the frame, while it also contains reverberation effects from previous frames. Moreover, the cepstrum of the clean speech signal $C_s(n)$ and the cepstrum of the acoustic impulse response $C_h(n)$ typically have a large overlap, resulting in signal distortion when using low time liftering. By using an exponential windowing procedure and cepstral averaging in order to identify the room impulse response $h(n)$ before inverse filtering, a significant improvement is possible [161,174]. However, in practice single-microphone cepstrum based techniques for dereverberation have a limited performance.

Additional signal enhancement can be obtained by combining the cepstrum based approach with multi-microphone beamforming techniques as described in [164,175]. The algorithm described in [175], for instance, factors the input signals into a minimum phase and an all-pass component. As the minimum phase components appear to be least affected by the reverberation, the minimum phase cepstra of the different microphone signals are averaged and the resulting signal is further enhanced with a low-time liftering. On the all-pass components, on the other hand, a spatial filtering (beamforming) operation is performed. The beamformer reduces the effect of the reverberation, which acts as uncorrelated additive noise to the all-pass components [176].

2.3.2.3 Temporal Envelope Filtering

Various one microphone algorithms are proposed using modulation transfer function (MTF) of source signals [170, 177–185]. These methods do not require that the impulse response of an environment be measured and use temporal envelope deconvolution through high pass filtering to remove the effect of reverberation. For example, Langhans and Strube [179] proposed an enhancement method for speech signals corrupted by reverberation or noise where they appropriately filtered the envelope

signals in critical frequency bands based on STFT and linear prediction. They used theoretically derived inverse MTF as highpass filtering. Similarly, Aveandano and Hermansky [180] attempt to recover the energy envelope of the original speech by applying theoretically derived inverse MTF and an optimum filter trained from clean and reverberant speech.

Mourjopoulos and Hammond [181] proposed another method to enhance reverberant speech by using multi-band processing for the envelope deconvolution. According to the multi-band envelope convolution method the envelope of the reverberant speech in each frequency band can be approximated by the convolution of the clean speech signal with the envelope of an acoustic impulse response. As such problem of enhancement reduces to the deconvolution of the room response envelope and the reconstruction of the speech signal. Recently, Unoki *et al.* proposed an improved technique based on the MTF concept for restoring the power envelope from a reverberant speech signal [182, 183].

2.3.2.4 LP Residual Enhancement

Yegnanarayana and Murthy developed a reverberant speech enhancement system by manipulating excitation source information (LP residual) based on the residual characteristics of speech [22]. Manipulation of the residual signal is more appropriate than the manipulation of speech signal, especially for short segments, as the residual signal samples are generally less correlated than the speech samples. On the other hand, for manipulation of the speech signal directly, the choice of the size and shape of the window may affect the results significantly. The processing method involves identifying and manipulating the linear prediction residual signal in different regions of the reverberant speech signal, namely, regions in which there is high SRR, low SRR and reverberant component only. A weight function is derived at gross and fine levels to modify the LP residual signal. The gross level identification is done using the entropy of the distribution of the samples in the LP residual signal and the fine level identification is done using the normalized prediction error. The authors also observed that there is a reduction in the flatness of the spectral envelope due to reverberation. Thus the LP coefficients are manipulated to increase the spectral flatness. Finally the enhanced speech signal is resynthesized from the processed LP residual signal and coefficients. In [23] authors proposed a multi-channel speech enhancement technique by exploiting the features of the excitation source in speech production. The most important property is that in voiced excitation the strength of excitation is largest around the instant of glottal closure. The HE of LP residual is used to derive the information of the strength of excitation. A weight function was derived by coherently combining the delay compensated HEs of

the LP residual signals from the different microphones. The enhanced speech was again obtained by exciting the time-varying all-pole filter with the LP residual modified by the weight function.

In [186, 187], the authors presented a spatiotemporal averaging method for the enhancement of reverberant speech. The basis is that the waveform of the LP residual between adjacent larynx-cycles varies slowly, so that each such cycle can be replaced by an average of itself and its nearest neighboring cycle. The averaging results in the suppression of spurious peaks in the LP residual caused by room reverberation. Finally, a speech signal with reduced reverberation is synthesized with the enhanced LP residual. The dynamic programming projected phase-slope algorithm (DYPSA) algorithm [188] is employed for automatic estimation of glottal closure (GC) instants in voiced speech. However no attempt is made to eliminate spurious instants detected in the unvoiced and silence regions by DYPSA algorithm.

Most of the LP residual techniques rely on the important assumption that the calculated LP coefficients of the all-pole filter are unaffected by the multi-path reflections of the room. Gaubitch and Naylor showed that this assumption holds only in a spatially averaged sense [189], and that it cannot be guaranteed at a single point in space for a given room. Recently Gaubitch *et al.* used statistical room acoustic theory for the analysis of the auto regressive (AR) modeling of reverberant speech [190]. They investigated and showed that in terms of spatial expectation, the AR coefficients calculated from reverberant speech are approximately equivalent to those from anechoic speech both in the single channel case and in the case when the coefficients are calculated jointly from an M-channel observation. It is expected that proper calculation of the LP coefficients, i.e., using spatially averaged LP coefficients, improves the quality of LP residual enhancement techniques.

2.3.3 Multi-Stage Algorithms

During last decade several multi-stage algorithms are proposed for enhancement of reverberant speech. In [153], Nakatani and Miyoshi proposed a system capable of blind dereverberation of a one microphone speech by employing the harmonic structure of speech. In this system, a sinusoidal representation is used to approximate the direct sound in the reverberant environments and adaptive harmonic filters are first employed to estimate the voiced clean speech from the reverberant speech signal. This estimation although crude is then used to derive a dereverberation filter. As the number of reverberant speech data sets increases, the estimation of the dereverberation filter becomes more precise. This method, however, requires accurate estimation of the fundamental frequency from the

reverberant speech, and they pointed out that it is difficult to meet this requirement [191]. Wu and Wang [20] proposed a two stage model to enhance reverberant speech. In the first stage, an inverse filter of the room impulse response is estimated, to increase the SRR by maximizing the kurtosis of the LP residual. In the second stage long term reverberation effects are removed by spectral subtraction approach. In [192, 193], a hybrid dereverberation method is proposed that combines correlation based blind deconvolution and modified spectral subtraction to suppress the tail of reverberation and improve the processed speech quality. Inverse filtering reduces the early reflection that constitutes most of the power of the reverberation. Then, the modified spectral subtraction suppresses the tail of the inverse-filtered speech.

In [194], the authors proposed a reverberant speech enhancement algorithm using spatiotemporal and spectral processing. The speech signals are first spatially averaged followed by temporal larynx cycle averaging of LP residual of voiced speech to primarily attenuate the early reverberation. This is followed by spectral subtraction to attenuate the late reverberation. This method takes the advantage of a multi-microphone system for spatial averaging. A similar two-stage single-microphone system is also developed in [195]. In the first stage, the spectral processing technique proposed in [19] is used to suppress late reverberation. In the second stage, the early reflections are suppressed by the LP residual processing in a similar way as in [186].

Furthermore, several methods have been proposed to achieve blind source separation (BSS) of convolutive mixtures, estimating the original signals using only the information of the convolutive mixtures received by the microphones [168,193,196–201]. These BSS methods achieve the enhancement in either temporal or spectral domain.

2.4 Enhancement of Multi-Speaker Speech

Signal separation remains one of the most challenging and compelling problems in auditory perception. In multi-speaker scenario humans have the remarkable ability to selectively focus onto the speech of desired speaker while ignoring speech of other speakers as well as background noise and reverberation. Alternatively, the signal collected by a microphone in such conditions is a mixture of speech signals from several speakers and other degradations. It is a signal processing challenge to separate the speech component corresponding to each speaker, while retaining the quality and intelligibility as much as possible. Processing speech for enhancement in such conditions is a challenging task, as the speech of the other speakers act as noise, against which the speech of the desired speaker needs to be enhanced. The difficulty in achieving this enhancement is due to the similarity of the spectral characteristics of the speech signals from different speakers. Therefore, speaker separation is a very difficult problem, primarily because

(i) The pitch and formants of different talkers may cross or overlap

(ii) The number of talkers is usually not known

(iii) Each talker amplitudes vary within the utterance

In literature, the words separation and enhancement are used interchangeably, which in a larger perspective refers to the goal of enhancing speech in a multi-speaker environment. A significant amount of research is happening across speech and signal processing community to develop methods for processing speech from multi-speaker environment. We have categorized the available speech separation methods into three categories: blind source separation (BSS) using independent component analysis (ICA), computational auditory scene analysis (CASA) and speech specific approaches (SSA). Here the first two categories are well known to the speech processing community. Speech processing in a multi-speaker environment is also attempted by speech processing community with an aim to use available speech-specific knowledge for achieving speech separation. We group them as speech specific approaches. These approaches use speech-specific knowledge like short time spectrum analysis, gross characteristics of excitation (voiced and unvoiced features), cepstrum, fundamental frequency, segmentation and masking in time-frequency planes and also exploiting inherent time structures of sound sources for separation. Similar to noisy speech and reverberant speech, these methods also grouped into temporal or spectral processing methods.

2.4.1 Spectral Processing Methods

2.4.1.1 Speech Specific Approaches

Pearson [24] proposed a harmonic selection method for co-channel speech separation. In this method first the spectral peaks are identified from the windowed mixed speech spectrum. The peaks are accumulated in a table that was used to construct a histogram. The fundamental frequency (F_0) of a first speaker is determined from the histogram and the F_0 of the second speaker was obtained by removing the harmonics belonging to the first speaker from the peak table and repeating the histogram calculation for remaining peaks. The speech of each speaker is then resynthesized by taking IDFT of separated pitch and harmonics. Morgan *et al.* [25] proposed a harmonic enhancement and suppression algorithm for separating two speakers. The idea behind this approach is to recover the stronger talkers speech by enhancing their harmonics and formants given a multi resolution pitch estimate. The weaker talker speech is then obtained from the residual signal created when the harmonics and formants of the stronger talker are suppressed. When both talkers have the same instantaneous pitch, the algorithm will place both talkers on one channel and neither talker on the other channel. When there are more than two talkers in the co-channel signal, only the stronger talker can be separated, and the separation is predicated on that talker always being stronger and voiced. In summary these approaches have taken the harmonic structure of voiced speech as the basis for separation. Voiced speech signals have a periodic nature which can be used as a discriminative feature when speech signals with different periods are mixed. Thus, the primary goal is to develop algorithms that extract the fundamental frequency of the underlying signals [202]. After determining the fundamental frequencies of the underlying signals, the time-frequency cells that lie within the extracted fundamental frequencies or their harmonics are grouped into two speech signals. Several methods have been proposed for estimation pitch. These methods are mainly based on the auto-correlation [203–212] or cepstrum [213, 214] or harmonic structure in the short-time spectrum [215, 216]. Despite extensive efforts by the research community, however, accurate pitch estimation from a sound mixture has proven to be a very difficult task because interference often corrupts target pitch information.

Lee and Childers [217] investigated a minimum cross-entropy spectral analysis which uses the harmonic magnitude suppression technique [218] at the front-end to make initial spectral estimates of each talker, and then minimizes the cross entropy of the two talkers to obtain better spectral separation. In [219], the authors proposed a two-stage scheme similar to [217] using the multi-signal

minimum cross entropy spectral analysis. The use of sinusoidal modeling using least square estimation algorithm to determine the sinusoidal components of each of the talkers has been exploited in [220]. The enhancement is achieved by synthesizing a waveform from the sine waves of desired speaker with the help of *a priori* sine wave frequencies or *a priori* pitch contour and least square estimation technique. Recently, a time domain method to precisely estimate the sinusoidal model parameters of co-channel speech is studied in [221]. Speech separation algorithm based on modeling the complex spectrum of the co-channel speech has been proposed in [222]. The basic requirement of all these methods is that voices to be separated must be periodic. Generally the separation of unvoiced speech is more difficult compared to voiced speech. This is mainly because of two reasons; first, unvoiced speech lacks harmonic structure and is often acoustically noise like. Second, the energy of unvoiced speech is usually much weaker than that of voiced speech. Recently, Radfar *et al.* [202] exploited vocal-tract filter characteristics to separate two voices that has the potential to deal with unvoiced speech. However by the nature of the speech production most of speech produced is of voiced type, and hence nearly all the information is perceived from the voiced sounds itself.

2.4.1.2 CASA Methods

While speech enhancement using signal processing methods with satisfactory performance remains a challenge, the natural ability to enhance sounds of interest selectively by human auditory system inspired researchers to approach this issue in a different way. In 1990, Bregman proposed the concept of auditory scene analysis (ASA) to segregate acoustic signal into streams, which correspond to different sources [223]. A typical ASA system generally consists of two main stages: Segmentation (analysis) and grouping (synthesis). In the first stage, the mixture sound is segmented into the time-frequency cells. Segmentation is performed using either the STFT [224] or the gammatone filter bank [225]. The segments are then grouped based on the cues that are mainly onset, offset, harmonicity, and position cues [226]. This ASA account has inspired a series of computational ASA (CASA) systems for sound segregation [226–232]. A main advantage of CASA is that it does not make strong assumptions about interference. Generally a typical CASA system contains four stages: peripheral analysis, feature extraction, segmentation, and grouping [226]. The peripheral processing decomposes the auditory scene into a time-frequency (T-F) representation via bandpass filtering and time windowing. The second stage extracts auditory features corresponding to ASA cues. In segmentation and grouping, the system generates segments for both target and interference and groups the segments originating

from the target into a target stream. Finally, the waveform of segregated target is synthesized from the target stream [226].

The techniques based on CASA suffer from two problems. First, these techniques are not able to separate unvoiced segments and almost in all reported results one or both underlying signals are fully voiced [233]. Second, the vocal-tract related filter characteristics are not included in the discriminative cues for separation [202]. In other words, in CASA techniques the role of the excitation signal is more important than the vocal tract shape. Another problem with these techniques is that they cannot replicate the entire process performed in the auditory system since the process beyond the auditory nerve is not known well [234]. Attempts are also being made to compare the CASA and BSS based methods for speech separation [235]. It was concluded in this study that a blend of CASA and BSS to take advantage of the merit of each approach may help in improving the performance.

2.4.2 Temporal Processing Methods

2.4.2.1 LP Residual Enhancement

The usefulness of excitation source information for processing speech degraded by background noise and reverberation has been demonstrated in [17, 18, 22, 23]. A method for processing speech from a multi-speaker environment using excitation source information is also proposed by the authors in [9,27]. The speech of each speaker is enhanced with respect to the speech of other by performing the relative emphasis of speech signal around each instant of significant excitation of the desired speaker. The relative emphasis is achieved by giving a larger weight to the LP residual samples in the region around the instants of significant excitation and lower weight to the samples in the other regions [27]. The temporal processing approach proposed in [9, 27] composes of following steps

(i) Identification of instants of significant excitation for determining the short high energy regions corresponding to each speaker

(ii) Classification of extracted instants into two speaker classes

(iii) Weighting the LP residual to enhance the excitation characteristics of desired speaker

(iv) Synthesize the enhanced speech by exciting the time-varying all-pole filter with the LP residual modified by the weight function.

The HE of LP residual is used as a representation for the sequence of impulses corresponding to the instants of significant excitation of the vocal tract system [9,27]. When these sequences are added coherently using the knowledge of the time-delay of each speaker, the strengths of the excitation of the desired speaker are enhanced relative to the strengths of excitation of other speakers. From the coherently added sequence of impulses, a weight function is derived, which is used to derive a modified excitation signal. This modified excitation signal is used to synthesize speech using the vocal-tract system characteristics derived from the degraded speech.

2.4.2.2 Cepstral Processing

Stubbs and Summerfield [236–238] compare the harmonic selection procedure suggested by Parsons [24] with the cepstral transformation of speech. It is known fact that cepstral transformation maps the spectral envelope to a region near the origin of the cepstral domain, and maps the harmonic excitation to a position well separated from the origin and, therefore, away from the cepstral components [239, 240]. For voiced speech the harmonic excitation simply an impulse with cepstrum frequency, or quefrency, equal to the pitch frequency [239]. If the pitch peak in the cepstrum is attenuated, the harmonic excitation is reduced. In [236–238], the authors exploited this fact to attenuate an interfering voice. The success of the filtering operation usually requires one voice to be more intense than the interfering voice [236]. In all pitch based separation methods, speech separation not only depends on the processing method used but also on the nature of the input (i.e., degraded) signal. The more separated in the pitch and harmonics of each talker, the better the result to be expected.

2.4.3 BSS and ICA Methods

BSS is one of the most commonly used approach to estimate original source signals using only the information of mixed signals observed in each input channel, where the independence between the source signals is mainly used for the separation. Typically, mixed signals are acquired by a number of sensors, where each sensor receives a different combination of the source signals. The term *blind* refers to the fact that only the recorded mixtures are known [241]. The early contributory works on the BSS have been performed by considering high-order statistics of the signals as the measurement for independence [242–244]. Later, Comon [245] has clearly defined the term Independent Component Analysis (ICA) and presented an algorithm that measures the independence between the source signals.

ICA performs BSS of statistically independent sources, assuming linear mixing of sources at the

sensors, generally using techniques involving higher-order statistics or temporal decorrelation [246,247]. The basic ICA approach uses the following linear model [248, 249]

$$\mathbf{X} = \mathbf{AS} \tag{2.14}$$

where the vector $\mathbf{S}$ represents m independent sources, the matrix $\mathbf{A}$ represents the linear mixing of the sources, and the vector $\mathbf{X}$ is composed of m observed signals. The idea of ICA is to recover the original sources by assuming that they are statistically independent. The independence assumption means that the joint PDF is the product of the densities for all sources.

$$P(S) = \prod_i p(s_i) \tag{2.15}$$

where $p(s_i)$ is the PDF of source i and $P(S)$ is the joint probability density function.

Denoting the output vector by $\mathbf{V}$, the aim of ICA algorithm is to find a matrix $\mathbf{U}$ to undo the mixing effect. That is, the output will be given by [250]

$$\mathbf{V} = \mathbf{UX} \tag{2.16}$$

where $\mathbf{V}$ is an estimate of the sources. The sources can be exactly recovered if $\mathbf{U}$ is inverse of $\mathbf{A}$.

ICA methods have several drawbacks. Often, it is required that the number of source signals is known in advance and only few have addressed the problem of determining the number of sources in a mixture. Further, standard formulation requires that the number of source signals does not exceed the number of microphones [251]. If the number of sources is greater than the number of mixtures, the mixture is called under-determined (or over-complete). In this case, the independent components cannot be recovered exactly without incorporating additional assumptions, even if the mixing process is known [251].

In BSS research there are two important problems that are generally considered: Instantaneous BSS and convolutive BSS. In the instantaneous BSS case, signals are mixed instantaneously and ICA algorithms can be directly employed to separate the mixtures. However, in a realistic environment, signals are always mixed in convolutive manner because of propagation delay and reverberation effects [252]. Therefore, much research deals with convolutive blind source separation based on extending instantaneous BSS to convolutive case [252].

The origin of the BSS technique in speech signal separation is first attempted by Cardoso [253]

and Jutten [242] using the principle of statistical independence of the sources [241]. Blind separation of multiple speakers is attempted in [254] where the coefficients of finite impulse response (FIR) filters are used to represent the linear mixing of the sources. These algorithms are based on higher order statistics of the signals mutual independence measure among the independent components. Later, numerous approaches have been presented using ICA in BSS for speech separation [247, 255–257]. Also, various methods have been proposed to address the convolutive mixture case [258–262]. Even though variety of algorithms are proposed, all ICA algorithms are fundamentally similar. The main difference between different ICA algorithms is the numerical algorithm used for measuring the signal independence.

BSS using ICA achieves near perfect reconstruction of independent sources in case of synthetic mixture of speech signals. However, when applied to mixture of speech signals collected from real acoustic environments, the performance degrades severally due to the effect of reverberation and background noise. Several methods have been proposed and being proposed in the framework of BSS using ICA to improve the performance in real acoustic environments [255, 261, 263]. Inspite of these sustained efforts the performance is still not satisfactory and there is a belief that using more *a priori* information about speech may help to improve the performance [264, 265].

ICA based algorithms for separation of speech signal have been developed both in time domain and in frequency domain. The time domain approach achieves good separation results, once the algorithm converges. However, these methods suffer from a large computational load to compute convolution of long filters. The frequency domain BSS has a great advantage that the convolution in the time domain becomes multiplication in the frequency domain and it can be easily implemented using FFT with lesser number of computations [252,266]. However, the problems with frequency domain approach are indeterminacy of scaling and permutation [252].

2.5 Summary and Scope for Present Work

In this chapter a brief review of existing speech enhancement techniques is made. In particular, the review is mainly focused from the temporal and spectral processing perspective. Various temporal and spectral processing approaches for the enhancement of noisy speech, reverberant speech and multi-speaker speech are discussed. In summary,

(i) For noisy speech enhancement, most of the spectral processing methods (like spectral subtraction and MMSE estimators) first estimate the spectral characteristics of the background noise and derive the gain function for the noisy speech signal to attenuate the noise spectral components. Therefore the effectiveness of these methods depends on the consistency and accuracy of noise magnitude spectrum estimates. In addition, the noise estimate needs to be continuously updated for non-stationary noise environments. In contrast to spectral processing methods, excitation source information based temporal processing methods first identify the speech-specific features at the gross and fine levels and enhance those features to obtain the enhancement. Hence information about the degradation is not mandatory in the enhancement process. Simplified block diagrams showing the important steps of temporal and spectral processing are shown in Figs. 2.2 and 2.3, respectively. In Fig. 2.3, OLA refers to overlap-add synthesis method.

(ii) In the case of reverberant speech enhancement, spectral processing methods, in particular, spectral subtraction based methods estimate the late reverberant spectral density and subtract it from the reverberant speech spectra to obtain the enhanced signal. Compared to noisy speech enhancement the only difference is that the noise estimation block in Fig. 2.3 is replaced with the late reverberant spectral density estimator. In the same way the approach followed in temporal processing is same as that of noisy speech. However, the speech-specific features used for identifying the gross and fine weight functions may differ.

(iii) In multi-speaker case, spectral based speech separation methods depend on the differences in fundamental frequency characteristics to enhance the spectral features of individual speakers. In temporal processing, speech of each speaker is enhanced with respect to the other by relatively emphasizing the speech around the instants of significant excitation of desired speaker by deriving speaker-specific weight function.

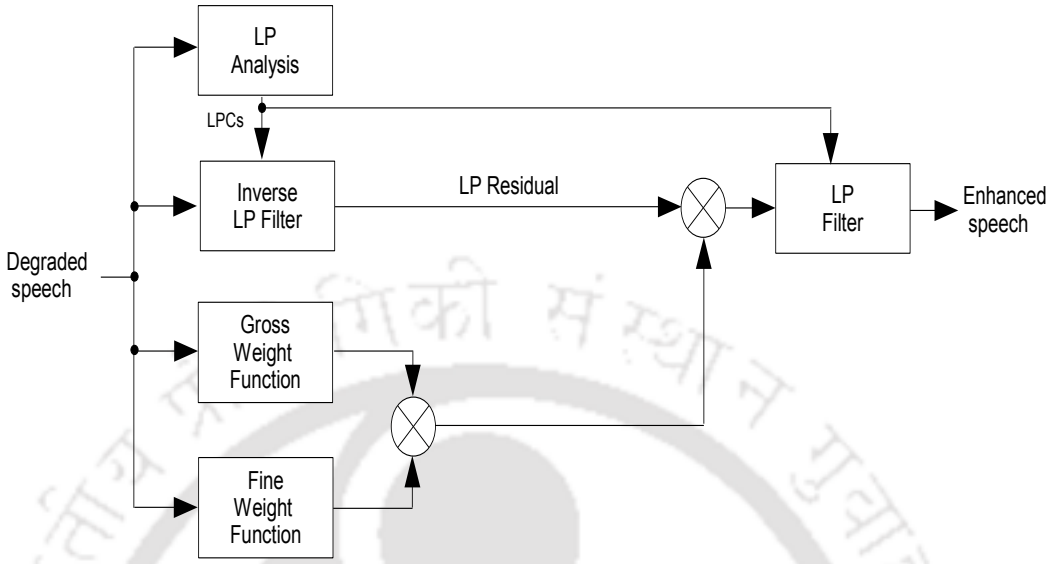


Figure 2.2: Basic block diagram of the temporal processing approach.

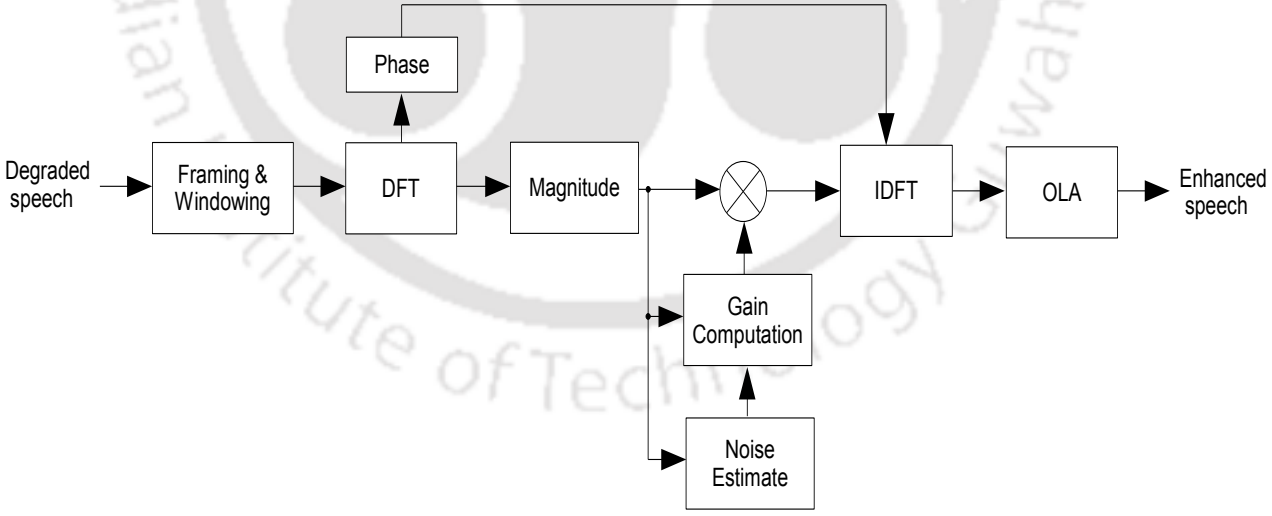


Figure 2.3: Basic block diagram of the spectral processing approach.

As it can be observed from the above discussion, the underlying principle of processing degraded speech is different in each domain of processing. The temporal processing is based on the identification and enhancement of speech-specific features. Alternatively, most of the spectral processing is based on the estimation and elimination of degradation. Further each domain of processing uses independent speech-specific features for processing. That is, excitation source features for temporal processing and vocal-tract based spectral features for spectral processing. Thus it may be possible to combine these two processing approaches for exploiting their independent nature of processing degraded speech. This may result in improved performance compared to any one of them. It may also be possible that one domain of processing may aid an other domain of processing in minimizing the demerit. For instance,

(i) In noisy speech enhancement, the difficulty in estimating degradation in the highly non-stationary environment for spectral processing may be carried out by the gross weight function derived from the temporal processing as a voice activity detector to identify non-speech regions.

(ii) In reverberant speech enhancement, spectral subtraction-based spectral processing methods reduce the late reverberation (i.e., tail of the impulse response) by estimating and subtracting the late reverberant spectrum from the degraded speech spectrum. On the other hand, these methods may not provide better enhancement in high signal to reverberation ratio (SRR) regions, i.e., early reverberant part of the signal. However, the excitation source information-based temporal processing methods mainly enhance the speech-specific features of high SRR regions in the temporal domain.

(iii) In multi-speaker case, spectral based speech separation methods depend on the differences in the fundamental frequency characteristics to enhance the spectral features of individual speakers. The main difficulty is the accurate estimation of the fundamental frequency of desired speaker. The spectral processing preceded by temporal processing may improve the accuracy of the pitch estimation of desired speaker. This is mainly because temporal processing enhances the instants of significant excitation of desired speaker relative to other speakers. As a result, pitch specific cues of interfering speaker(s) will be deemphasized with reference to the desired speaker.

Finally the proposed work would also like to identify and enhance the perceptually significant features in both the domains for improving the perceptual quality of processed speech.

2.6 Organization of the Work

Chapter 3 presents the proposed combined temporal and spectral processing method for the enhancement of noisy speech. The temporal enhancement is achieved by identifying and emphasizing the high SNR regions of noisy speech signal. First, a method for gross level identification of high SNR regions of noisy speech is proposed. Subsequently a method to identify the high SNR regions at the sub-segmental level (1-3 ms) is proposed. These two steps are combined to obtain temporally processed signal. Further, to obtain better noise reduction, the proposed temporal processing method is combined with conventional spectral processing methods like spectral subtraction and MMSE estimators. An additional spectral enhancement method is also proposed to further improve the perceptual quality of the enhanced speech signal. Finally the performance of the proposed method is evaluated using the composite objective quality measures.

Chapter 4 proposes the combined temporal and spectral processing method for the enhancement of reverberant speech. The individual temporal and spectral processing methods are used for reducing early and late reverberation, respectively. In this study, first the late reverberant components are estimated and subtracted using conventional spectral subtraction method proposed in the literature. In the next step, to eliminate the early reverberation a new temporal processing method is proposed by considering speech-specific features at different levels. Finally various experimental studies performed on the spectral processing, gross level temporal processing, fine level temporal processing and the combined processing are reported.

The proposed combined temporal and spectral processing method for two speaker separation is described in Chapter 5. The temporal processing method is proposed to identify and enhance glottal closure instants of desired speaker. The proposed method makes use of the time-delay between the two microphone signals for identification of desired speaker instants. At the second level, the spectral enhancement is performed by estimating the pitch and harmonics of desired speaker from temporally processed speech. Lastly, different objective and subjective measures performed on the proposed method are described.

Apart from the various objective quality measures presented in each chapter, the performance of the proposed methods is also evaluated in the speaker recognition under degraded conditions. First speaker models are created using clean speech data of TIMIT database by the universal background model (UBM)-Gaussian mixture model (GMM) concept. While testing the synthetic degraded signals

are generated for each degradation type and passed through the individual and combined processing methods before the feature extraction step to observe the improvement in the speaker recognition performance. The obtained results are reported in Chapter 6.

The last chapter, that is, Chapter 7 deals with the following: Discussions on the summary of various methods developed in this work are first described. The major contributions of the thesis in developing the combined temporal and spectral processing methods are then mentioned. Finally, possible scopes for the future research directions are mentioned.

3

Combined TSP for Noisy Speech Enhancement

Contents

3.1 Objective of Combined TSP for Noisy Speech Enhancement

The objective of the work presented in this chapter is to develop a combined temporal and spectral processing method for noisy speech enhancement. This is achieved by identifying and enhancing speech-specific features from the noisy speech present both in the temporal and spectral domains. Temporal processing involves identifying and enhancing speech-specific features present at the gross and fine temporal levels. The gross level features are identified by estimating the following speech parameters: sum of the peaks in the discrete Fourier transform (DFT) spectrum, the smoothed Hilbert envelope of the LP residual and the modulation spectrum values from the noisy speech signal. The fine level features are identified using the knowledge of the instants of significant excitation. A weight function is derived from the gross and fine weight functions to obtain the temporally processed speech signal. The temporally processed speech is further subjected to spectral domain processing involving conventional spectral processing and then enhancing speech-specific features in the spectral domain. As indicated by the different experimental studies, the proposed combined temporal and spectral processing method provides better enhancement, compared to either temporal or spectral processing alone.

3.2 Introduction to Noisy Speech Enhancement

The problem of enhancing noisy speech received considerable attention and in the literature variety of methods has been proposed. The noisy speech enhancement methods available may be broadly classified into two categories, namely, *spectral and temporal domain enhancement methods*. The spectral domain enhancement methods attempt to suppress the noise. These include spectral subtraction methods [13, 33–56, 58–60] and minimum mean square error (MMSE) short-time spectral amplitude (STSA) estimator methods [14, 15, 73, 75–78, 80–83]. The temporal domain enhancement methods enhance the characteristics of the speech signal in the time domain. These include linear prediction (LP) residual enhancement [17, 18, 131] and event based analysis methods [9].

Spectral subtraction is a popular noise suppression method for reducing the effect of additive background noise [13]. Spectral subtraction is performed by subtracting the average magnitude of the noise spectrum from the spectrum of the noisy speech to estimate the magnitude of the enhanced speech spectrum. The noise spectrum is estimated by averaging short term magnitude spectra of the non-speech segments. One of the serious drawbacks of this method is that it produces musical noise in

the enhanced speech. This noise arises because of randomly spaced peaks in the time frequency plane due to the deviation of the estimated spectrum of noise from the instantaneous noise spectrum [44]. Several modifications are proposed in the spectral subtraction approach to reduce the effect of musical noise. One of the most commonly used spectral based noisy speech enhancement methods is the MMSE estimator proposed by Ephraim and Malah [14,15]. This estimator is derived based on the assumption that speech and noise may be modeled as independent, zero-mean, Gaussian random variables.

Yegnanarayana *et al.* proposed an enhancement method by exploiting the characteristics of excitation source signal such as LP residual [17]. The basic approach for speech enhancement is to identify the high SNR portions in the noisy speech signal, and enhance those portions relative to the low SNR portions, without causing significant distortion in the enhanced speech. A weight function is derived for the residual signal which will reduce the energy in the low SNR regions relative to the high SNR regions of the noisy signal. The residual signal samples are multiplied with the weight function and the weighted LP residual is used to excite the time-varying all-pole filter derived from the noisy speech to generate the enhanced speech. A multi channel speech enhancement method using the glottal closure (GC) events of speech signal is also proposed using the Hilbert envelope (HE) of the LP residual and the enhancement is achieved by emphasizing the excitation around the GC events [9].

In the noise suppression methods more emphasis is given to suppress the noise components by estimating the noise characteristics from the degraded speech signal. The merit of this approach is the effectiveness for noise removal. While attempting to reduce the effect of noise, these methods introduce their own degradation, more commonly known in the literature as *musical noise*. Alternatively, the objective of temporal domain enhancement methods that use the characteristics of speech is to enhance the speech components by identifying the high signal to noise ratio (SNR) regions, so that the resulting speech is perceived less noisy. The merit of this approach is the effectiveness in enhancing speech-specific features and also they do not require any explicit noise modeling techniques. The limitation of this approach is that the level of noise removal achieved may not be significant as in the case of noise suppression methods.

In general, spectral based noise suppression methods provide better noise suppression compared to the LP residual based methods. But the main limitation is the musical noise. It may not also perform well for highly non-stationary environments. The LP residual methods provide less perceptual distortion like musical noise, but the noise suppression level may be low. This work proposes a method

for enhancement of noisy speech by the *combined temporal and spectral processing* to provide better noise suppression as well as better enhancement in the speech regions. The temporal processing involves identifying and enhancing the speech-specific features in the noisy speech signal present at the gross and fine temporal levels. This is achieved by first identifying the high SNR speech regions at gross level using speech-specific parameters like the sum of ten largest peaks in the DFT spectrum, smoothed Hilbert envelope of the LP residual and modulation spectrum values from the noisy speech signal. The high SNR speech-specific features at the fine level are identified using the knowledge of the instants of significant excitation. A weight function is derived as a result of this process, which is then multiplied with the LP residual of the noisy speech signal to enhance the speech-specific features in the temporal domain. The temporally processed signal is then subjected to spectral processing which involves conventional spectral processing (spectral subtraction or MMSE estimator based methods) and then the proposed spectral enhancement technique.

The rest of the chapter is organized as follows: Section 3.3 discusses the temporal processing of noisy speech signal. The spectral processing of noisy speech signal is described in Section 3.4. Experimental results of the proposed method and the objective studies performed on the experimental results are given in Section 3.5. Section 3.6 gives summary of the work presented in this chapter.

3.3 Temporal Processing of Noisy Speech

Temporal processing refers to the processing of excitation source information in the temporal domain. This involves identification and enhancement of higher SNR regions. The basis for the temporal processing approach is that human beings perceive speech by capturing features present from the high SNR regions and then extrapolating the features in the low SNR regions [17]. Accordingly, the temporal processing involves identifying and enhancing the speech-specific features present at the gross and fine temporal levels. The main objective of the gross level processing is to identify and enhance the speech components at the sound units (100-300) ms level and the objective of the fine level processing is to identify and enhance the speech-specific features at the sub-segmental (2-3) ms level.

The first step in the temporal processing is the extraction of excitation source information from the speech signal. The LP residual, obtained by inverse filtering the speech signal, has long been used as a tool for acquiring the source characteristics of the speech production system [267]. A brief discussion

on the LP analysis is given in Appendix-B. The next step in temporal processing is the identification of instants of significant excitation in the LP residual. The instants of significant excitation correspond to instants of glottal closure or epochs during voiced speech and onset of events like burst, frication and aspiration during unvoiced speech. These instants are quasi-periodic in nature during voiced speech and random in nature during unvoiced speech [268–272]. The instants of significant excitation can be determined by identifying large error regions in the LP residual [271]. This is mainly because the LP residual removes the second order correlations among the samples of the signal, and produces large amplitude fluctuations around the instants of significant excitation. Though the LP residual mostly contains information about the excitation source, there are difficulties in using it directly for further processing. This is due to the phase of the residual which results in either polarity around the instants of significant excitation. The phase alteration in the LP residual produces positive and negative peaks at the instants of significant excitation which makes it difficult to locate the instants directly. Therefore, instead of using the LP residual directly, the Hilbert envelope (HE) of the LP residual can be used [268]. Here, the HE is defined as the magnitude of the analytic signal. The analytic signal is the complex temporal representation of the real signal and is given by [273]

$$\tilde{e}(n) = e(n) + je_h(n) \tag{3.1}$$

where $e_h(n)$ is the Hilbert transform of $e(n)$ [273]. Since the real and imaginary parts of an analytic signal related through the Hilbert transform have positive and negative samples, the HE of the signal is a positive function, giving the envelope of the signal [274]. For example, the HE of an unit sample sequence or its derivative has a peak at the same instant. Thus the properties of HE can be exploited to derive the impulse-like characteristics of the excitation. The HE of the LP residual $e(n)$ is computed as

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{3.2}$$

where

$$e_h(n) = IDFT[E_h(k)] \tag{3.3}$$

and

$$E_h(k) = \begin{cases} -jE(k), & k = 0, 1, ..., \left(\frac{N}{2}\right) - 1 \\ jE(k), & k = \left(\frac{N}{2}\right), \left(\frac{N}{2}\right) + 1, ..., (N-1) \end{cases} \tag{3.4}$$

where $E(k)$ is computed as the DFT of $e(n)$ and N is the number of points used for computing the

DFT [268].

The advantage of using the HE of the LP residual for better identification of impulse-like instants around the instants of significant excitation may be further explained as follows: Let the LP residual $e(n)$ be represented in magnitude and phase form as

$$e(n) = |E(n)| \cos \varphi(n) \tag{3.5}$$

where $|E(n)|$ and $\varphi(n)$ are the magnitude and phase of residual at the given time instant n. The effect of phase $\varphi(n)$ is to introduce bipolar swing to the residual, which leads to ambiguity in locating the impulse-like instants around the instants of significant excitation. For instance, let at time $n = N_1$ the instant of significant excitation occurs and its magnitude is unity i.e., $|E(n)| = 1$. Suppose the phase value at this instant is $\pi/2$ (i.e., $\varphi(n) = \pi/2$), then the residual value will become zero instead of unity. Alternatively, the analytic signal of $e(n)$ represented as

$$x(n) = |E(n)| \cos \varphi(n) + j|E(n)| \sin \varphi(n) \tag{3.6}$$

will have unit magnitude. Since we are considering envelope or magnitude of analytic signal, we will have an unambiguous peak at $n = N_1$ and hence the effect of $\varphi(n)$ is minimized using the HE of the LP residual. For other phase values, both $e(n)$ and $e_h(n)$ will be non-zero, but with a phase shift of $\pi/2$. Hence the effect of phase to reduce the magnitude of peak by one component is compensated by the magnitude of peak by other component.

Apart from the large amplitudes around the instants of significant excitation, the HE also contains a large number of small positive values. The regions around the instants of significant excitation are further emphasized by dividing the square of each sample of the HE by the moving average of the HE, computed over a short window around the sample. The emphasized HE is computed as

$$h_{em}(n) = \frac{h_e^2(n)}{\frac{1}{2M+1} \sum_{p=n-M}^{n+M} h_e(p)} \tag{3.7}$$

where $M = (2 \times F_s/1000)$, $h_{em}(n)$ is the emphasized HE of $h_e(n)$ and F_s is the sampling frequency in Hz. Hereinafter $h_{em}(n)$ is referred as HE of the LP residual. For illustration, segment of voiced speech, corresponding LP residual and HE for a speech signal collected in noisy environment are shown in Figs. 3.1(a)-(c), respectively. The effect of emphasizing regions around the instants of significant excitation is shown in Fig. 3.1(d) for the HE given in Fig. 3.1(c). From the figure it can be observed

that the ambiguity present around the instants in the LP residual is reduced significantly in the HE.

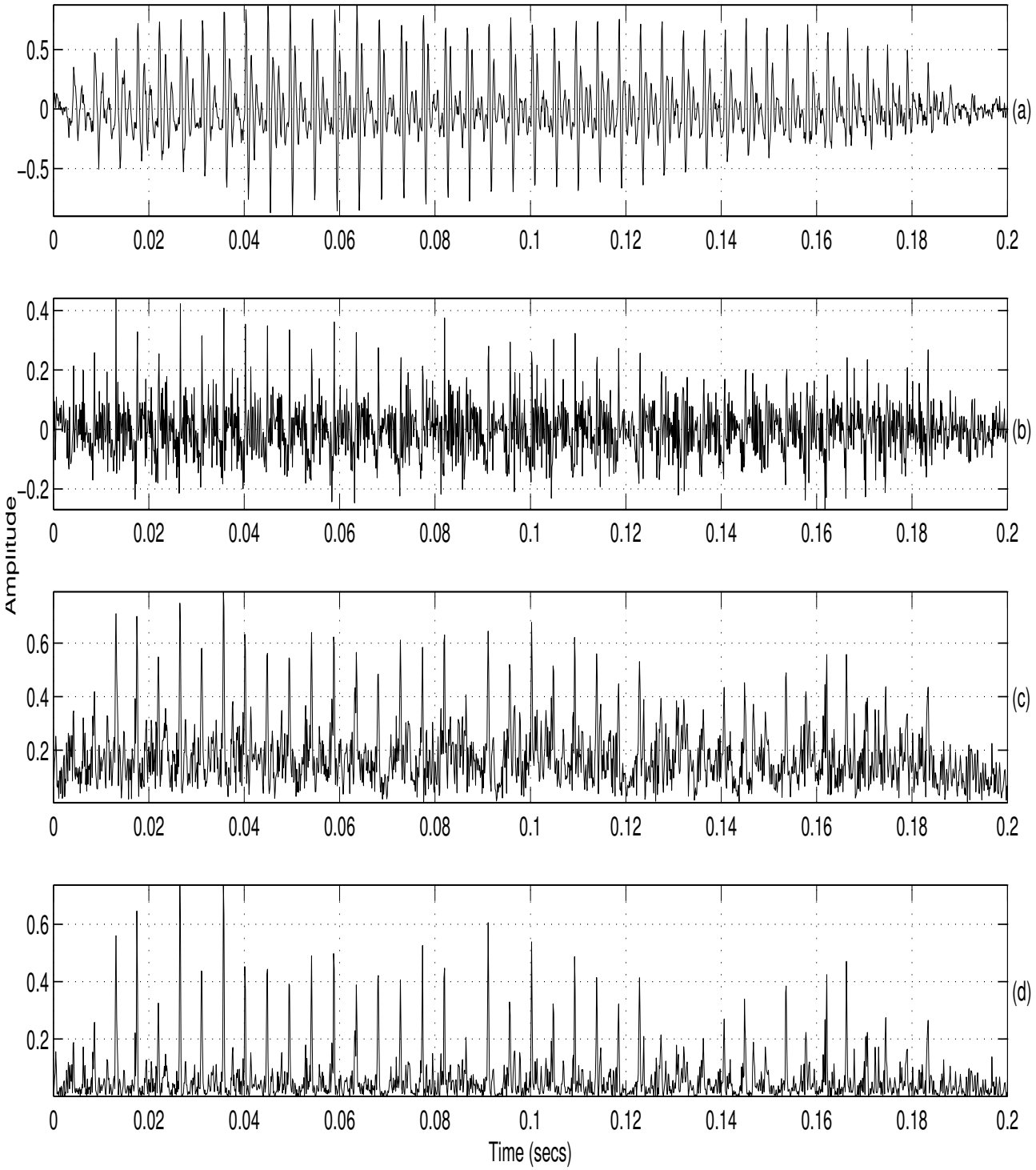


Figure 3.1: Extraction of excitation source information (a) A segment of voiced speech signal, (b) corresponding LP residual, (c) Hilbert envelope (HE) of LP residual, and (d) emphasized HE of the LP residual.

Note that, the HE of the LP residual gives only an approximate location of instants of significant excitation. Therefore hereinafter the term instants of significant excitation refer to ***approximate instants of significant excitation***. From speech enhancement perspective, an approximate location of instants is sufficient. This is because the enhancement is normally performed by emphasizing the residual signal in the regions around the instants of significant excitation.

3.3.1 Gross Level Temporal Processing of Noisy Speech

The high SNR regions at gross level are identified by using the sum of ten largest peaks in the DFT spectrum, smoothed HE of the LP residual and modulation spectrum values of the noisy speech signal. The motivation behind using these three indicators is that they represent different aspects of the speech production mechanism. The sum of the peaks in the DFT spectrum represents predominantly the vocal tract information. The smoothed HE of the LP residual represents predominantly the excitation source information. The modulation spectrum represents the long term (supra segmental) information. Since the origin of these three indicators is different, combining them may improve the robustness and also the detection accuracy as compared to any one of them. Further, the proposed method relies mainly on the characteristics of the speech production process, rather than on the properties of the noise spectrum. Moreover, the proposed method does not assume any noise characteristics, and does not depend on parameters estimated from the noise spectrum. Hence, the proposed method can be applied irrespective of the noise environment. In addition it is reported that the excitation source based methods performs better than the spectrum based methods under babble noise environment. The spectrum based methods perform poorly in babble noise environment because of its speech-like spectral properties. But, the excitation source information and the periodicity of the GC instants are not preserved in the babble noise. However, the excitation source based method gives relatively less performance than the spectrum based methods under white noise environment. This poor performance is due to the limitations of LP analysis under high degradation due to white noise [275]. The modulation spectrum essentially represents the syllable rate. Therefore it is independent of noise environment. However, the weak voiced regions may not be correctly identified using modulation spectrum alone due to the use of long-term window for its computation. Therefore combination of these three features improves the identification accuracy for all noise environments. However a detailed comparison on the proposed gross level features with the recently proposed VAD features needs to be exploited further to validate the performance of proposed features with the existing voice

activity detection methods.

(i) **Sum of Peaks in the DFT Spectrum**

A voiced speech signal is produced by passing a quasi-periodic excitation signal through a linear time varying vocal tract system. The quasi periodic excitation results in a periodic spectrum with peaks at multiples of the pitch frequency. The spectral envelope of the excitation spectrum has an average of -12 dB/octave rolloff for the male speaker and -18 dB/octave rolloff for the female speaker [16]. The vocal-tract system modulates the excitation source by formant frequencies, which depend on the sound unit being generated. Because of the damped sinusoidal nature of the resonance, the formant frequency appears as a broad resonant peak in the frequency domain [276]. As a result the DFT magnitude spectrum of a voiced frame has the same harmonic structure as the excitation source spectrum, but the amplitudes of the harmonics have been shaped according to the frequency response of the vocal tract. The resultant DFT spectrum will have the peaks at pitch and harmonics location and also stronger peaks at formant locations. Hence the sum of amplitudes of the major peak locations will be higher in high SNR regions than other SNR regions. This property is exploited in the identification of high SNR regions of the noisy speech. Mathematically, it is expressed as

$$s_d(l) = \sum_{m=1}^{10} |Y(k_m, l)| \tag{3.8}$$

where l is the frame index, k_m represents the frequency indexes of the largest ten spectral peaks and $Y(k,l)$ represents the DFT of a frame of noisy speech and is computed as

$$Y(k,l) = \sum_{n=0}^{N-1} y(n)w(n-lR)e^{-\frac{j2\pi nk}{N}} \tag{3.9}$$

where $w(n)$ is a Hamming window, N is the number of points used for computing the DFT and R is the frame shift in samples.

Fig. 3.2(a) and (c) show a frame of voiced and silence portion of the speech and the corresponding DFT spectrum values are plotted in Fig. 3.2(b) and (d), respectively. It can be observed that the voiced portion of speech spectrum have dominant peaks at the formants and harmonics location compared to the silence potion. The summed values therefore show large difference in speech and non-speech region. This is further illustrated in Fig. 3.7. For further illustration, the speech data spoken by a female speaker sampled at 8 kHz with a resolution of 16 bits/sample

is taken and a white Gaussian noise is added such that the SNR of the signal is 3 dB and shown in Fig. 3.7(a). The sum of the ten largest peaks of the DFT spectrum of the Hamming windowed signal is calculated using a window of 20 ms duration and 10 ms overlap between the frames. The sum of peaks in the DFT spectrum computed for every frame is repeated 80 times (corresponding to frame shift of 10 ms at F_s= 8 kHz) to make the indicator length equal to that of the speech signal and plotted in Fig. 3.7(b). Note that here the sum of largest peaks is used to obtain the evidence about the vocal tract information for the identification of high SNR regions. The amplitudes of the formants may be estimated by picking some of the largest peaks in the spectrum. Since the objective is only to have gross information about the vocal tract shape, we have used only the ten largest peaks.

(ii) **Smoothed Hilbert Envelope of the LP Residual**

As already mentioned, the high SNR regions around the glottal closure instants can be highlighted by computing the HE of the LP residual. Since the HE of the LP residual signal shows large amplitudes around the instants where the residual error is large, these values are smoothed using a mean filter of 50 ms to smooth out the smaller variations and then used as evidence for high SNR region identification. Fig. 3.7(c) shows the smoothed HE of the LP residual. The residual signal is derived by inverse filtering of the speech signal, and the inverse filter is obtained using LP analysis [267]. Note that to obtain the LP residual signal LP analysis is performed on the speech signal using a 10^{th} order prediction using a frame size of 20 ms with shift of 10 ms. For speech signals sampled at 8 kHz, an LP order in the range 8-16 is found to be most suitable for extracting the LP residual [277].

The mean filter of length greater than pitch period is chosen to smooth out the smaller variations across the pitch periods. As our proposed high SNR evidence enhancement algorithm relies on peaks in the evidence plot (i.e., plot of HE). Therefore smaller variations in the HE (As shown in Fig. 3.3 & 3.4) lead to detection of more number of unwanted spurious peaks. This will degrade the correct detection accuracy of high SNR regions. Therefore mean filter of length 50 ms is experimentally chosen to smooth out smaller variations. However the choice of 50 ms is not critical any filter of length $L \geq 50$ ms may be considered (Fig. 3.5). In order to reduce the computational complexity filter length of 50 ms is selected.

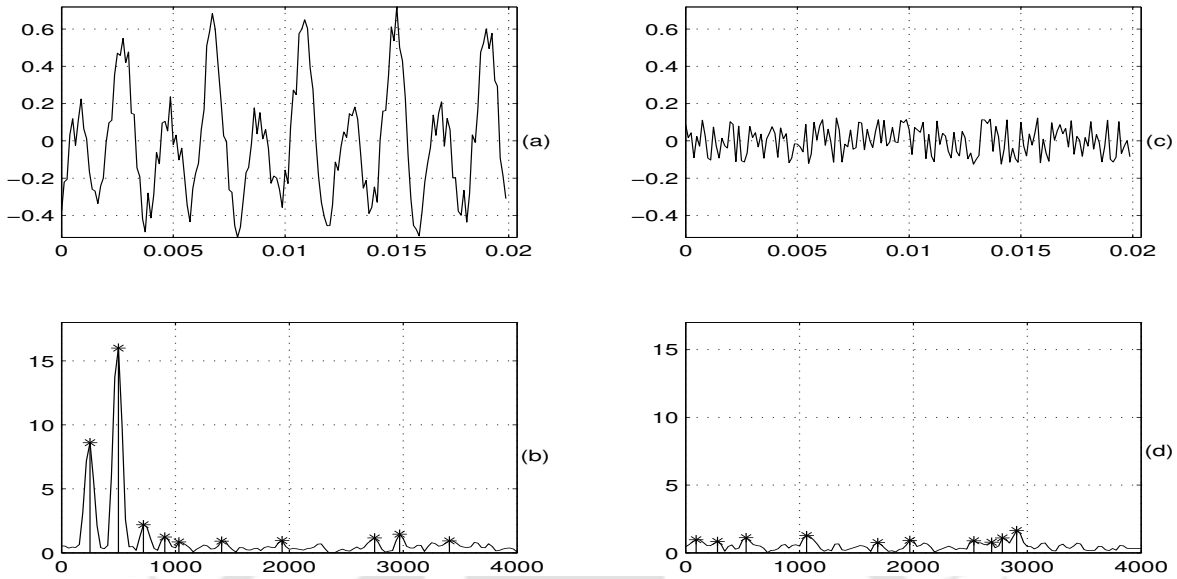


Figure 3.2: Sum of the largest ten peaks of DFT spectrum: (a) a frame of voiced portion of noisy speech, (b) DFT spectrum of signal in (a), (c) a frame of silence portion of noisy speech and (d) DFT spectrum of signal in (c).

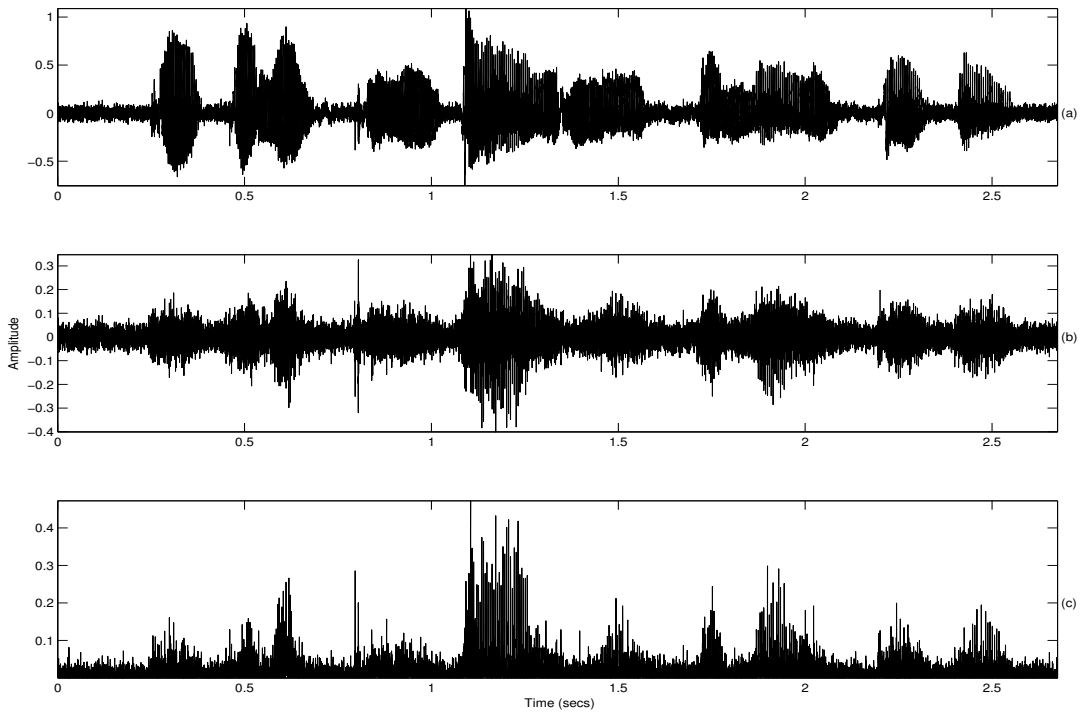


Figure 3.3: HE of the LP residual: (a) degraded speech (SNR = 3 dB), (b) LP residual and (c) Hilbert Envelope (HE).

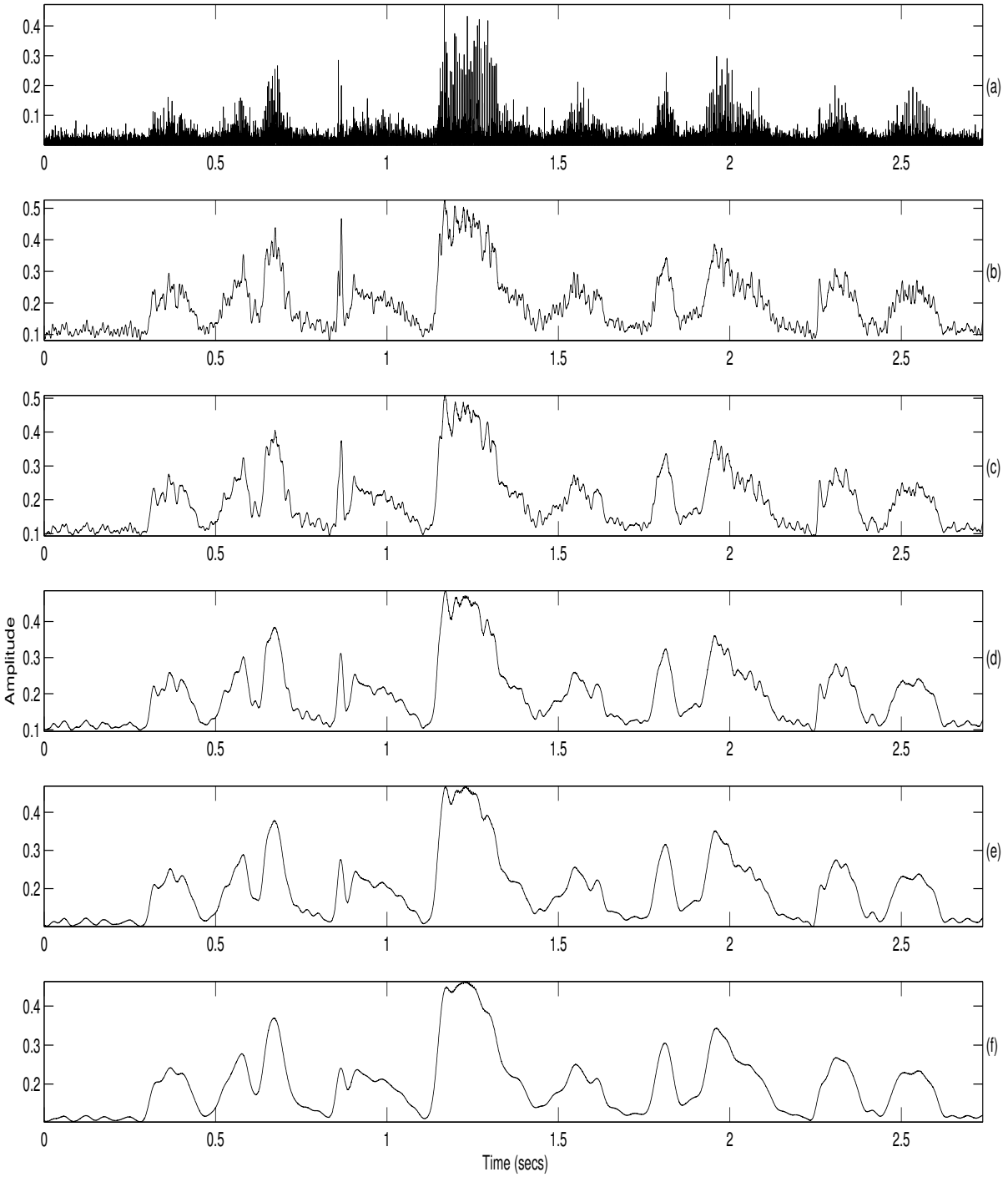


Figure 3.4: Mean smoothed HE of the LP residual: (a) HE of the LP residual, (b)-(f) smoothed HE with a filter length of 10, 20, 30, 40 & 50 ms.

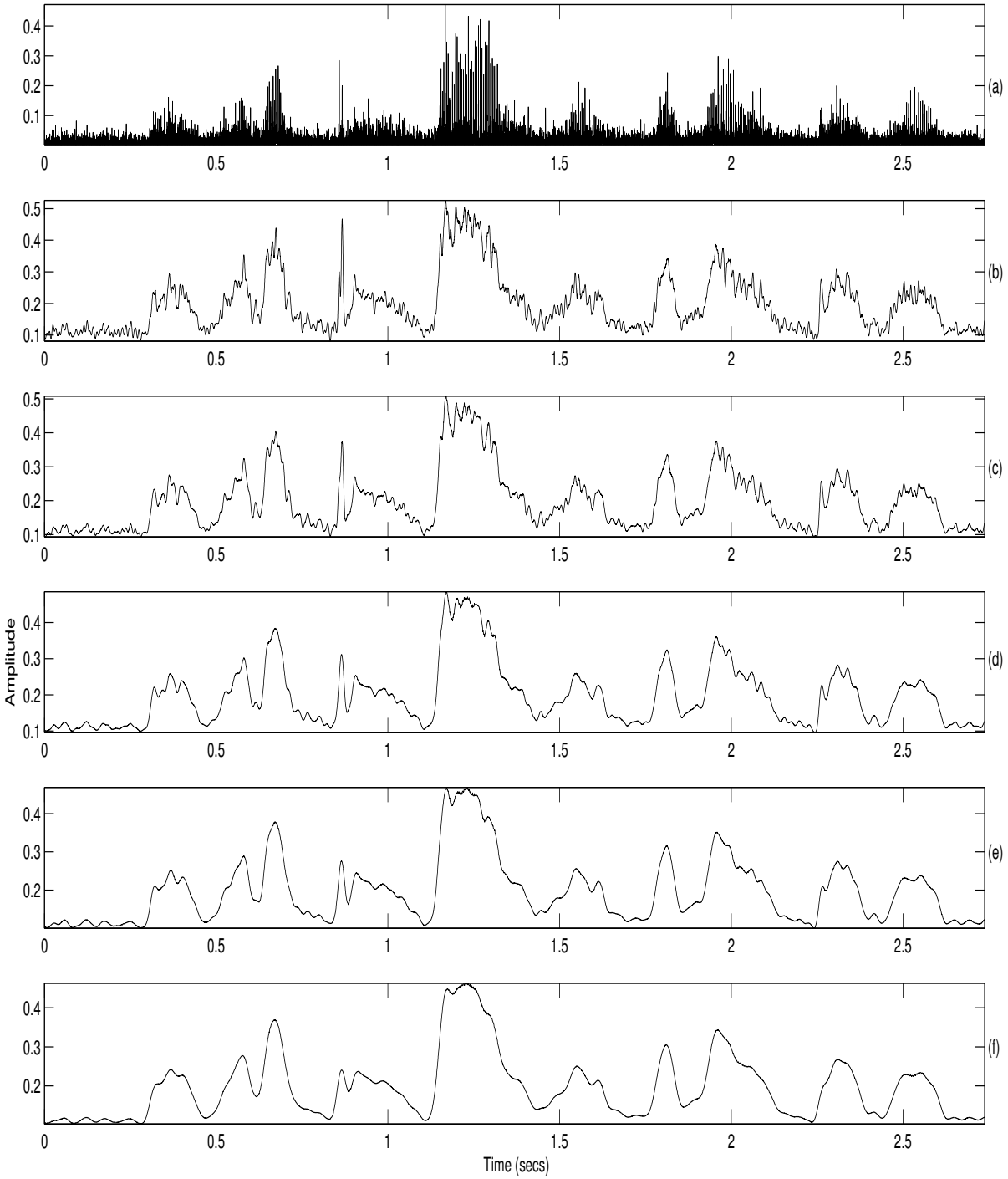


Figure 3.5: Mean smoothed HE of the LP residual: (a) HE of the LP residual, (b)-(f) smoothed HE with a filter length of 60, 70, 80, 90 & 100 ms.

(iii) **Modulation Spectrum**

The modulation spectrum is the spectral representation of the temporal envelope of the speech signal [278]. Dominant components of the modulation spectrum indicate the dominant rate of change of the vocal tract shape. The modulation spectrum of speech is dominated by components between 2 and 16 Hz centered at 4 Hz, which is the most important range for human perception of speech [279–282]. This reflects the syllabic temporal structure of speech. This property is used to identify the high SNR regions of the noisy speech. The modulation spectrum is computed by performing spectral analysis of the analytic signal of the given signal and then normalizing by the average energy of the signal. It is defined as

$$|m(\omega)| = \frac{1}{\langle s(n) \rangle} \left| \sum_{n=-\infty}^{\infty} \tilde{s}(n) e^{-j\omega n} \right| \tag{3.10}$$

where $\tilde{s}(n)$ is the analytic signal and $\langle s(n) \rangle$ is the average value of the signal $s(n)$ [278, 283]. The modulation spectrum of the signal can be calculated using the following steps [278, 279]:

(a) The speech signal is first analyzed into approximately 18 critical band filters. The filters are trapezoidal in shape, and there is minimal overlap between adjacent bands.

(b) In each band, an amplitude envelope signal is computed by half-wave rectification and low pass filtering with a cutoff frequency of 28 Hz and then downsampled to a sampling frequency of 80 Hz (the down sampling factor $D = 100$). The low-pass filtering removes any fine structure components of the speech signal, such as the fundamental frequency of the speaker. The resulting signal will be referred to as the envelope signal. Here, in order to pass the components between [0 - 16] Hz without any attenuation filter cutoff off frequency (half power frequency) of 28 Hz is chosen for a linear phase FIR filter of order ten (Refer to the filter response shown in Fig. 3.6). This choice also not critical. For higher order filter this cutoff frequency may be reduced.

(c) For each critical band the power spectral density (PSD) of the envelope of the signal is estimated using the window length of 250 ms, a Hamming window with an overlap of 12.5 ms.

(d) Since the energies for the frequencies between the 2 - 16 Hz represent important components for the speech signal. The intensity values of the PSD are summed between 2 to 16 Hz for

each critical band and normalized using the total energy.

Mathematically, the modulation transfer function energies is expressed as [284]

$$m(i) = \sum_{j=1}^{18} \left[\sum_{k=k_1}^{k=k_2} \left| \tilde{S}_j(k,i) \right|^2 \right] \tag{3.11}$$

where i is the frame index, j represents the critical band number, k_1 and k_2 represent frequency index of 2 Hz and 16 Hz, respectively. $\hat{S}_j(k)$ is computed as

$$\hat{S}_j(k) = \sum_{n=0}^{N-1} \tilde{s}_j(n) w(n) e^{-\frac{j2\pi nk}{N}}; \qquad j = 1, 2, ..., 18. \tag{3.12}$$

where $\tilde{s}_j(n)$ represents the normalized envelope of j^{th} filter output, $w(n)$ is a Hamming window and N is the number of points used for computing the DFT. Finally the modulation energies of all bands are summed up and the sum of modulation energy components computed for each frame is converted back to the original sampling rate, upsampling by D and is shown in Fig. 3.7(d).

Note that in this work the modulation spectrum is used for identification of high and low SNR regions. Therefore for every frame first the modulation spectral components values in the range of 2 - 16 Hz are computed and summed. The resultant summed value indicates whether that corresponding frame is high SNR or not. Therefore the modulation spectrum values of each frame are plotted as a function of time. Hence in the plot horizontal axis represents time and vertical axis represent modulation spectrum values

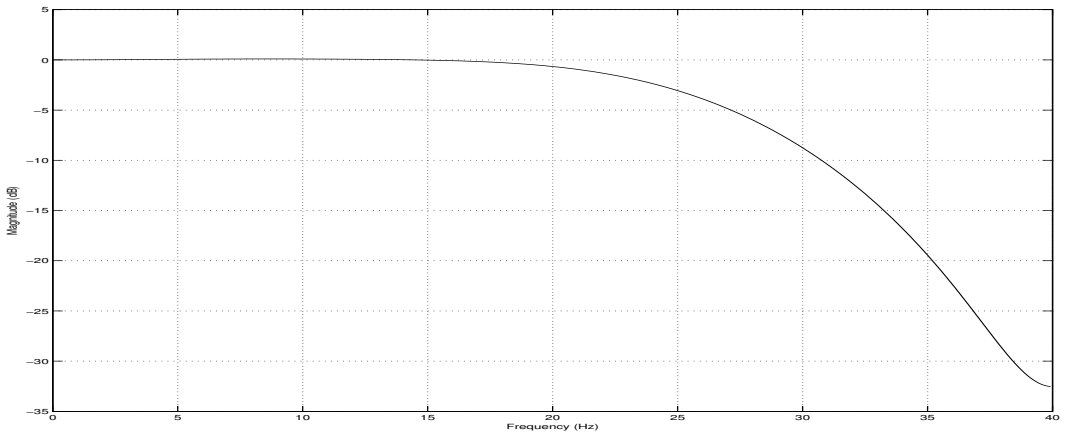


Figure 3.6: Magnitude response of tenth order linear phase FIR filter.

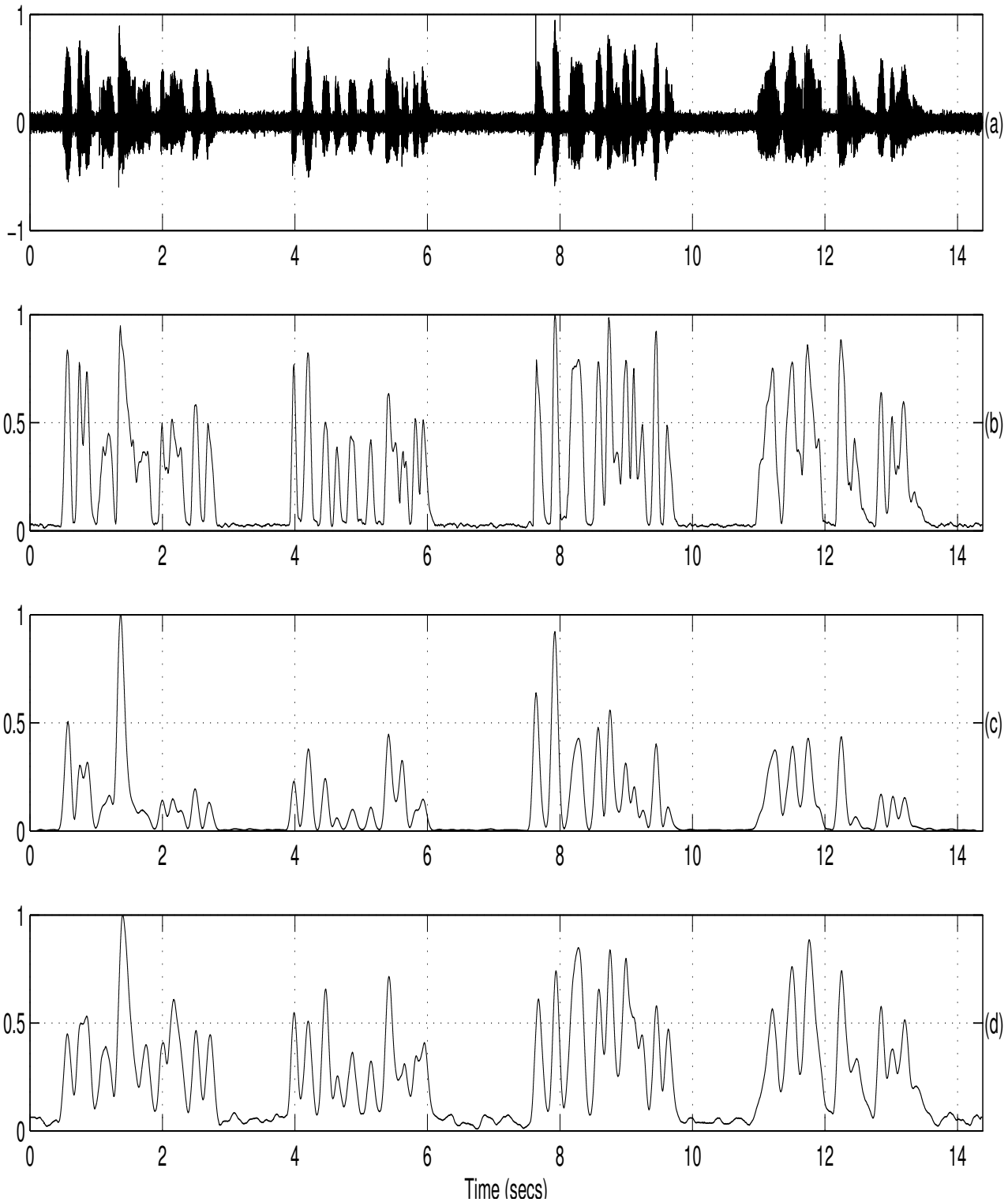


Figure 3.7: Gross level features: (a) noisy speech (SNR = 3 dB), (b) sum of the peaks in the DFT spectrum, (c) smoothed HE of the LP residual and (d) modulation spectrum.

(iv) **Gross Level Weight Function**

Each of the indicators computed above has different information about high SNR regions as is illustrated by their different shape (Figs. 3.7(b)-(d)). They may be combined to get robust evidence. The direct combination may not be very effective due to significant variation in their individual values. In the proposed method, the indicators of the high SNR regions are first enhanced and then combined to identify the gross level features. This is achieved with the help of the first order difference (FOD) of the indicators obtained. The steps involved in the enhancement of high SNR indicators are explained for the sum of peaks in the DFT spectrum with the help of Fig. 3.9. Since FOD represents the slope, the positive to negative going zero transition in FOD locates the peaks in the sum of DFT spectrum values. Fig. 3.9(a) shows first 6 secs duration of the speech signal shown in Fig. 3.7(a). The sum of the DFT spectrum values and its FOD values are shown in Figs. 3.9(b) and (c), respectively. The positive to negative going zero transition points and the corresponding local peaks are represented by a star (*) symbol in Figs. 3.9(b) and (c). The unwanted zero crossings that are detected at the low SNR regions are eliminated by finding the sum of absolute FOD values for a duration of 5 ms on either side with reference to each positive to negative going zero crossing point and are given in Fig. 3.9(d). The peaks with the lower FOD values are eliminated by setting the threshold at 0.5 times the mean value of the FOD. In the next step, if two successive peaks occur within 50 ms then the peak with the lower FOD value is eliminated based on the assumption that it is unlikely that two high SNR regions occur within a 50 ms interval. The star (*) symbols in Fig. 3.9(e) show the peak locations after eliminating the undesirable peaks. With respect to each of these local peaks the nearest negative to positive going zero transition points on either side are identified and are marked by circles in Fig. 3.9(e). The regions between the circles are enhanced by taking the normalized value of that particular region and is shown in Fig. 3.9(f). The same procedure is repeated for both the smoothed HE and the modulation spectrum.

Finally, to derive a gross weight function, the enhanced values of all the three indicators are summed and normalized with respect to the maximum value of the sum. The normalized sum values are then smoothed using a Hamming window of 50 ms and the smoothed values are further

processed using a sigmoid non-linear function given by

$$w_g(n) = \frac{1}{1 + e^{-\lambda(s_i(n)-T)}} \tag{3.13}$$

where λ is the slope parameter set at 20 and $w_g(n)$ is the nonlinearly mapped values of normalized sum $s_i(n)$ and T is the average value of $s_i(n)$. The $w_g(n)$ is termed as the gross weight function. Fig. 3.8 shows the characteristics of sigmoid non-linear function for different values of λ with $T = 0.4$. Note that the value of λ is not very critical, as long as it is in a range which gives desired emphasis and deemphasis. Figs. 3.10(b)-(d) show the sum of peaks in the DFT spectrum, smoothed HE, modulation spectrum values of the speech signal plotted in Fig. 3.10(a) (same as Fig. 3.9(a)) and the corresponding enhanced high SNR indicator plots are shown in Figs. 3.10(f)-(g), respectively. The normalized sum and the nonlinearly mapped values are given in Figs. 3.10(h) and (i), respectively. Figs. 3.11(a)-(d) show the sum of peaks in the DFT spectrum, the smoothed HE, the modulation spectrum values and gross weight function values of the speech data spoken by a male speaker recorded in a real noisy environment, respectively.

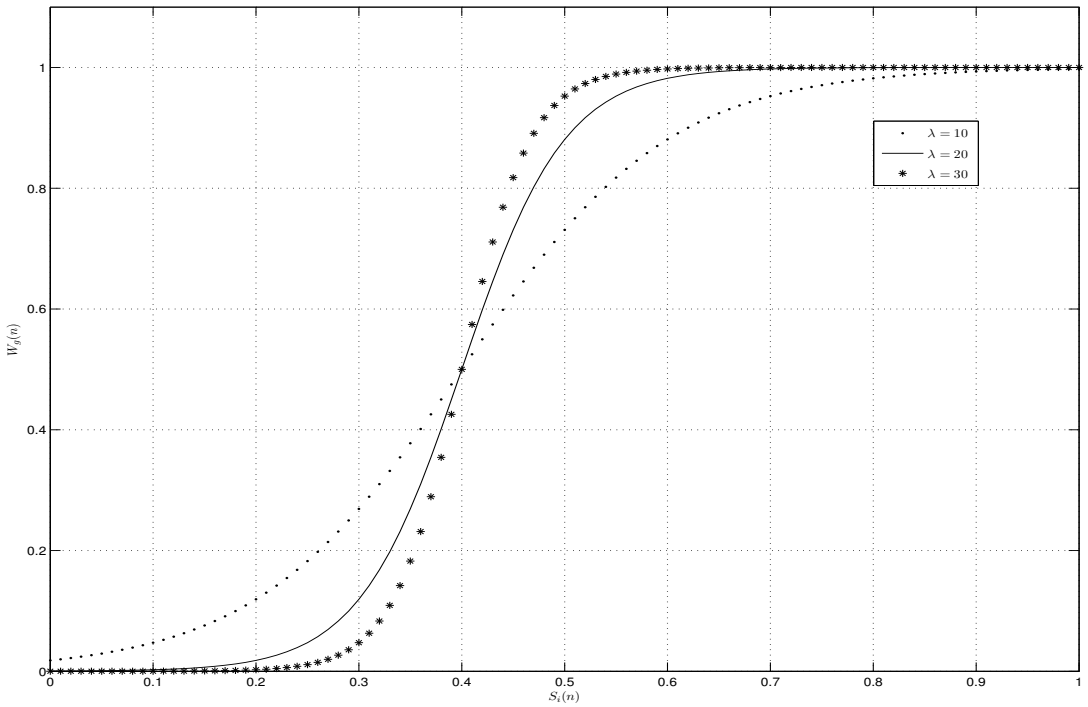


Figure 3.8: Characteristics of sigmoid non-linear function for different values of λ with $T = 0.4$.

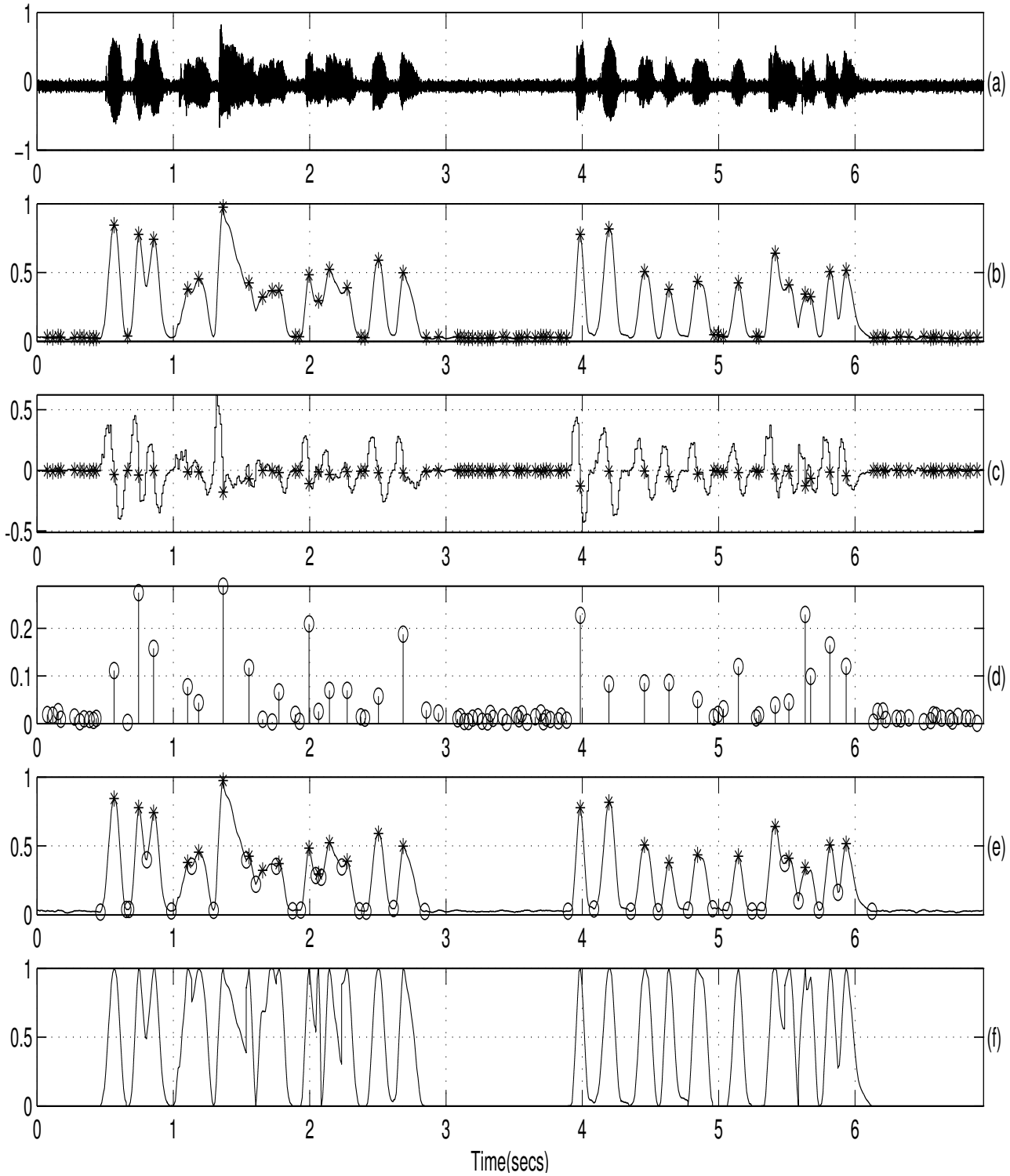


Figure 3.9: High SNR regions enhancement: (a) noisy speech (SNR = 3 dB), (b) normalized sum of peaks in the DFT spectrum, (c) First Order Difference (FOD) values, (d) sum of absolute FOD values computed for a duration of 5 ms on either side with reference to each positive to negative going zero crossing point, (e) sum of peaks in the DFT spectrum and high SNR region locations and (f) enhanced sum of peaks in the DFT spectrum values. In the figures * and o represent the peaks and their boundaries of high SNR regions, respectively.

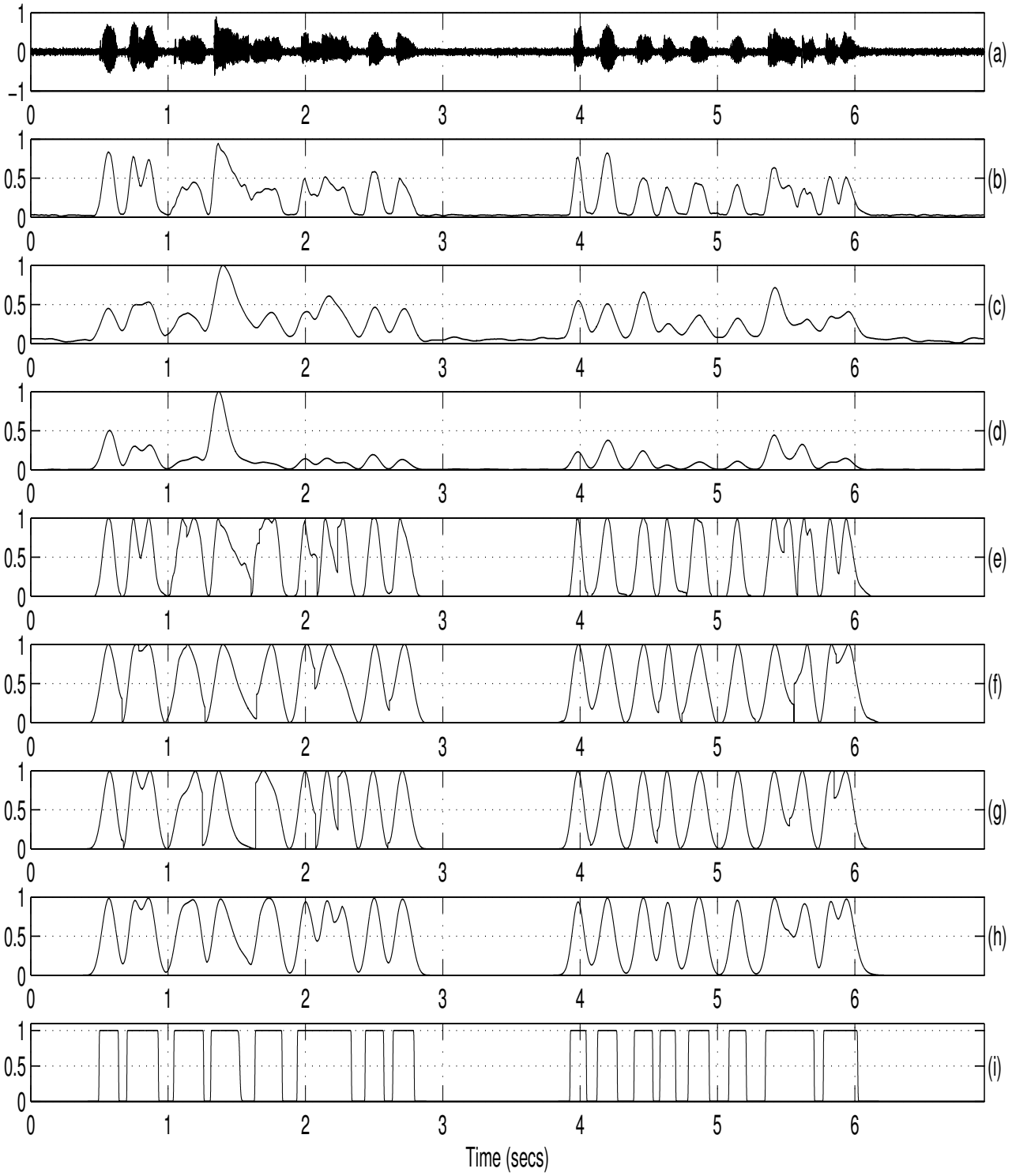


Figure 3.10: Gross level features identification for real noisy speech signal: (a) noisy speech, (b) sum of the peaks in the DFT spectrum, (c) smoothed HE of the LP residual, (d) modulation spectrum, (e) enhanced DFT spectrum values, (f) enhanced smoothed HE values, (g) enhanced modulation spectrum values, (h) normalized sum and (i) nonlinearly mapped values.

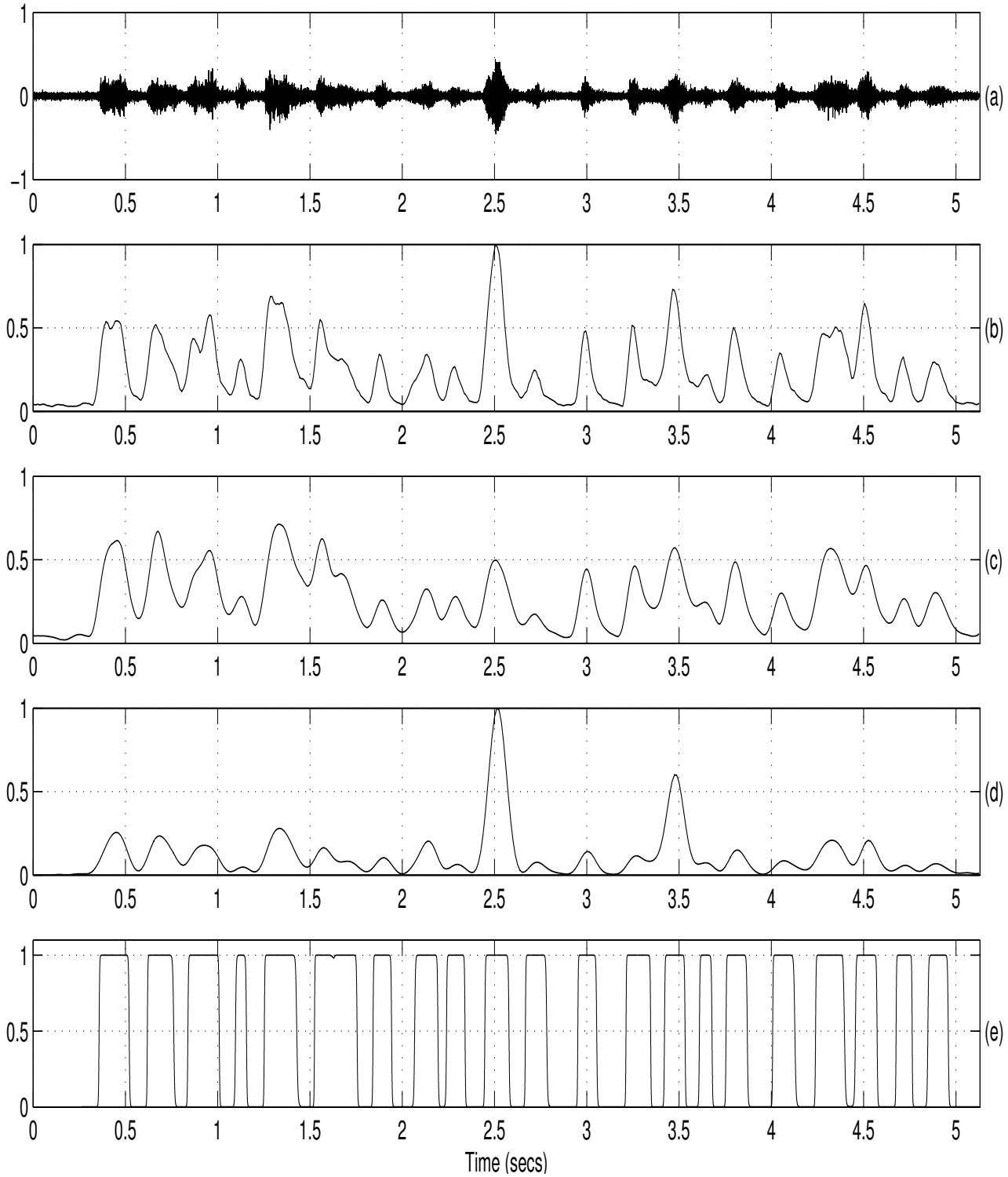


Figure 3.11: Gross level features identification for real noisy speech signal: (a) noisy speech, (b) sum of the peaks in the DFT spectrum, (c) smoothed HE of the LP residual, (d) modulation spectrum and (e) gross weight function.

3.3.2 Fine Level Temporal Processing of Noisy Speech

The basis for the fine level temporal enhancement is that the voiced speech is produced as a result of excitation of quasi periodic glottal pulses and unvoiced speech is produced as a result of excitation of onset of events like burst, frication and aspiration. The significant excitation in each glottal cycle takes place at the instant of glottal closure [268, 269]. The relative spacing between the glottal closure (GC) events is not affected by degradations. Therefore by locating the instants of significant excitation, it is possible to enhance speech around the instants relative to other regions. A weight function is derived for the LP residual from the instants of significant excitation to enhance the excitation source information around these instants relative to other regions. The identification of instants of significant excitation directly from the noisy speech is difficult due to the presence of noise components. If the degraded speech HE envelope is directly used, depending on the magnitude of the peak value, spurious peaks due to the noise components also get detected as the instants of significant excitation. Therefore first the sinusoidal analysis is performed on the noisy speech signal so that most of the noise components get eliminated.

Sinusoidal analysis [285] is performed on the noisy speech signal and only largest 8 peaks are considered for synthesizing the speech signal, so that most of the noise components get eliminated. A detailed description of the sinusoidal analysis is given in Appendix-C. An experiment is conducted on ten different male and female speakers to determine the deviation in the peak locations with respect to their clean speech locations by considering different number of peaks like 4, 8, 16 and 32 peaks per frame. The result of the analysis is given in Table 3.1. In table 4, 8, I6, and 32 correspond to the number of the sinusoidal components used. The percentage values show the ratio of total number of noisy speech peak locations detected at the same locations of clean speech (allowing the frequency deviation of $\pm$ 10 Hz) to the total number of clean speech peak locations. These values are determined by considering only high SNR regions, since the fine weight function is applied only for the high SNR regions of noisy speech. It can be observed that if we consider lower number of peaks per frame, most of the peak locations of degraded speech are not affected with reference to clean speech. Since these peaks mainly represent formants, pitch and its harmonics which has high energy, the effect of noise on these locations will be less. Alternatively, if we consider more number of peaks, then more spurious peaks will be detected from the noisy speech spectra. This shows that as the number of components increase, the peaks in the spectrum are different from that of the clean speech. Therefore only eight

peaks are chosen in this study.

Fig. 3.12(a) shows a voiced portion of degraded speech and the speech signal synthesized by considering eight sinusoidal components is given in Fig. 3.12(b). It can be observed that the high frequency noise variations due to noise are minimized in the reconstructed signal.

Note that even though four peaks gives best performance, we have chosen eight peaks/frame mainly because the LP residual obtained from four sinusoidal components may not contain evidences about all the instants of significant excitation as illustrated in Fig. 3.13. The reason is that, in majority of the cases the largest four peaks represent first two/three formants location, pitch and its first few harmonics location. Therefore the LP residual obtained from pitch and its first few harmonics peaks may not contain the sufficient information related to the instants of significant excitation.

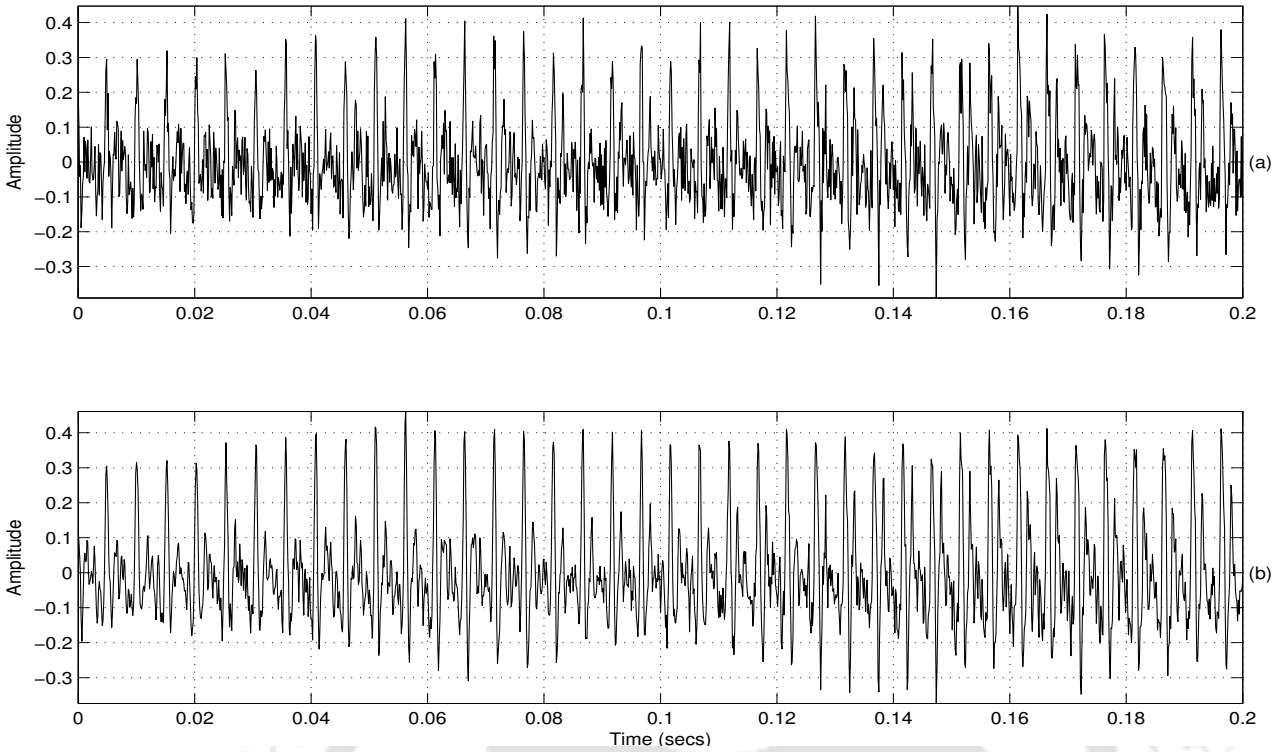


Figure 3.12: Sinusoidal synthesis: (a) 200 ms frame of noisy speech and (b) speech signal synthesized using 8 sinusoidal components.

Table 3.1: Percentage of noisy speech peak locations detected at the same locations of clean speech for different number of peaks per frame.

SNR Level	No. of Peaks/Frame (%)				No. of Peaks/Frame (%)			
	White Noise				Babble Noise			
	4	**8**	**16**	**32**	**4**	**8**	**16**	**32**
0 dB	91.25	78.59	63.28	50.43	93.75	88.91	88.98	80.74
3 dB	93.75	82.97	67.81	53.95	96.25	91.25	90.70	83.24
6 dB	95.00	86.41	72.42	58.48	96.56	93.13	92.34	84.88
9 dB	95.31	89.53	77.11	63.48	97.50	94.53	93.44	86.80
	Factory Noise				Pink Noise			
0 dB	94.06	90.78	88.98	80.39	90.63	82.97	69.53	57.19
3 dB	95.63	93.44	91.17	83.36	94.69	87.03	75.00	62.58
6 dB	96.88	94.69	92.97	85.35	97.19	89.69	79.06	67.19
9 dB	97.19	95.63	94.53	86.76	97.19	92.03	84.45	71.99

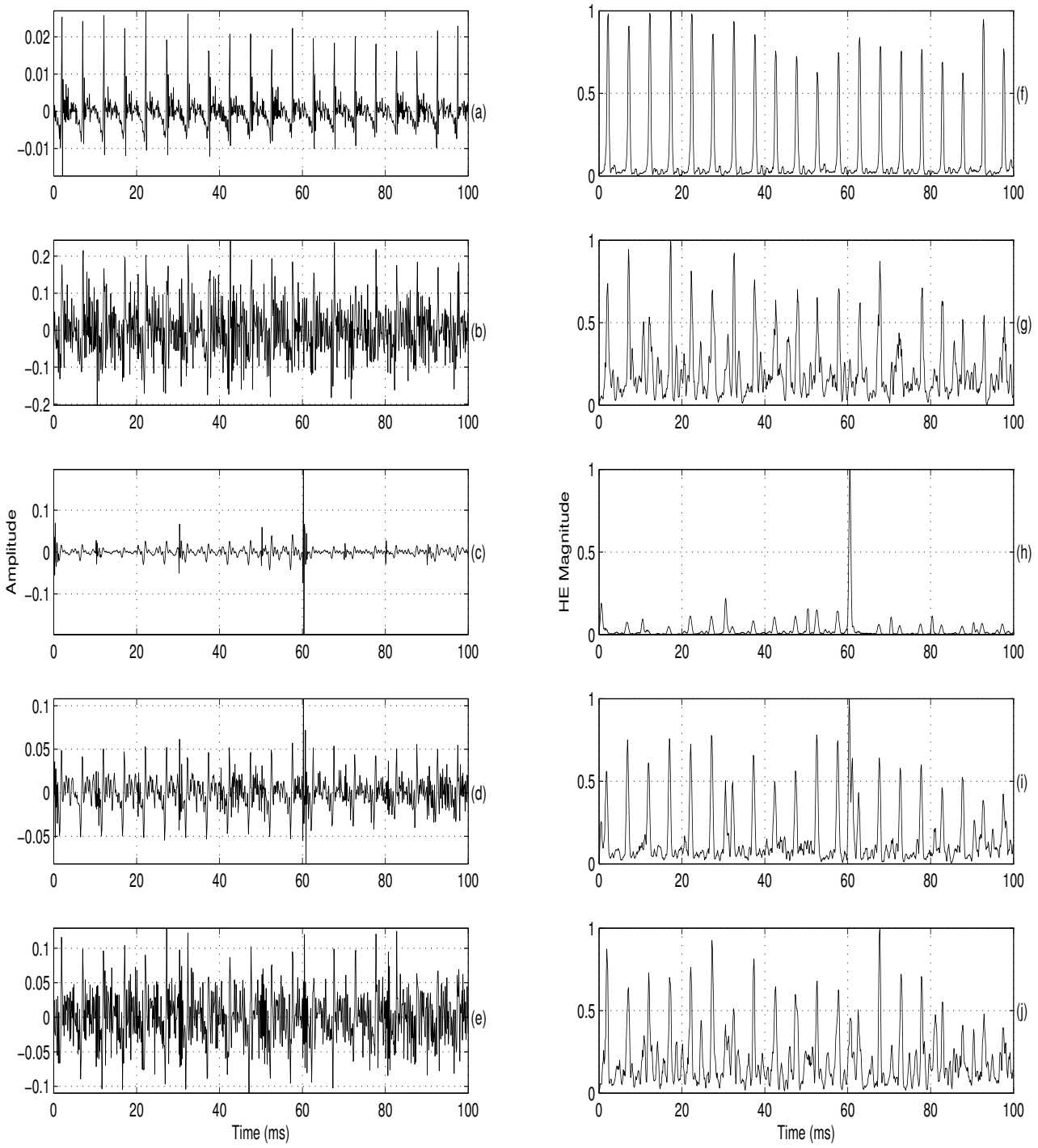


Figure 3.13: LP residual of: (a) clean speech, (b) degraded speech (SNR=3 dB), (c)-(e) speech signal synthesized using 4, 8 & 16 sinusoidal components, (f)-(j) HEs of respective signal shown in Figs. (a)-(e)

In the next step, to determine the approximate locations of the instants of significant excitation, a 10^{th} order LP analysis is performed on the speech signal synthesized from the 8 sinusoidal components, sampled at 8 kHz using frame size of 20 ms and shift of 10 ms to extract the LP residual signal. Then the HE of the LP residual is computed and mean smoothed using 1 ms rectangular window. The peaks in the large error regions, representing the instants of significant excitation are detected using the first order Gaussian differentiator (FOGD) [286]. Because of the anti-symmetric nature of the Gaussian differentiator, it gives a zero-crossing around the peaks in the HE of the LP residual. In discrete-time case the FOGD is defined as [286]

$$g_d(n) = \frac{1}{\sigma\sqrt{2\pi}}\left[e^{-\frac{(n+1)^2}{2\sigma^2}} - e^{-\frac{n^2}{2\sigma^2}}\right], \quad 1 \leq n \leq L_g \tag{3.14}$$

where L_g is length of Gaussian window and σ is standard deviation. FOGD is obtained from a Gaussian window of length $L_g = 80$ samples using $\sigma = 8$ [Fig. 3.14].

The negative of FOGD is convolved with the mean smoothed HE of the LP residual. The zero crossings accompanied by negative to positive transition are detected as the candidates for the instants of significant excitation [286]. It is experimentally verified that, to detect the impulse train of duration T_d using the FOGD, the value of T_d will be $\geq 2.35\sigma$. Since the pitch period lies between 2.5 - 20 ms [287], so the standard deviation of Gaussian window is selected as 8 to detect the events with the minimum interval of 2.5 ms. A fine weight function is derived to enhance the region around the instants of significant excitation by convolving them with the Hamming window of 3 ms duration, since in the LP residual the regions around 3 ms of instants of significant excitation are high energy regions [8]. The minimum value of the fine weight function is kept as 0.4 to reduce the perceptual distortion. A segment of LP residual corresponding to speech signal shown in Fig. 3.12(b), its mean smoothed HE, convolved output with the negative FOGD, the detected instants of significant excitation locations and the fine weight function are shown in Figs. 3.15(a)-(e), respectively.

The final weight function for the noisy speech LP residual is derived by multiplying gross weight function with the fine weight function. The noisy speech residual signal samples are then multiplied with the final weight function. The residual samples are weighted rather than the speech samples mainly because the residual samples are relatively less correlated and hence weighting may lead to less perceptual distortion [17]. The modified residual signal is used to excite the time-varying all-pole filter derived from the noisy speech to generate the enhanced speech which is termed as temporally

processed speech signal. A speech signal synthesized from the 8 sinusoidal components, its LP residual, mean smoothed HE, gross weight function, fine weight function and the final weight function are given in Fig. 3.16(a)-(f), respectively. Figs. 3.17(a)-(c) show the LP residuals of clean speech, noisy speech and enhanced residual signal obtained by multiplying the noisy speech residual signal using a weight function and it shows the enhancement in epoch locations as compared to the degraded speech residual signal.

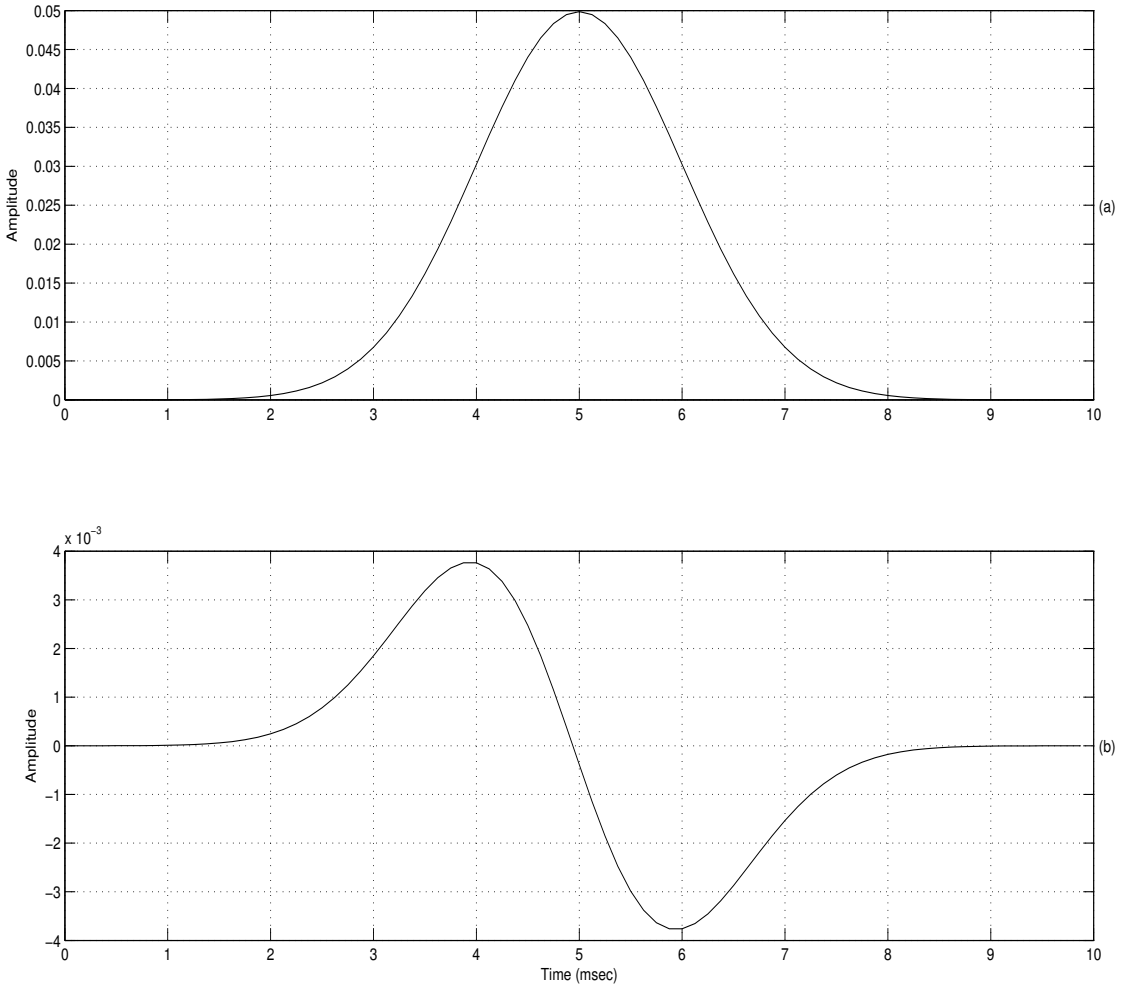


Figure 3.14: First order Gaussian differentiator (FOGD): (a) Gaussian window, (b) FOGD.

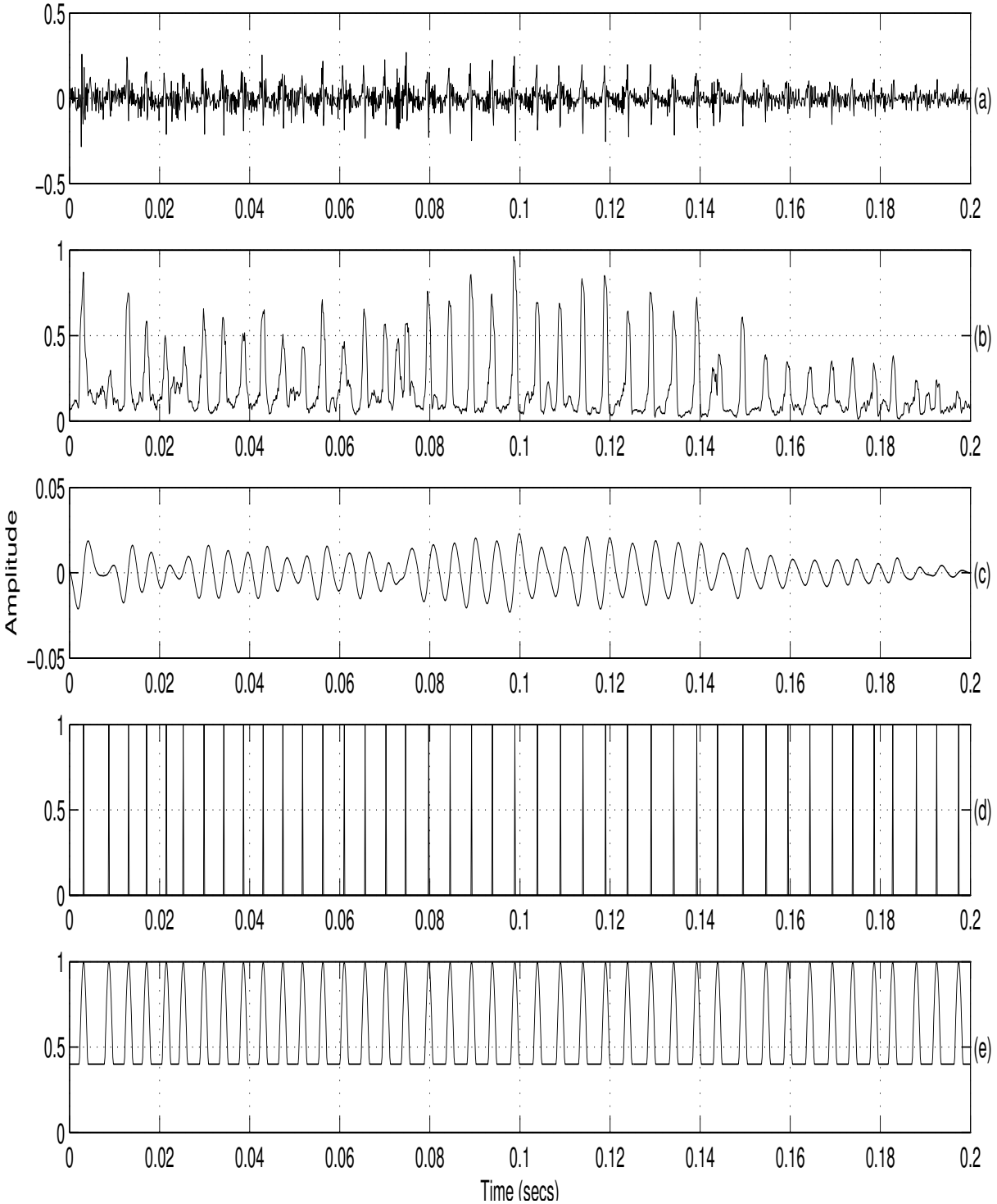


Figure 3.15: Determination of fine weight function: (a) LP residual of speech signal shown in Fig. 3.12(b), (b) mean smoothed HE, (c) convolved output of mean smoothed HE with negative of FOGD operator, (d) instants of significant excitation and (e) fine weight function.

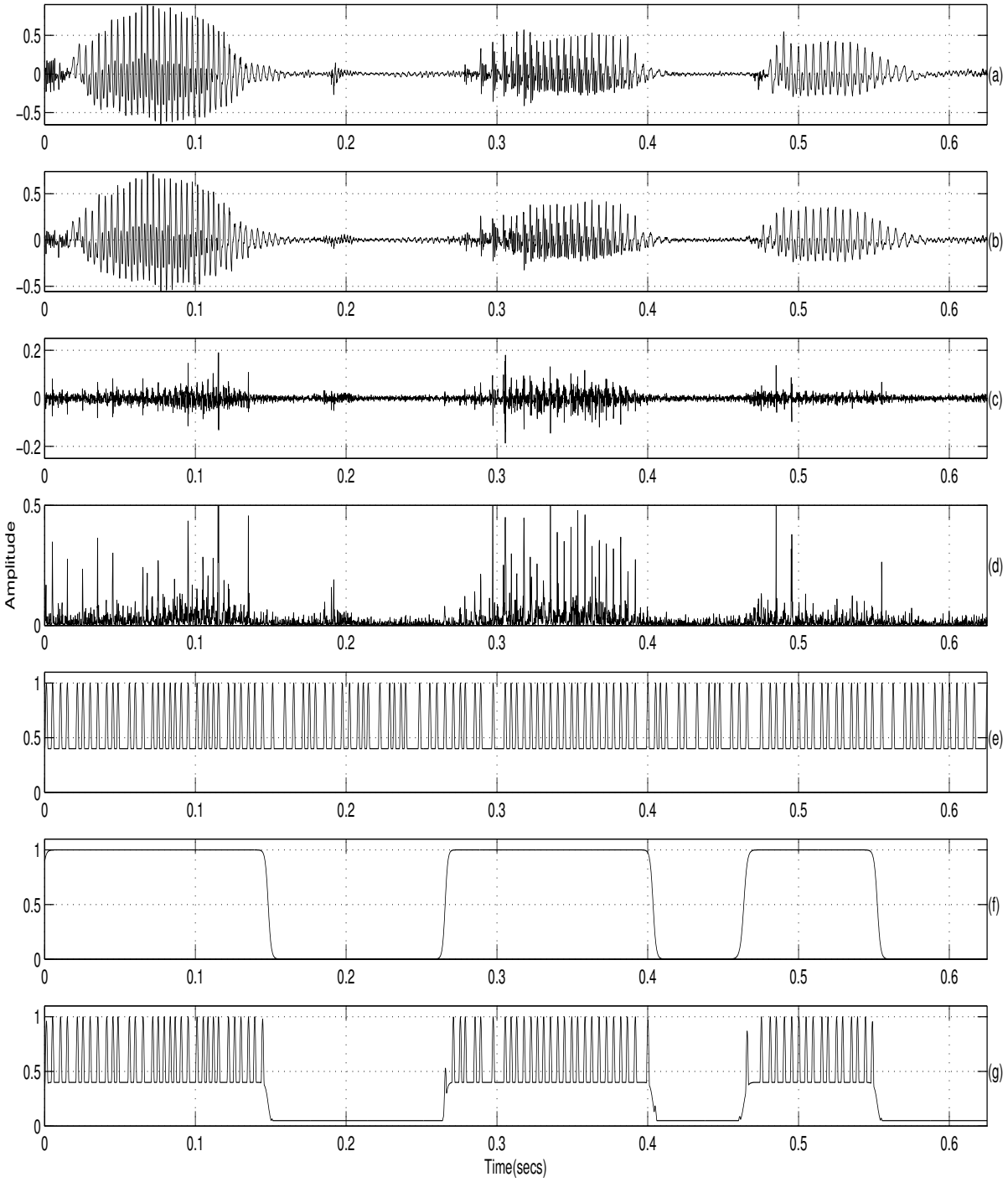


Figure 3.16: LP residual weight function determination: (a) clean speech, (b) speech signal synthesized using 8 sinusoidal components, (c) LP residual of signal shown in (b), (d) mean smoothed HE, (e) fine weight function, (f) gross weight function and (g) final weight function.

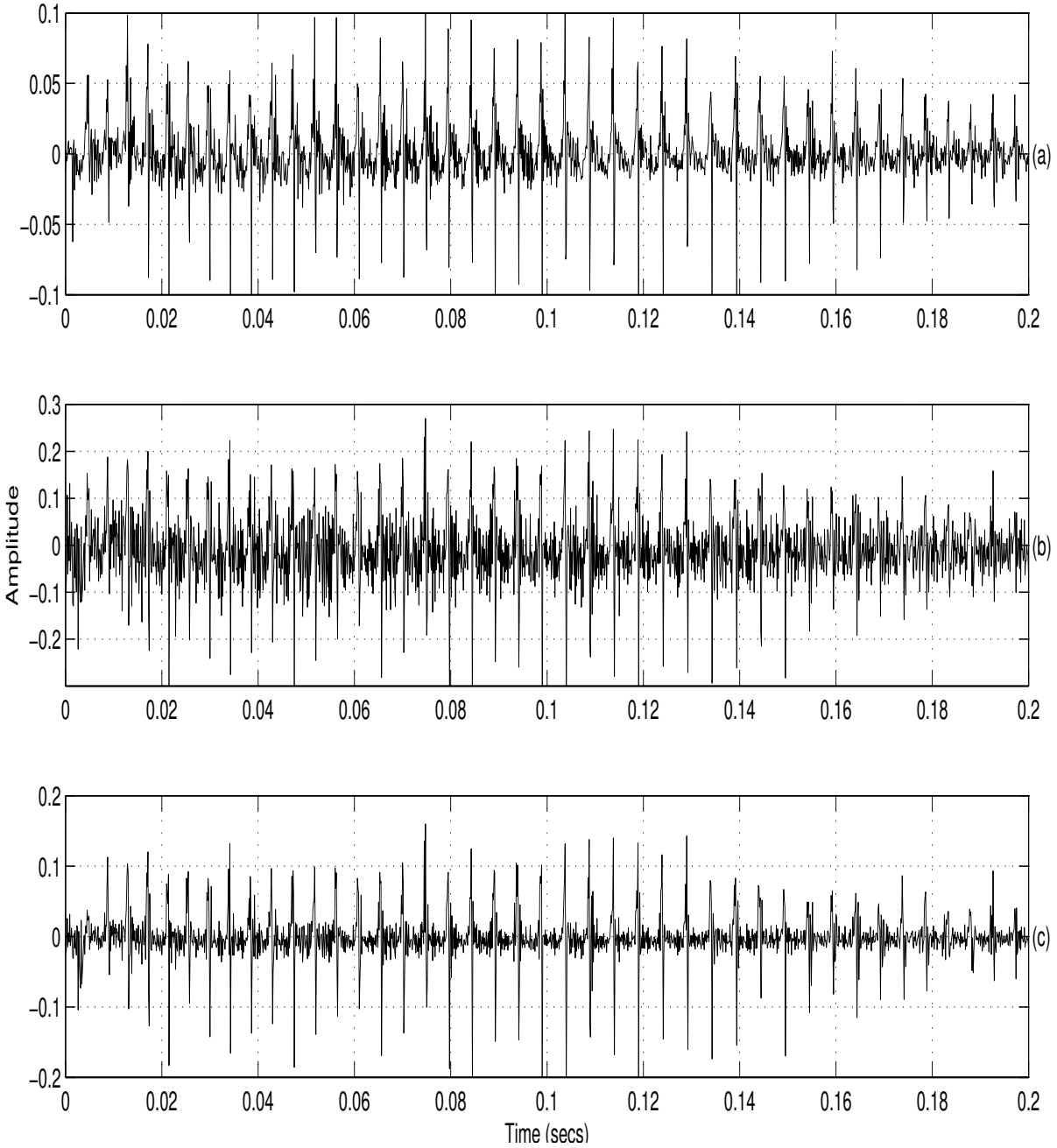


Figure 3.17: LP residual weighting : (a) LP residual of clean speech, (b) LP residual of noisy speech (SNR = 3 dB) and (c) LP residual shown in (b) weighted by a weight function shown in Fig. 3.15(e).

3.4 Spectral Processing of Noisy Speech

Temporal processing method enhances speech-specific features in the temporal domain. This includes high SNR regions at gross level and regions around the instants of significant excitation. This is achieved by multiplying the LP residual of the noisy speech signal by the weight function. Even though speech-specific features are emphasized in the temporally processed speech, the noise suppression may not be significant mainly due to the use of all-pole filters derived from the noisy speech. To further improve the vocal-tract response characteristics at the spectral level and to provide better noise suppression, the spectral processing is performed on the temporally processed speech. This involves conventional spectral processing and then the proposed spectral enhancement technique.

3.4.1 Conventional Spectral Processing Methods

The proposed temporal processing method is combined with four different spectral based speech enhancement algorithms namely, spectral subtraction [13], multi-band spectral subtraction [39], MMSE-STSA estimator [14] and MMSE-LSA estimator [15]. A brief description of these spectral processing algorithms is given below.

(i) **Spectral Subtraction [13]**

Spectral subtraction is performed by subtracting the average magnitude of the noise spectrum from the spectrum of the noisy speech to estimate the magnitude of the enhanced speech spectrum. An estimate of the enhanced speech spectrum $\hat{S}(k)$ is given by [13]

$$\left|\hat{S}(k)\right| = |Y(k)| - \left|\hat{D}(k)\right| \tag{3.15}$$

where $\hat{D}(k)$ is the time-average of the magnitude spectrum of noise calculated during silence. The enhanced speech spectrum obtained will contain some isolated residual noise levels of large variance due to the errors in the estimated noise spectrum. To reduce the auditory effects of this spectral error, magnitude averaging, half wave rectification and residual noise reduction steps are also included to obtain the modified spectrum. The modified spectrum is combined with the phase information of the noisy speech signal to reconstruct the signal in the time domain using the IDFT in conjunction with the overlap add (OLA) method.

(ii) **Multi-band Spectral Subtraction [39]**

The spectrum is first divided into a number of frequency bands, from which the *a posteriori* segmental SNR is estimated in each band. A subtraction factor is derived according to the segmental SNR in each band. The estimate of the clean speech spectrum $\hat{S}_j(k)$ at the frequency bin k and band j is obtained as follows [39]

$$\left|\hat{S}_j(k)\right|^2 = |Y_j(k)|^2 - \alpha_j \delta_j \left|\hat{D}_j(k)\right|^2, b_j \leq k \leq e_j \tag{3.16}$$

where b_j and e_j are the beginning and ending frequency bins of the j^{th} frequency band, α_j is the over-subtraction factor of the j^{th} band and δ_j is a tweaking factor that can be individually set for each frequency band to customize the noise removal properties. A total of eight linearly spaced bands are used in Eqn. (3.16). The band-specific subtraction factor α_j is a piecewise linear function of the segmental SNR of band j and is calculated as follows [39]

$$\alpha_j = \begin{cases} 5, & SNR_j < -5 \\ 4 - \frac{3}{20}(SNR_j), & -5 \leq SNR_j \leq 20 \\ 1, & SNR_j > 20 \end{cases} \tag{3.17}$$

where

$$SNR_j(dB) = 10 \log 10 \left(\frac{\sum_{k=b_j}^{e_j} |Y_j(k)|^2}{\sum_{k=b_j}^{e_j} \left|\hat{D}_j(k)\right|^2} \right) \tag{3.18}$$

The value δ_j is empirically set to [39]

$$\delta_j = \begin{cases} 1, & f_j < 1kHz \\ 2.5, & 1kHz \leq f_j \leq \frac{F_s}{2} - 2kHz \\ 1.5, & f_j > \frac{F_s}{2} - 2kHz. \end{cases} \tag{3.19}$$

The negative values in the enhanced spectrum are floored to the noisy spectrum as [33]

$$\left|\hat{S}_j(k)\right|^2 = \begin{cases} \left|\hat{S}_j(k)\right|^2, & \left|\hat{S}_j(k)\right|^2 > 0 \\ \beta |Y_j(k)|^2, & \text{else} \end{cases} \tag{3.20}$$

where β is the spectral floor factor and is chosen as 0.02.

(iii) **MMSE-STSA Estimator [14]**

Ephraim and Malah proposed this estimator for the short time spectral amplitude component of speech in noise based on a MMSE criterion. An estimate for clean speech is obtained by applying a spectral gain function to the corresponding noisy spectral component. The following equations describe the method [14]

$$\hat{S}(k) = H(k)Y(k) \tag{3.21}$$

$$H(k) = \left(\frac{\sqrt{\pi}}{2}\right)\frac{\sqrt{\nu_k}}{\gamma_k}\exp\left(-\frac{\nu_k}{2}\right)\left[(1+\nu_k)I_0\left(\frac{\nu_k}{2}\right)+\nu_k I_1\left(\frac{\nu_k}{2}\right)\right] \tag{3.22}$$

where $I_0(.)$ and $I_1(.)$ denote the zero and first order modified Bessel functions respectively, ν_k is defined as

$$\nu_k = \frac{\xi_k}{1+\xi_k}\gamma_k \tag{3.23}$$

where ξ_k and γ_k are defined by

$$\xi_k = \frac{\lambda_S(k)}{\lambda_D(k)} \tag{3.24}$$

$$\gamma_k = \frac{|Y(k)|^2}{\lambda_D(k)} \tag{3.25}$$

where $\lambda_S(k) = E\left\{|S(k)|^2\right\}$ and $\lambda_D(k) = E\left\{|\hat{D}(k)|^2\right\}$. The parameters ξ_k and γ_k are interpreted as *a priori SNR* and *a posteriori SNR*, respectively [14].

The *a priori SNR* is calculated as (decision directed approach) [14]

$$\xi_k(l) = \alpha\left(\frac{\left|\hat{S}(k,l-1)\right|^2}{\lambda_D(k,l-1)}\right)+(1-\alpha)P(\gamma_k(l)-1); \quad \alpha = 0.98. \tag{3.26}$$

Here, l is the frame index with

$$P(x) = \begin{cases} x, & x > 0 \\ 0, & otherwise \end{cases} \tag{3.27}$$

The following initial condition is used for the first frame

$$\xi_k(0) = \alpha + (1-\alpha)P(\gamma_k(0)-1). \tag{3.28}$$

(iv) **MMSE-LSA Estimator [15]**

MMSE-LSA estimator minimizes the mean squared error of the logarithmic spectra of the original undisturbed speech signal and the processed output signal. The spectral gain function for the MMSE-LSA estimator is given by [15]

$$H(k) = \frac{\xi_k}{1+\xi_k} \exp\left(\frac{1}{2}\int_{\nu_k}^{\infty} \frac{e^{-x}}{x} dx\right). \tag{3.29}$$

In the above equation, the exponential integral is numerically evaluated using the "expint" function in MATLAB and the values of ξ_k and ν_k are computed as defined before.

3.4.2 Proposed Spectral Enhancement Method

From human perception point of view, the high SNR regions in the temporal domain (instants of significant excitation) and the peaks in the short-time spectrum, specifically, formants play central importance in the perception of speech [8,10]. The temporal processing approach enhances the region around the instants of significant excitation and the subsequent spectral processing suppresses the noise spectral components. To further improve the perceptual quality of the speech, this work proposes the spectral enhancement technique on the high SNR regions of the spectrally processed speech. Here the spectrally processed speech refers to the speech processed by the combined temporal and conventional spectral processing method. Since most of the speech energy is concentrated at the harmonics of fundamental frequency [26], the region around the pitch and harmonic peaks of the spectrally processed speech is enhanced so that the speech components get enhanced. This may tend to reduce the perceptual level of residual noise.

In this work, the pitch period of the high SNR region is determined from the autocorrelation of HE of the temporally processed LP residual [211]. Since the speech is already temporally processed, the estimation of the pitch will be robust. Let $s_t(n)$ be the enhanced speech signal by temporal processing method and $h(n)$ be the HE of LP residual of $s_t(n)$. For each block of 40 ms with shift of 10 ms, the normalized autocorrelation is obtained as [288]

$$R(\tau) = \frac{\sum_{n=0}^{L-1-l} h_m(n) h_m(n+\tau)}{\sum_{n=0}^{L-1} h_m^2(n)}; \quad \tau = 0, 1, 2, ..., L-1 \tag{3.30}$$

where $L = 320$ for $F_s = 8$ kHz and

$$h_m(n) = h(n) - E\{h(n)\} \tag{3.31}$$

where $E\{.\}$ denotes the expected value operator. The first major peak with reference to zero time lag is considered as pitch period of speaker. The autocorrelation methods need at least two pitch periods to detect pitch and hence frame size of 40 ms is chosen.

After obtaining the pitch, its harmonic frequencies are derived from the estimated pitch information. The amplitude spectrum of the desired speech components are constructed by sampling the spectrally processed speech spectrum at pitch and harmonic instants. The pitch and harmonics are sampled using the double-sided exponential function

$$w_d(k) = e^{-\upsilon|k|}; \quad -\frac{L_p}{4} \leq k \leq \frac{L_p}{4} \tag{3.32}$$

where L_p is the frequency index corresponding to the pitch frequency and the value of υ is experimentally determined as 0.5. The sampled spectrum is added with the spectrally processed speech spectrum. The resultant speech spectra is recombined with the original noisy speech phase spectra and converted back to the time domain by an IDFT. The proposed spectral enhancement steps are illustrated in Fig. 3.18 with reference to the spectral subtraction method. Figs. 3.18(a), (b) and (c) show the spectrum of a frame of voiced portion of clean speech, noisy speech and the spectral subtracted speech, respectively. Fig. 3.18(d) shows the window function used for sampling the spectrum derived from the pitch and harmonic locations. The sampled spectrum is added to the spectrally subtracted speech spectrum and is shown in Fig. 3.18(e). Enhanced spectral peaks may be observed at pitch and harmonic instants. Lastly, the various steps involved in the proposed temporal and spectral processing method are illustrated in Fig. 3.19.

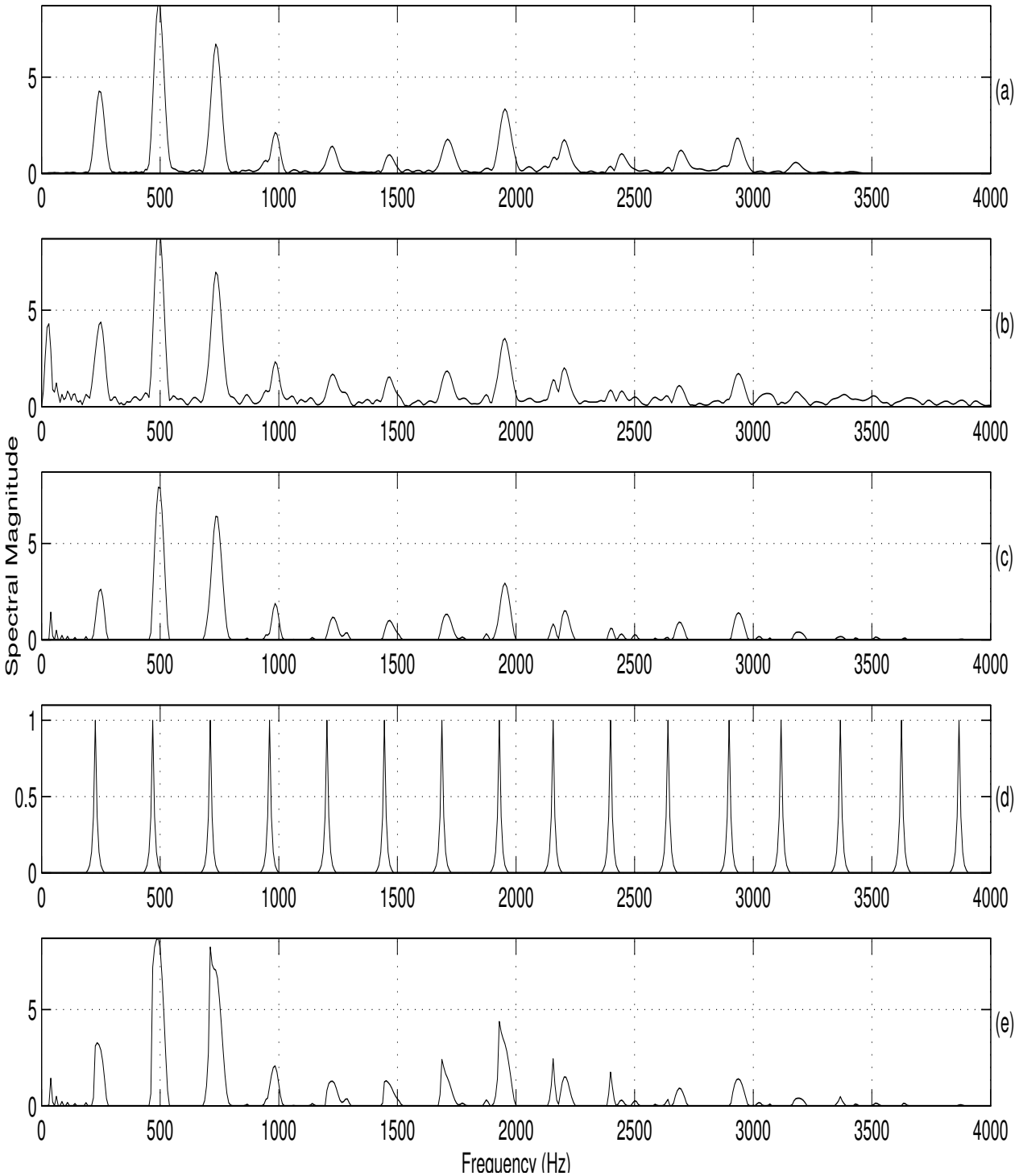


Figure 3.18: Speech Components Enhancement: (a) clean speech spectrum, (b) noisy speech spectrum (SNR = 3 dB), (c) spectral subtracted speech Spectrum, (d) window function for sampling the spectral subtracted speech spectrum and (e) enhanced spectrum.

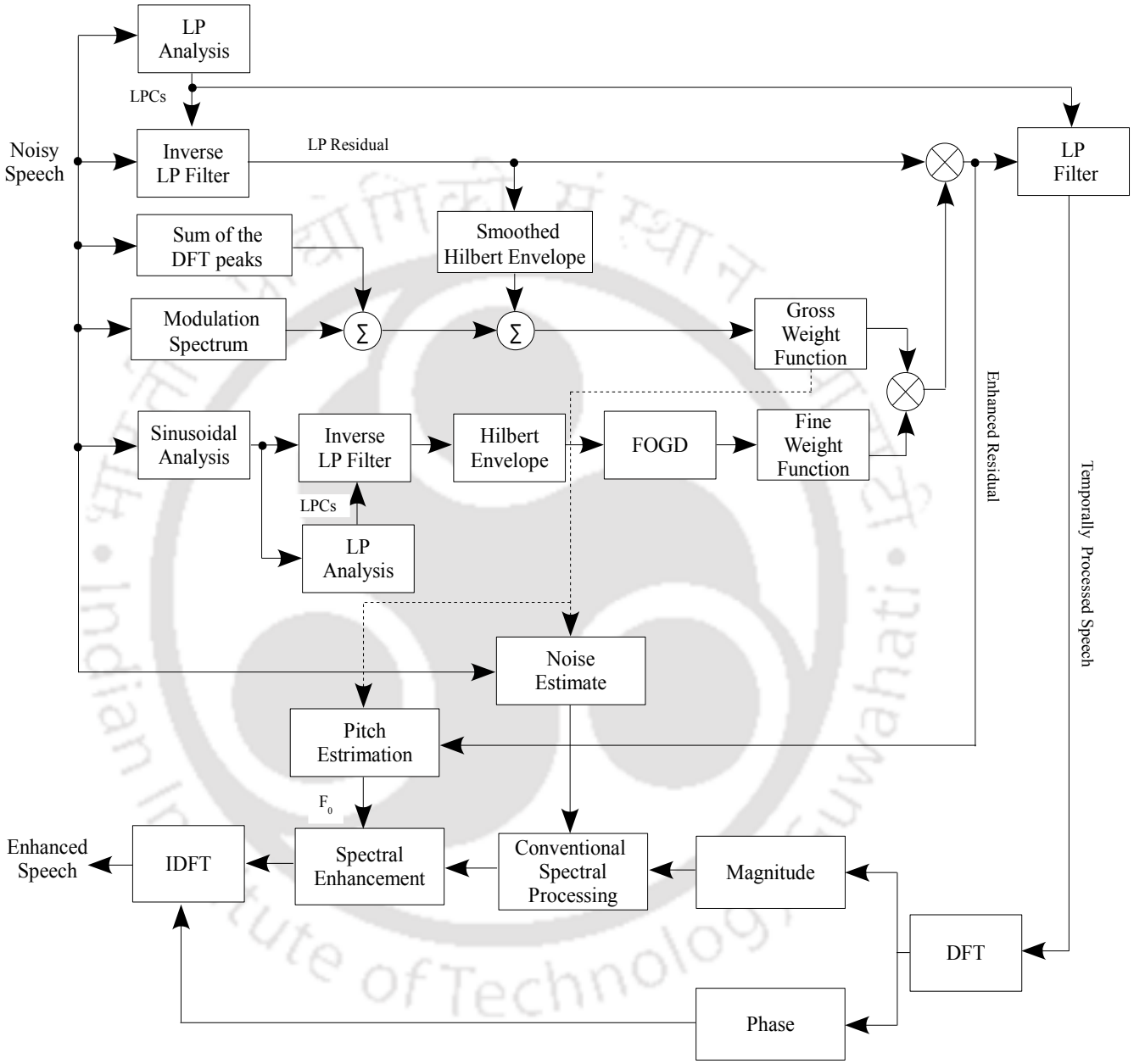


Figure 3.19: Block diagram of the proposed combined TSP method for noisy speech enhancement.

3.5 Experimental Results and Performance Evaluation

Different experiments are carried to evaluate the performance of various stages of the proposed algorithm. For evaluation of the proposed method, ten different samples (five male and five female) from the TIMIT database [289, 290], ten different samples from the NOIZEUS database [291, 292] and the data recorded in the laboratory environment under noisy and noise free conditions are used. The TIMIT sentences are downsampled to 8 kHz before noise is added. The NOIZEUS database contains 30 IEEE sentences produced by three male and three female speakers. Out of this 10 speech samples are randomly selected. The sentences were originally sampled at 25 kHz and downsampled to 8 kHz [291]. Four different noise sources (white Gaussian noise, babble noise, factory noise and pink noise) are taken from NOISEX-92 database [293] and the energy level of the noise is scaled such that the overall SNR of the noisy speech is maintained at 0, 3, 6 and 9 dB.

3.5.1 Time Domain Waveforms and Spectrograms

The speech data spoken by a female speaker is selected and white Gaussian noise is added to make global SNR of the signal as 3 dB and shown in Fig. 3.20(a) (same as last 2.5 sec of the signal given in Fig. 3.7(a)). The degraded signal is processed by the conventional spectral processing and the proposed combined TSP method as described earlier. Fig. 3.20(b)-(d) show the speech processed by the temporal processing, conventional spectral processing and the proposed combined TSP method, respectively. Figs. 3.20 and 3.21 show the comparisons of the speech spectrograms obtained by different enhancement methods. All the speech spectrograms presented in this section used Hamming window of 128 samples with an overlap of 64 samples. The spectrogram of processed signal by the proposed method (Fig. 3.20(j) and Fig. 3.21(j)) shows significant improvement and also noticeable reduction of random peaks compared to conventional spectral processing methods. Fig. 3.22 illustrates the performance of the proposed method at the segmental level. For illustration a voiced segment of 40 ms duration has been chosen. Figs. 3.22(a)-(e) show the HE of LP residual of voiced portion of clean, degraded, and speech processed by the different processing methods. Similarly, the vocal-tract (LP) spectrum of clean, degraded and speech proceed by the individual and combined processing methods are plotted in Figs. 3.22(f)-(j). It can be observed that the combined processing shows improvement in both the excitation source signal and vocal-tract spectrum, whereas individual processing methods show major improvement either at the excitation source signal or at the vocal-tract spectrum only.

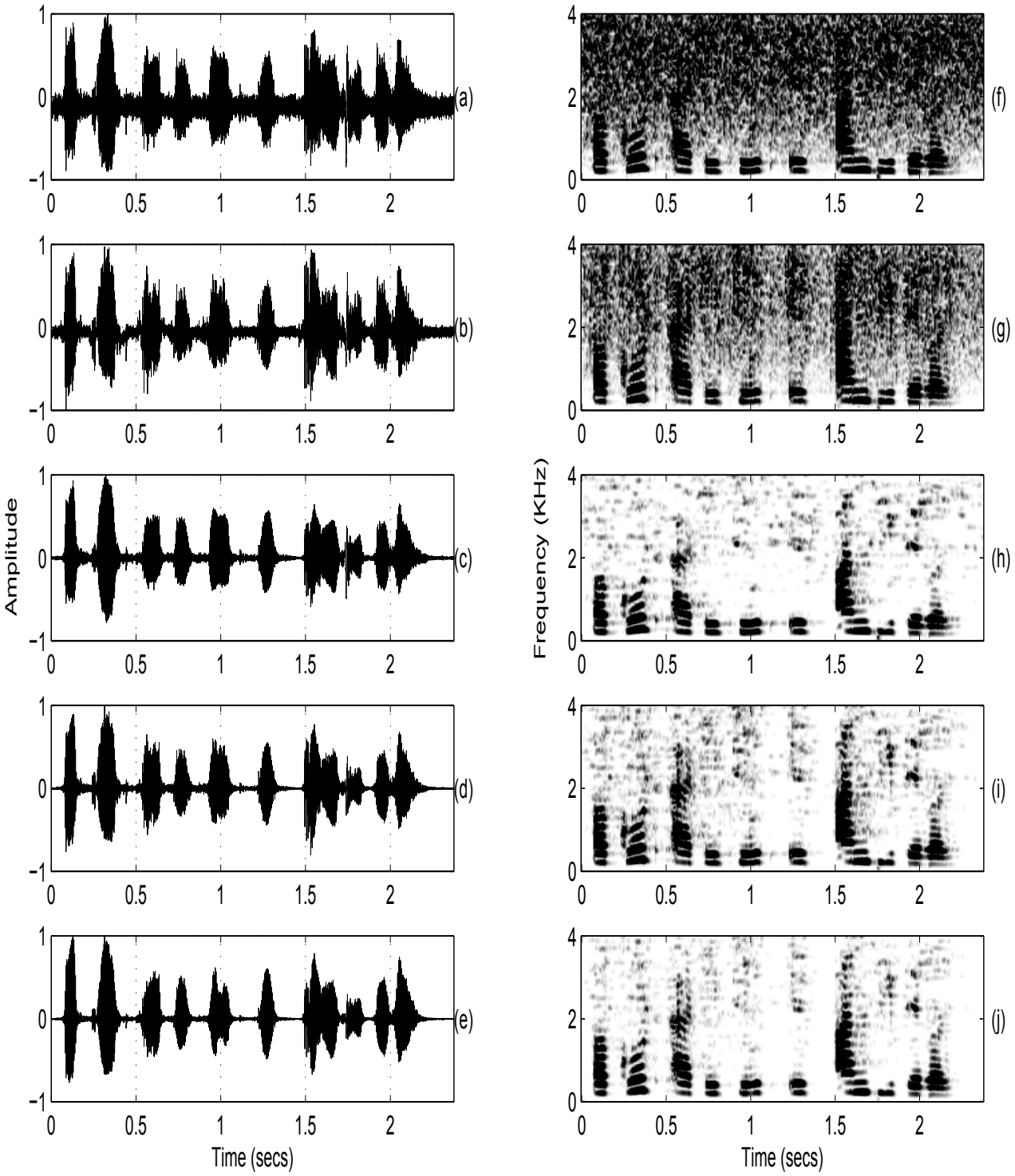


Figure 3.20: Results of enhancement of noisy speech by temporal and multi-band spectral subtraction: (a) degraded speech (SNR = 3 dB), (b) speech processed by temporal processing, (c) speech processed by spectral processing (multi-band spectral subtraction), (d) speech processed by temporal and spectral processing (e) speech processed by temporal and spectral processing with spectral enhancement and (f)-(j) spectrograms of the respective signals shown in (a)-(e).

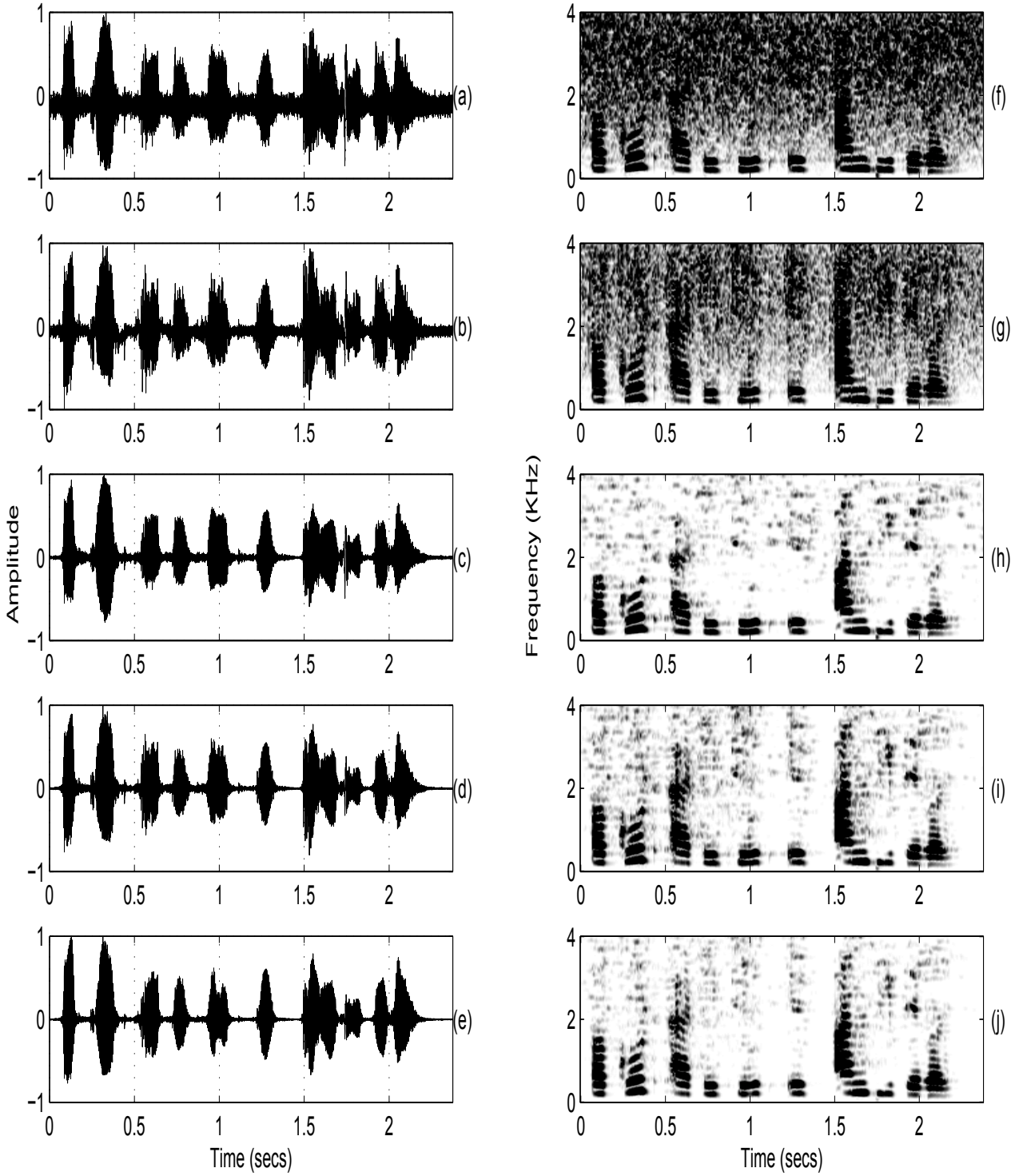


Figure 3.21: Results of enhancement of noisy speech by temporal and MMSE-STSA estimator: (a) degraded speech (SNR = 3 dB), (b) speech processed by temporal processing, (c) speech processed by spectral processing (MMSE-STSA estimator), (d) speech processed by temporal and spectral processing (e) speech processed by temporal and spectral processing with spectral enhancement and (f)-(j) spectrograms of the respective signals shown in (a)-(e).

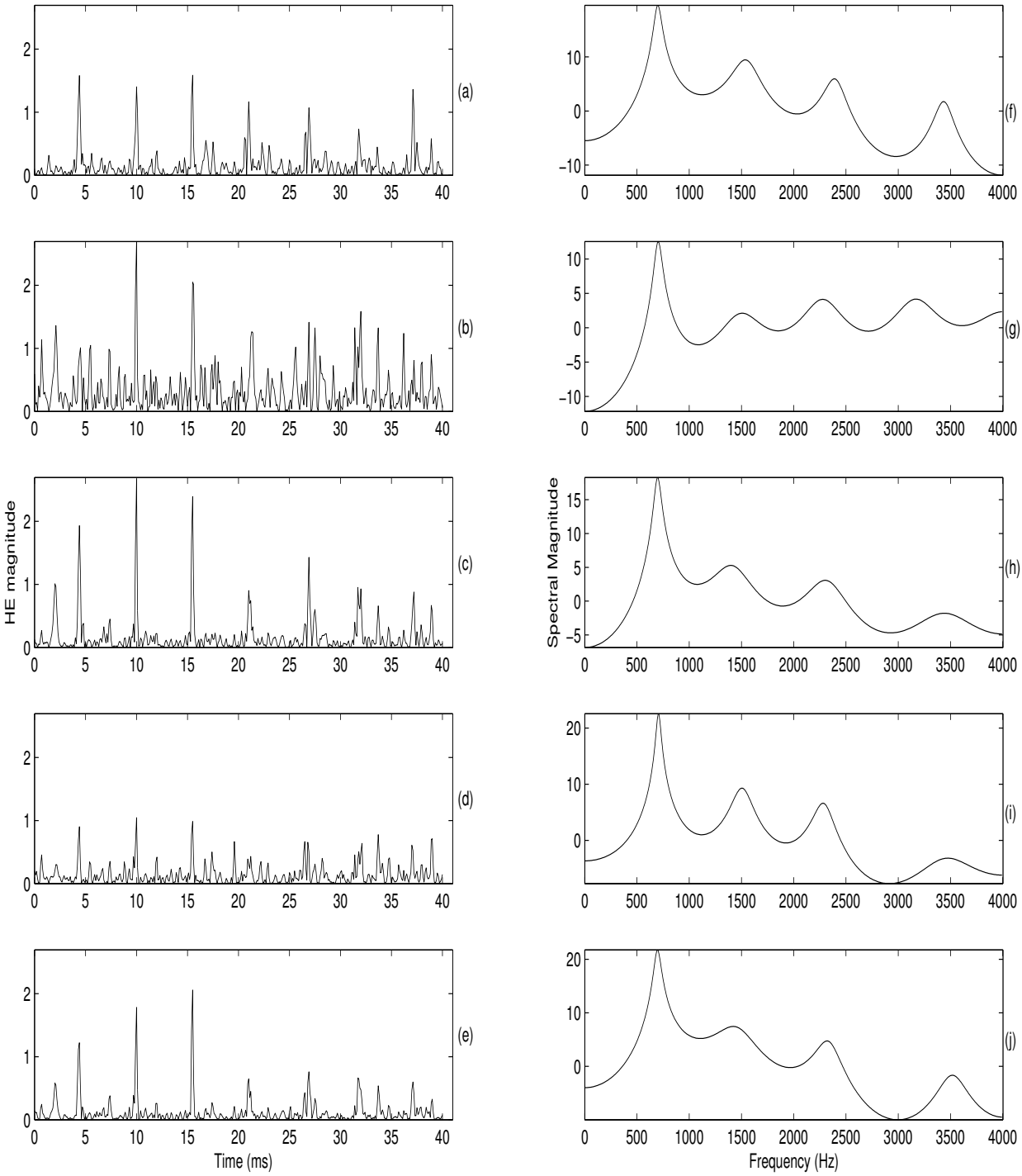


Figure 3.22: HE of the LP residual of (a) clean speech, (b) degraded speech, (c) speech processed by temporal processing, (d) speech processed by spectral processing, (e) speech processed by temporal and spectral processing, and Vocal-tract (LP) spectrum of a frame of (f) clean speech, (g) degraded speech, (h) speech processed by temporal processing, (i) speech processed by spectral processing, and (j) speech processed by temporal and spectral processing.

3.5.2 Gross Weight Function Performance

The performance of the gross weight function detection algorithm is evaluated using the manually labelled high SNR and low SNR regions. The manual labeling is done independently by three different persons and the final decision about high SNR and low SNR regions is taken based on the majority logic. For comparison, the gross weight function is computed for each of the indicators independently and also for combined one. Table 3.2 shows the percentage of correct detection accuracy (P_c) and false alarm probabilities (P_f) for various SNR levels and different noises taken from NOISEX-92 database. The value of P_c is computed as

$$P_c = \frac{N_c}{N_t} \times 100 \tag{3.33}$$

where N_c is the total number of correctly detected low and high SNR frames and N_t is the total number of frames. The value of P_f is computed as

$$P_f = \frac{N_f}{N_{tl}} \times 100 \tag{3.34}$$

where N_f and N_{tl} are the total number of incorrectly classified low SNR frames and the total number of low SNR frames, respectively. The last column entries of the table indicate the performance of the gross weight function algorithm for the combined one. The proposed method is also evaluated for the speech data recorded in the real noisy environment. The last row entries in the table show the performance of the gross weight function algorithm for the real noisy speech data. As mentioned earlier, the following observations can be made from the table (i) smoothed HE performs well under Babble noise, (ii) sum of the DFT spectral peaks performs well under white and pink noise and (iii) modulation spectrum gives consistent performance under all noise environments. But the average performance is less compared to that of DFT spectrum or smoothed HE alone. The combined method consistently yields better performance than the individual parameters.

The performance of the proposed gross weight function method is compared with two basic voice methods: (1) short time energy and zero crossing rate profile [100], and (2) inverse spectral flatness profile [17].

Short-time energy and zero crossing rate profile [100]: The short-term signal energy and zero-crossing rate have long been used as simple acoustic features for voice activity detection [100]. In this algorithm it is assumed that during the first 100 ms of the recording interval there is no speech present. During this interval the statistics of the background noise are measured. These measurements

include the average and standard deviation of the average energy and the zero crossing rates. From these measurements three different thresholds are computed. They are: (i) ITU - Upper energy threshold, (ii) ITL - Lower energy threshold and (iii) IZCT - Zero crossings rate threshold. Then the method proceeds as follows. Search from the beginning until the energy crosses ITU. Then search backward towards the signal beginning until the first point at which the energy falls below ITL is reached. This is assumed to be a provisional beginning point (N_1). The end point (N_2) is selected in a similar way. For the beginning point, now the the zero-crossing rate of the previous 250 ms is examined. If this measure exceeds the IZCT threshold 3 or more times, beginning point (N_1) is moved to the first point at which the IZCT threshold is exceeded. Again, perform a similar method for the end point N_2 [100]. The gross weight function result based on the short-time energy and zero crossing rate is shown in Fig. 3.23 for the same noisy speech shown in Fig. 3.7.

Inverse spectral flatness profile [17]: Inverse spectral flatness is the ratio of speech energy to the linear prediction (LP) residual energy. For each small segment of the residual signal, this ratio gives an indication of the amount of reduction in the correlation of the signal samples. This also gives an indication of how much the signal spectrum is flattened in the residual. If the signal spectrum is already flat, then the energies of the noisy signal and the residual signal in the short segment will be nearly unity. Otherwise, the ratio will be quite large. Thus the computation of inverse spectral flatness gives an indication of the high SNR regions. For evaluation the inverse spectral flatness is computed with a non overlapping frame of 2 ms and smoothed using a 17-point Hamming window. The smoothed inverse spectral flatness values are non-linearly mapped to enhance the contrast between the high SNR and low SNR regions [17]. Fig. 3.24 depicts the various processing steps involved in this method.

The percentage of correct detection accuracy (P_c) of these two basic methods are given in Table 3.3. For comparison the P_c values obtained with the proposed method also shown. The comparison indicates that our proposed method performs better than the conventional methods under all noisy conditions. We also observed that the performance of the conventional methods depends on the SNR of noisy speech. However the proposed method is independent of the noise level. It can be seen from the results that the proposed method shows a maximum of only ± 2 % deviation across all the SNR levels. But if we consider the case of basic methods there is large deviation (maximum of ± 10 % for short time energy and zero crossing rate and ± 8 % for inverse spectral flatness) in the performance.

Table 3.2: Gross weight function performance. In the table abbreviations SDFT, SHE, MS and COMB refer to sum of peaks in the DFT spectrum, smoothed HE of the LP residual, modulation spectrum and combination of all three indicators, respectively.

Noise Type	SNR Level	SDFT		SHE		MS		COMB	
		P_c	P_f	P_c	P_f	P_c	P_f	P_c	P_f
White Gaussian Noise	0 dB	88.09	6.60	77.16	19.39	82.45	9.37	88.58	6.33
	3 dB	87.12	6.33	80.64	16.09	82.59	9.23	88.44	6.73
	6 dB	87.88	3.83	81.55	13.72	81.48	9.76	87.81	3.69
	9 dB	88.65	3.56	83.01	12.66	81.96	9.37	89.07	3.83
Babble Noise	0 dB	74.44	22.16	82.45	11.48	82.87	8.05	85.86	8.05
	3 dB	81.55	11.35	82.24	11.21	82.59	8.58	86.35	8.31
	6 dB	85.79	6.60	83.36	9.76	82.38	8.71	86.91	8.05
	9 dB	86.49	6.92	83.22	9.76	82.17	8.97	87.26	6.46
Factory Noise	0 dB	81.82	12.14	82.87	11.08	82.66	8.31	87.33	6.99
	3 dB	84.61	9.37	83.50	9.76	82.52	8.44	87.40	7.39
	6 dB	86.91	8.05	82.87	9.76	82.24	8.84	87.53	7.65
	9 dB	87.40	6.33	82.66	9.89	82.17	9.10	87.26	5.67
Pink Noise	0 dB	78.34	16.09	79.25	13.06	84.26	7.12	86.63	7.12
	3 dB	83.98	9.37	83.22	9.89	84.26	7.39	87.95	8.39
	6 dB	86.56	7.65	83.36	8.84	83.50	8.18	87.95	7.86
	9 dB	87.60	6.60	83.29	9.10	82.45	8.58	87.40	6.91
Real Data		84.47	7.79	84.96	11.53	83.77	9.85	87.49	7.39

Table 3.3: Comparison of gross weight function performance with the simpler methods. In the table abbreviations STEZCR and ISF refer to short-time energy and zero crossing rate profile and the inverse spectral flatness profile, respectively.

SNR Level	Correct detection accuracy (P_c)			Correct detection accuracy (P_c)		
	STEZCR	**ISF**	**Proposed**	**STEZCR**	**ISF**	**Proposed**
	White Noise			Babble Noise		
0 dB	71.08	82.87	88.58	79.55	76.89	85.86
3 dB	75.13	84.84	88.44	80.19	79.69	86.35
6 dB	77.05	86.28	87.81	80.05	82.46	86.91
9 dB	76.32	87.53	89.07	79.19	84.40	87.26
	Factory Noise			Pink Noise		
0 dB	81.49	82.45	87.30	70.78	82.28	87.95
3 dB	81.69	86.60	87.40	79.22	85.28	87.95
6 dB	81.37	89.23	87.53	81.86	87.73	87.40
9 dB	81.01	90.63	87.26	81.99	89.61	87.49

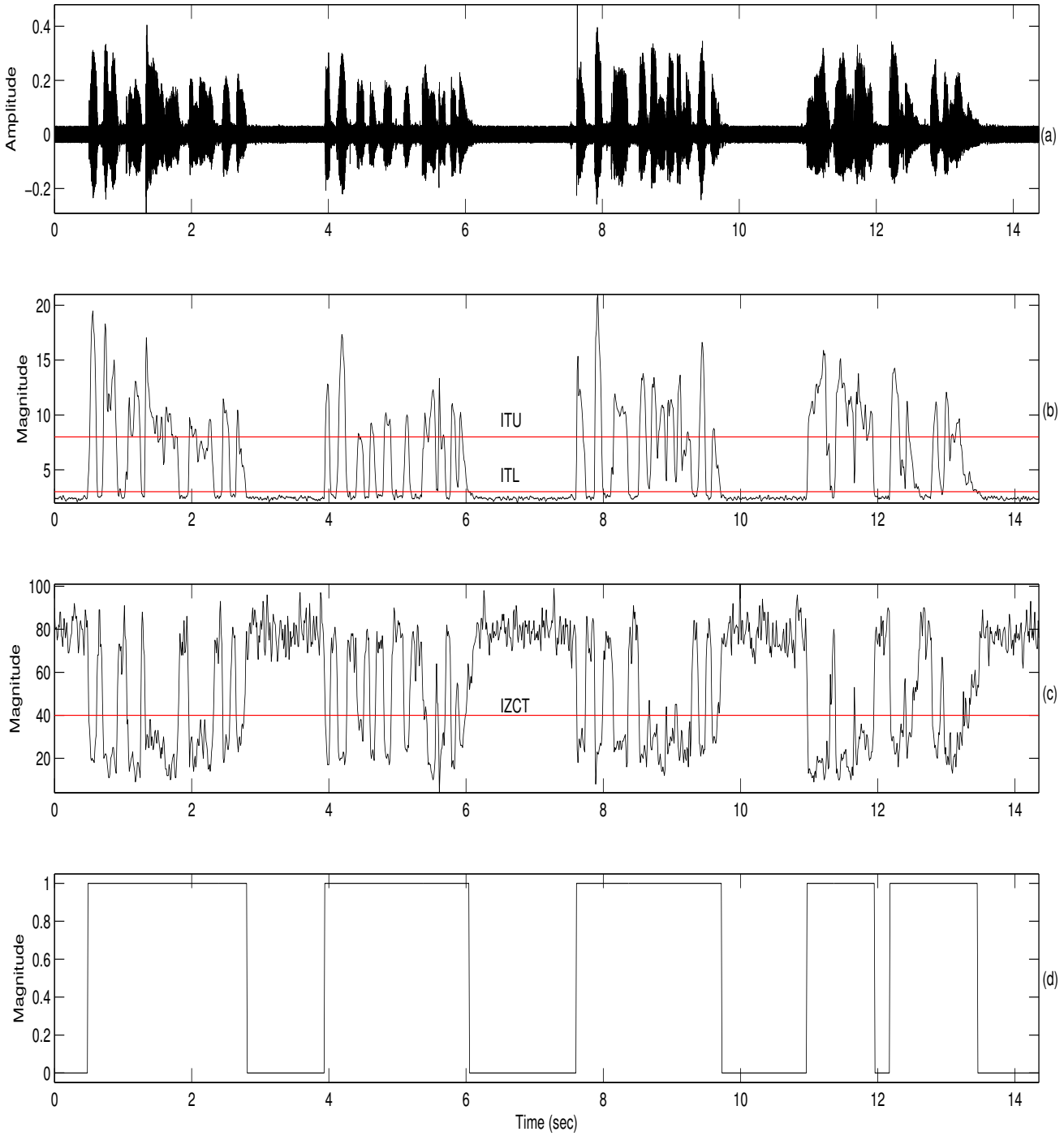


Figure 3.23: Gross level features identification using short-time energy and zero crossing rate profile: (a) noisy speech (SNR= 3 dB), (b) short-time energy, (c) short-time zero crossing rate, and (d) gross weight function.

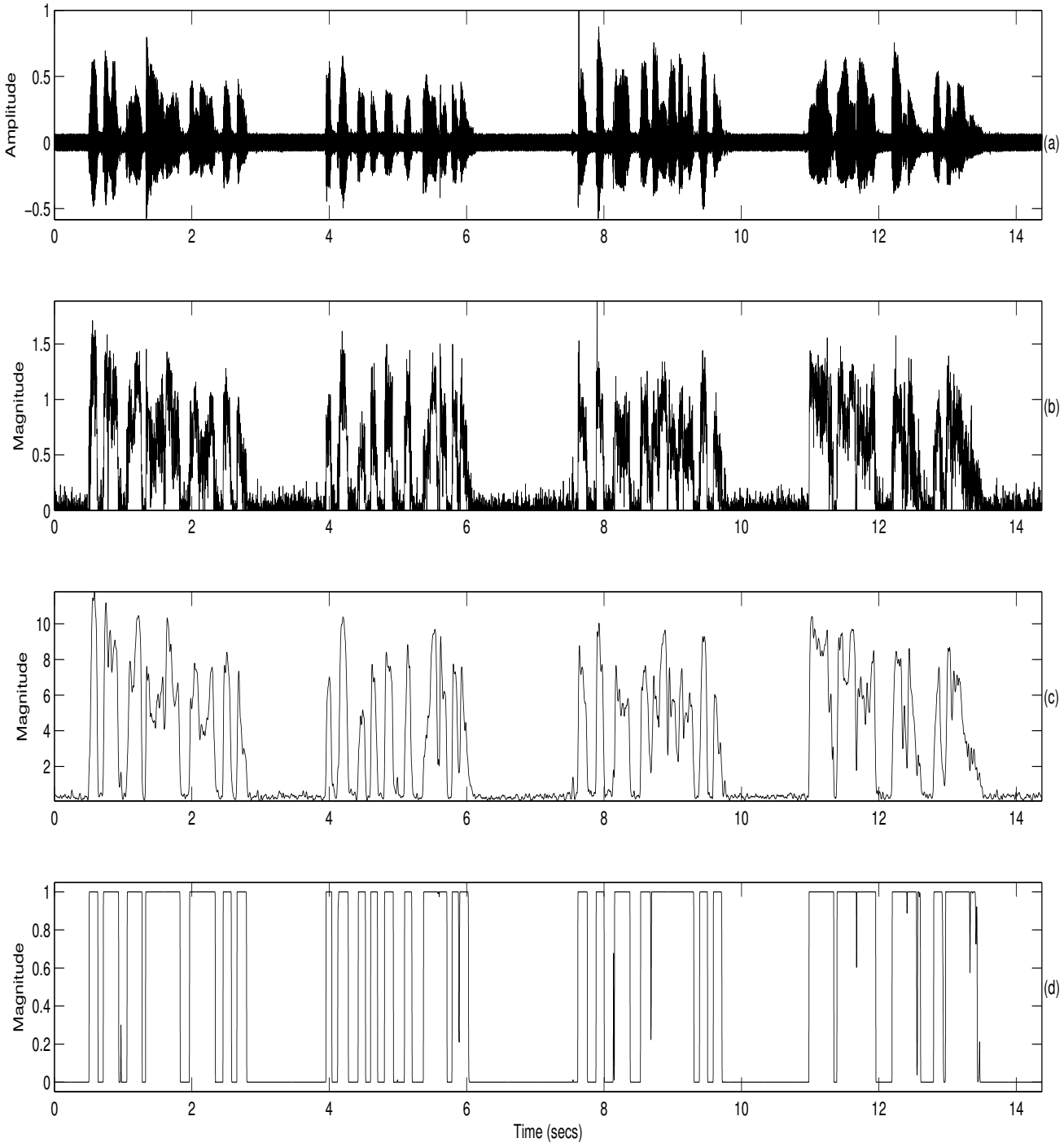


Figure 3.24: Gross level features identification using inverse spectral flatness profile: (a) noisy speech (SNR= 3 dB), (b) inverse spectral flatness, (c) smoothed inverse spectral flatness, and (d) nonlinearly mapped inverse spectral flatness.

3.5.3 Fine Weight Function Performance

The performance of the fine weight function detection algorithm is evaluated by computing the deviation in the approximate epoch locations of the proposed method with respect to their clean speech instants. First the the approximate instants location of the clean speech signal are computed using the FOGD as described earlier. The percentage of accuracy (P_a) in determining the instant of significant excitation is found as

$$P_a = \frac{N_{tc}}{N_{ti}} \times 100 \tag{3.35}$$

where N_{tc} represents the total number of instants detected at the same locations (or within the specified time resolution) of direct signal instants and N_{ti} is the total number of clean signal instants. The result of the percentage of accuracy analysis is given in Table 3.4. The entries in the Table 3.4 show the percentage of approximate instants and their deviation with respect to the direct signal instants location. From the table it can be observed that most of the detected instants lie within the 2 ms interval with reference to clean speech instants. As already mentioned from speech enhancement perspective, an approximate location of instants is sufficient. This is because the enhancement is normally performed by emphasizing the residual signal in the regions around the instants of significant excitation.

Table 3.4: Percentage of approximate instants derived for different deviations with respect to clean speech instant locations.

SNR Level	Deviation in time				Deviation in time			
	White Noise				Babble Noise			
	0.5 ms	**1 ms**	**1.5 ms**	**2 ms**	**0.5 ms**	**1 ms**	**1.5 ms**	**2 ms**
0 dB	49.9	67.1	78	84.44	74.37	83.07	88.42	91.98
3 dB	54.69	70.6	80.95	87.05	74.85	83.96	88.97	92.6
6 dB	58.67	73.13	82.25	88.07	77.11	84.99	89.65	93.01
9 dB	61.62	76.56	84.24	89.24	77.31	86.15	89.92	92.67
	Factory Noise				Pink Noise			
0 dB	72.86	81.97	87.59	91.5	54.35	70.73	79.85	87.18
3 dB	74.64	83.41	88.21	91.71	60.66	73.47	81.91	89.03
6 dB	77.38	85.47	89.58	92.73	64.15	76.9	84.85	91.02
9 dB	78.68	86.36	89.72	92.53	67.99	78.68	85.26	90.82

3.5.4 Pitch Estimation Performance

The performance of pitch estimation is evaluated in terms of deviation between (i) pitch frequency of clean speech and degraded speech and (ii) pitch frequency of clean speech and temporally processed speech. The pitch frequency of the respective signal is estimated from the autocorrelation of the HE of LP residual. Then the accuracy of pitch estimation is measured as

$$P_e = \frac{N_{tp}}{N_{cs}} \times 100 \tag{3.36}$$

where N_{cs} is the total number of frames in the clean speech and N_{tp} is the total number of frames having $F_{cs} > 0$ & $|F_{cs} - F_{tp}| \leq F_r$. The abbreviations F_{cs} and F_{tp} represent pitch frequency (in Hz) of clean and temporally processed speech, respectively. F_r is frequency deviation considered for the evaluation. The results of this evaluation are given in Table 3.5 for different values of F_r (F_r=5, 10, 15 & 20 Hz). For comparison the same experiment is repeated with reference degraded speech and the results are tabulated in Table 3.6. From the results it can be seen that the pitch estimate obtained from temporally processed speech consistently gives a superior performance than the pitch estimate obtained from the degraded speech for all SNR levels. For higher levels of degradation, the estimation error of degraded speech is very high (nearly 50 %). On the other hand the pitch estimate obtained from the temporally processed speech shows acceptable performance. This is mainly due to the enhancement of the instants of significant excitation of voiced speech.

Table 3.5: Percentage of accuracy of the pitch estimation of temporally processed speech with reference to clean speech.

SNR	**Deviation in frequency (Hz)**				**Deviation in frequency (Hz)**			
Level	**± 5**	**± 10**	**± 15**	**± 20**	**± 5**	**± 10**	**± 15**	**± 20**
	White Noise				**Babble Noise**			
0 dB	75.58	78.47	78.67	78.84	89.69	90.54	90.74	90.84
3 dB	80.16	82.57	82.81	82.98	91.12	91.90	92.13	92.23
6 dB	83.93	85.72	85.93	86.10	92.23	92.85	93.05	93.08
9 dB	86.88	88.57	88.74	88.84	93.22	93.73	93.86	93.90
	Factory Noise				**Pink Noise**			
0 dB	85.62	87.15	87.42	87.49	81.45	83.49	83.55	83.69
3 dB	88.40	89.79	89.89	90.00	85.15	86.81	86.94	87.05
6 dB	90.30	91.42	91.45	91.52	87.96	89.32	89.52	89.66
9 dB	91.79	92.54	92.68	92.71	89.66	91.05	91.22	91.32

Table 3.6: Percentage of accuracy of the pitch estimation of degraded speech speech with reference to clean speech

SNR Level	Deviation in frequency (Hz)				Deviation in frequency (Hz)			
	± 5	± 10	± 15	± 20	± 5	± 10	± 15	± 20
	White Noise				Babble Noise			
0 dB	57.95	58.32	58.32	58.32	81.62	82.03	82.1	82.13
3 dB	62.36	62.94	62.94	62.94	85.69	86.06	86.13	86.2
6 dB	68.84	69.35	69.35	69.35	88.34	88.81	88.88	88.91
9 dB	73.89	74.53	74.57	74.57	90.13	90.57	90.67	90.71
	Factory Noise				Pink Noise			
0 dB	71.85	72.43	72.47	72.47	65.31	65.75	65.75	65.75
3 dB	77.18	77.86	77.89	77.89	70.97	71.62	71.62	71.65
6 dB	81.52	82.16	82.16	82.16	76.36	77.18	77.21	77.21
9 dB	85.89	86.4	86.44	86.47	80.91	81.62	81.69	81.72

3.5.5 Composite Objective Quality Measures

The proposed method is evaluated using the composite objective quality measures that have high degree of correlation with subjective quality [294–296]. This measure evaluates the quality of enhanced speech along three dimensions: signal distortion, noise distortion, and overall quality. The resultant objective score values are in between 1 to 5 like mean opinion score (MOS). The MOS score values obtained from this measure is based on ITU-T P.835 standard [297]. This standard rates the quality of the enhanced speech on three different measures. They are

(i) The speech signal alone using a five-point scale of signal distortion ((C_{sig})) [1-Very unnatural, 2-Farly unnatural, 3-Somewhat natural, 4-Fairly natural and 5-Very natural]

(ii) The background noise alone using a five-point scale of background intrusiveness ((C_{bak})) [1-very intrusive, 2-somewhat intrusive, 3-Noticeable but not intrusive, 4-Somewhat noticeable and 5-Not noticeable]

(iii) The overall effect using the scale of the Mean Opinion Score ((C_{ovl})) [1=bad, 2=poor, 3=fair, 4=good and 5=excellent].

A detailed description of the composite objective quality measure is given in Appendix-D.

Table 3.8, 3.9 and 3.10 show mean opinion score (MOS) for signal distortion level, background noise level and overall objective quality values obtained from the different algorithms. The abbreviations used in the Table 3.8 - 3.10 are listed in Table 3.7. Table 3.11 shows the percentage of improvement in each of the MOS score with reference to the degraded speech signal. The percentage improvement in MOS is calculated as

$$\%\text{Increment} = \frac{\text{Degraded Speech Score} - \text{Enhanced Speech Score}}{\text{Degraded Speech Score}} \times 100. \tag{3.37}$$

These values are calculated for all SNR levels in all four noise cases and finally averaged across the noise types. The following observations can be made from the contents of Tables 3.8 - 3.11.

(i) The combined TSP method shows improved performance, compared to the individual processing methods.

(ii) The temporal processing alone gives always less performance compared to the conventional spectral processing methods. It is expected mainly because there is no specific attempt is made to explicitly remove the background noise. The enhancement is achieved only by processing of speech-specific regions, i.e., instants of significant excitation. From perception point of view also the speech enhanced by the temporal processing method is also noisier than the ones enhanced by the spectral based methods.

(iii) In spectral processing method the order of best performing system in terms of overall objective quality scores are: (i) MMSE-LSA estimator, (ii) MMSE-STSA estimator, (iii) multi-band spectral subtraction, and (iv) conventional spectral subtraction. However with reference to signal distortion score MMSE-STSA estimator performs well than that of LSA estimator.

(iv) In spectral subtractive based algorithms multi-band spectral subtraction performed consistently better across all conditions.

(v) The combined TSP method (without spectral enhancement technique) itself gives higher MOS score for signal distortion, background noise level and the overall objective quality compared to the temporal or spectral processing alone. The results show that the additional spectral enhancement technique gives relatively higher improvement in background noise level score as compared to the signal distortion score. This reduction in the background noise level is achieved through the enhancement of speech specific features (region around pitch and its harmonics)

in the spectral domain. As mentioned earlier, the enhancement of the speech specific spectral amplitudes relatively reduces the noise spectral amplitudes in the high SNR regions. This result in higher MOS for background noise level compared to the combined temporal and spectral processing. However the same amount of relative increment is not evident in the overall quality score. This is theoretically interpreted as follows: Hu and Loizou treated the overall quality score as the dependent variable and the speech and noise scores as independent variables. By regression analysis they found the relationship between the three scores as [291, 292]:

$$C_{ovl} = -0.0783 + 0.571C_{sig} + 0.366C_{bak}. \tag{3.38}$$

This shows that the overall quality score has higher correlation with the signal distortion score as compared to the background noise level score. Due to this lower correlation the same amount of relative increment is not seen in the overall quality score. For lower correlation, the reason being stated was listeners seem to place more emphasis on the distortion imparted on the speech signal itself rather than on the background noise, when making judgments of overall quality [291].

(vi) In combined TSP method, the relative amount of increase in the performance reduces as the SNR of the noisy speech is increased from 0 dB to 9 dB. In addition, under white noise environment combined TSP methods result lower performance than that of other noise environments. This poor performance can be attributed to the limitations of LP analysis under high degradation due to white noise.

(vii) In general as the additive noise level in the speech signal is increased, the quality of the resulting enhanced speech decreases progressively due to loss of speech information in the low SNR, high noise regions.

Further to demonstrate the importance of the combined temporal weight function, we have compared the performance of speech signal with only silences suppressed and the proposed temporal processing method. For evaluation another set of ten different speech signals (five male & five female) from the TIMIT database is taken and the noise from Noisex-92 database is added to create the noisy speech data at SNR levels ranging from 0 to 25 dB. The noisy speech is processed by the proposed temporal processing and gross weight function alone. The overall MOS score values (C_{ovl}) obtained for the speech signal with only silences suppressed (i.e., weighted by gross weight function alone) and

speech processed by the combined weight function are shown in Fig. 3.25. It can be observed that the use of combined weight function improves the performance compared to that of speech signal with only silences suppressed. However, for higher SNR values, in particular for SNR $\geq$ 20 dB, the weighted signal results slightly poor performance than original speech alone. This is mainly because the underlying temporal processing method involves weighting of the excitation source signal for the enhancement. For higher SNR values, since noise level is very low, the weighting may disturb the actual signal itself and thus results in the slight reduction in performance.

The next experiment conducted is to analyze the performance of the proposed spectral enhancement method with the Ephraim and Malah noise suppression method [14]. The same speech signals considered in the earlier case are used and the obtained overall objective score values are shown in Fig. 3.26 for SNR levels ranging from 0 to 25 dB . The results show that the integration of proposed spectral enhancement operation further improves the performance of the conventional Ephraim and Malah noise suppression method. Similar to the previous case, for higher SNR level the spectral enhancement slightly reduces the performance. It is expected mainly because for very high SNR levels the proposed technique is equivalent to disturbing the pitch and its harmonics location by adding the spectral values. Thus the resultant objective quality score becomes lower.

Table 3.7: Abbreviations of the various symbols used in Table 3.8 - 3.11

Symbol	Abbreviations
DEG	Degraded Speech
TP	Temporal Processing
SP1	Spectral Subtraction
SP2	Multi-band Spectral Subtraction
SP3	MMSE-STSA Estimator
SP4	MMSE-LSA Estimator
TSP1	Combined TP and SP1
TSP2	Combined TP and SP2
TSP3	Combined TP and SP3
TSP4	Combined TP and SP4
TSP1E	TSP1 and Spectral Enhancement
TSP2E	TSP2 and Spectral Enhancement
TSP3E	TSP3 and Spectral Enhancement
TSP4E	TSP4 and Spectral Enhancement

Table 3.8: Signal distortion score for different speech signals of the examples collected from the TIMIT and NOIZEUS database. In the table, abbreviations DEG, TP, SP1, SP2, SP3 and SP4 refer to degraded Speech, temporal Processing, spectral subtraction, multi-band spectral subtraction, MMSE-STSA estimator and MMSE-LSA estimator, respectively. TSPx refers to combined temporal and respective spectral processing. Similarly, TSPxE refers to combined temporal and respective spectral processing with spectral enhancement.

Noise Type	SNR Level	DEG	TP	SP1	SP2	SP3	SP4	TSP1	TSP2	TSP3	TSP4	TSP1E	TSP2E	TSP3E	TSP4E
White Gaussian Noise	**0 dB**	2.09	2.43	2.45	2.53	3.02	2.90	2.85	3.00	3.36	3.44	2.98	3.13	3.45	3.47
	3 dB	2.46	2.78	2.84	2.91	3.37	3.28	3.27	3.42	3.63	3.70	3.38	3.51	3.70	3.72
	6 dB	2.83	3.12	3.23	3.27	3.78	3.73	3.60	3.74	3.88	3.95	3.70	3.82	3.95	3.96
	9 dB	3.19	3.43	3.66	3.69	4.06	4.01	3.88	4.01	4.11	4.16	3.97	4.08	4.16	4.17
Babble Noise	**0 dB**	3.03	3.30	3.40	3.46	3.62	3.57	3.54	3.64	3.72	3.71	3.67	3.78	3.81	3.79
	3 dB	3.33	3.56	3.70	3.68	3.91	3.90	3.81	3.81	3.96	3.96	3.92	3.92	4.05	4.03
	6 dB	3.62	3.81	3.99	3.97	4.19	4.19	4.06	4.06	4.19	4.21	4.16	4.16	4.28	4.28
	9 dB	3.90	4.04	4.26	4.24	4.45	4.45	4.29	4.30	4.41	4.43	4.38	4.39	4.48	4.49
Factory Noise	**0 dB**	3.29	3.46	3.73	3.76	3.87	3.87	3.78	3.83	3.97	4.02	3.91	3.95	4.07	4.10
	3 dB	3.56	3.71	4.02	4.04	4.13	4.15	4.04	4.08	4.19	4.26	4.16	4.20	4.29	4.32
	6 dB	3.84	3.96	4.30	4.32	4.37	4.39	4.29	4.33	4.40	4.46	4.40	4.44	4.48	4.52
	9 dB	4.10	4.19	4.56	4.58	4.61	4.63	4.51	4.55	4.59	4.64	4.61	4.65	4.67	4.69
Pink Noise	**0 dB**	2.76	3.01	3.08	3.09	3.41	3.34	3.35	3.44	3.70	3.75	3.48	3.55	3.79	3.80
	3 dB	3.09	3.32	3.42	3.41	3.75	3.70	3.69	3.78	3.93	3.98	3.81	3.89	4.01	4.02
	6 dB	3.41	3.61	3.79	3.78	4.06	4.03	3.97	4.05	4.15	4.20	4.08	4.15	4.23	4.23
	9 dB	3.72	3.88	4.15	4.16	4.32	4.31	4.22	4.30	4.36	4.40	4.32	4.38	4.42	4.42

Table 3.9: Background noise level score for different speech signals of the examples collected from the TIMIT and NOIZEUS database. In the table, abbreviations DEG, TP, SP1, SP2, SP3 and SP4 refer to degraded speech, temporal processing, spectral subtraction, multi-band spectral subtraction, MMSE-STSA estimator and MMSE-LSA estimator, respectively. TSPx refers to combined temporal and respective spectral processing. Similarly, TSPxE refers to combined temporal and respective spectral processing with spectral enhancement.

Noise Type	SNR Level	DEG	TP	SP1	SP2	SP3	SP4	TSP1	TSP2	TSP3	TSP4	TSP1E	TSP2E	TSP3E	TSP4E
White Gaussian Noise	**0 dB**	1.22	1.47	1.50	1.54	1.90	1.96	1.76	1.84	2.02	2.13	1.87	1.96	2.11	2.20
	3 dB	1.39	1.63	1.70	1.74	2.05	2.11	1.95	2.03	2.16	2.26	2.04	2.13	2.25	2.33
	6 dB	1.57	1.80	1.89	1.92	2.21	2.26	2.11	2.19	2.29	2.39	2.21	2.29	2.38	2.45
	9 dB	1.75	1.96	2.09	2.13	2.36	2.41	2.26	2.34	2.41	2.51	2.36	2.44	2.50	2.57
Babble Noise	**0 dB**	1.52	1.73	1.79	1.88	1.96	1.97	1.95	2.00	2.07	2.11	2.06	2.16	2.17	2.19
	3 dB	1.69	1.89	1.97	1.96	2.13	2.15	2.10	2.11	2.22	2.26	2.22	2.22	2.32	2.34
	6 dB	1.86	2.04	2.14	2.14	2.29	2.32	2.26	2.27	2.36	2.40	2.37	2.38	2.46	2.50
	9 dB	2.03	2.19	2.31	2.31	2.45	2.48	2.40	2.42	2.49	2.54	2.52	2.53	2.60	2.63
Factory Noise	**0 dB**	1.69	1.84	1.98	2.02	2.10	2.15	2.08	2.12	2.21	2.30	2.19	2.23	2.32	2.39
	3 dB	1.85	1.99	2.16	2.19	2.25	2.32	2.24	2.27	2.35	2.44	2.35	2.39	2.46	2.53
	6 dB	2.00	2.14	2.33	2.37	2.39	2.46	2.39	2.43	2.48	2.57	2.51	2.55	2.58	2.66
	9 dB	2.17	2.28	2.50	2.53	2.54	2.61	2.53	2.57	2.60	2.68	2.65	2.70	2.71	2.78
Pink Noise	**0 dB**	1.46	1.66	1.74	1.76	1.95	2.00	1.91	1.97	2.11	2.21	2.02	2.08	2.21	2.29
	3 dB	1.64	1.82	1.91	1.93	2.12	2.17	2.08	2.14	2.24	2.33	2.19	2.25	2.34	2.41
	6 dB	1.81	1.97	2.11	2.12	2.27	2.32	2.24	2.30	2.37	2.45	2.35	2.41	2.46	2.53
	9 dB	1.97	2.13	2.30	2.32	2.42	2.47	2.39	2.45	2.49	2.57	2.51	2.56	2.58	2.63

Table 3.10: Overall objective quality score for different speech signals of the examples collected from the TIMIT and NOIZEUS database. In the table, abbreviations DEG, TP, SP1, SP2, SP3 and SP4 refer to degraded speech, temporal processing, spectral subtraction, multi-band spectral subtraction, MMSE-STSA estimator and MMSE-LSA estimator, respectively. TSPx refers to combined temporal and respective spectral processing. Similarly, TSPxE refers to combined temporal and respective spectral processing with spectral enhancement.

Noise Type	SNR Level	DEG	TP	SP1	SP2	SP3	SP4	TSP1	TSP2	TSP3	TSP4	TSP1E	TSP2E	TSP3E	TSP4E
White Gaussian Noise	**0 dB**	1.82	2.12	2.12	2.17	2.59	2.54	2.47	2.60	2.88	2.99	2.59	2.72	2.97	3.03
	3 dB	2.13	2.41	2.45	2.50	2.90	2.87	2.83	2.96	3.12	3.22	2.93	3.06	3.19	3.25
	6 dB	2.44	2.70	2.78	2.81	3.26	3.27	3.12	3.26	3.34	3.44	3.22	3.34	3.41	3.46
	9 dB	2.74	2.97	3.17	3.21	3.52	3.53	3.37	3.50	3.54	3.63	3.47	3.58	3.61	3.65
Babble Noise	**0 dB**	2.56	2.79	2.87	2.95	3.05	3.04	3.00	3.11	3.14	3.16	3.13	3.22	3.24	3.24
	3 dB	2.83	3.03	3.15	3.14	3.32	3.34	3.24	3.24	3.36	3.39	3.36	3.36	3.46	3.47
	6 dB	3.10	3.26	3.42	3.42	3.58	3.61	3.48	3.48	3.58	3.63	3.59	3.59	3.68	3.70
	9 dB	3.36	3.48	3.68	3.68	3.83	3.86	3.70	3.71	3.89	3.84	3.80	3.81	3.88	3.91
Factory Noise	**0 dB**	2.78	2.93	3.18	3.23	3.32	3.36	3.22	3.27	3.38	3.47	3.35	3.40	3.49	3.55
	3 dB	3.03	3.16	3.46	3.49	3.56	3.61	3.47	3.51	3.59	3.69	3.60	3.64	3.70	3.76
	6 dB	3.28	3.39	3.73	3.77	3.78	3.84	3.70	3.75	3.79	3.89	3.83	3.88	3.89	3.94
	9 dB	3.53	3.62	3.98	4.03	4.01	4.07	3.92	4.04	3.98	4.06	4.04	4.08	4.07	4.11
Pink Noise	**0 dB**	2.31	2.54	2.61	2.62	2.90	2.88	2.85	2.93	3.13	3.22	2.97	3.04	3.23	3.28
	3 dB	2.60	2.82	2.91	2.91	3.20	3.20	3.15	3.24	3.34	3.43	3.27	3.35	3.43	3.48
	6 dB	2.89	3.07	3.25	3.25	3.48	3.50	3.41	3.49	3.55	3.63	3.52	3.60	3.63	3.67
	9 dB	3.16	3.31	3.59	3.60	3.72	3.76	3.65	3.73	3.74	3.82	3.76	3.83	3.82	3.84

Table 3.11: Percentage improvement in signal distortion, background noise level and overall objective quality score with reference to the degraded speech. In the table, abbreviations TP, SP1, SP2, SP3 and SP4 refer to temporal processing, spectral subtraction, multi-band spectral subtraction, MMSE-STSA estimator and MMSE-LSA estimator, respectively. TSPx refers to combined temporal and respective spectral processing. Similarly, TSPxE refers to combined temporal and respective spectral processing with spectral enhancement.

SNR Level	TP	SP1	SP2	SP3	SP4	TSP1	TSP2	TSP3	TSP4	TSP1E	TSP2E	TSP3E	TSP4E
Percentage of Improvement in Signal Distortion Score													
0 dB	9.85	13.60	15.37	26.29	23.81	22.37	26.18	34.57	36.27	27.16	30.80	37.96	38.35
3 dB	7.89	12.54	13.16	22.94	21.69	20.06	22.59	27.84	29.45	23.82	26.07	30.58	30.92
6 dB	6.12	11.87	12.14	20.54	20.01	16.88	18.96	22.28	23.80	19.97	21.81	24.63	24.98
9 dB	4.40	11.69	11.98	17.49	17.15	13.77	15.63	17.77	18.86	16.33	17.91	19.50	19.76
Percentage of Improvement in Background Noise Level Score													
0 dB	14.22	19.26	22.50	35.63	38.62	31.61	35.69	44.26	50.22	39.19	44.29	51.09	55.67
3 dB	11.91	18.02	19.30	31.10	34.18	28.11	31.02	37.59	42.57	34.67	37.75	43.70	47.45
6 dB	10.04	17.13	18.24	27.20	29.96	24.79	27.53	31.92	36.28	30.88	33.62	37.19	40.81
9 dB	8.27	16.30	17.47	23.86	26.38	21.32	23.93	26.65	30.63	27.13	29.61	31.70	34.51
Percentage of Improvement in Overall Quality Score													
0 dB	10.21	13.99	16.02	26.60	25.96	23.03	27.20	34.49	37.98	28.41	32.28	38.78	40.68
3 dB	8.24	13.11	13.86	23.51	23.75	20.76	23.48	28.04	31.17	25.22	27.84	31.52	33.28
6 dB	6.35	12.61	13.22	21.19	22.16	17.73	20.24	22.69	25.57	21.59	23.89	25.67	27.07
9 dB	4.82	12.89	13.69	18.44	19.50	14.92	17.16	19.02	20.67	18.29	20.21	20.85	21.88

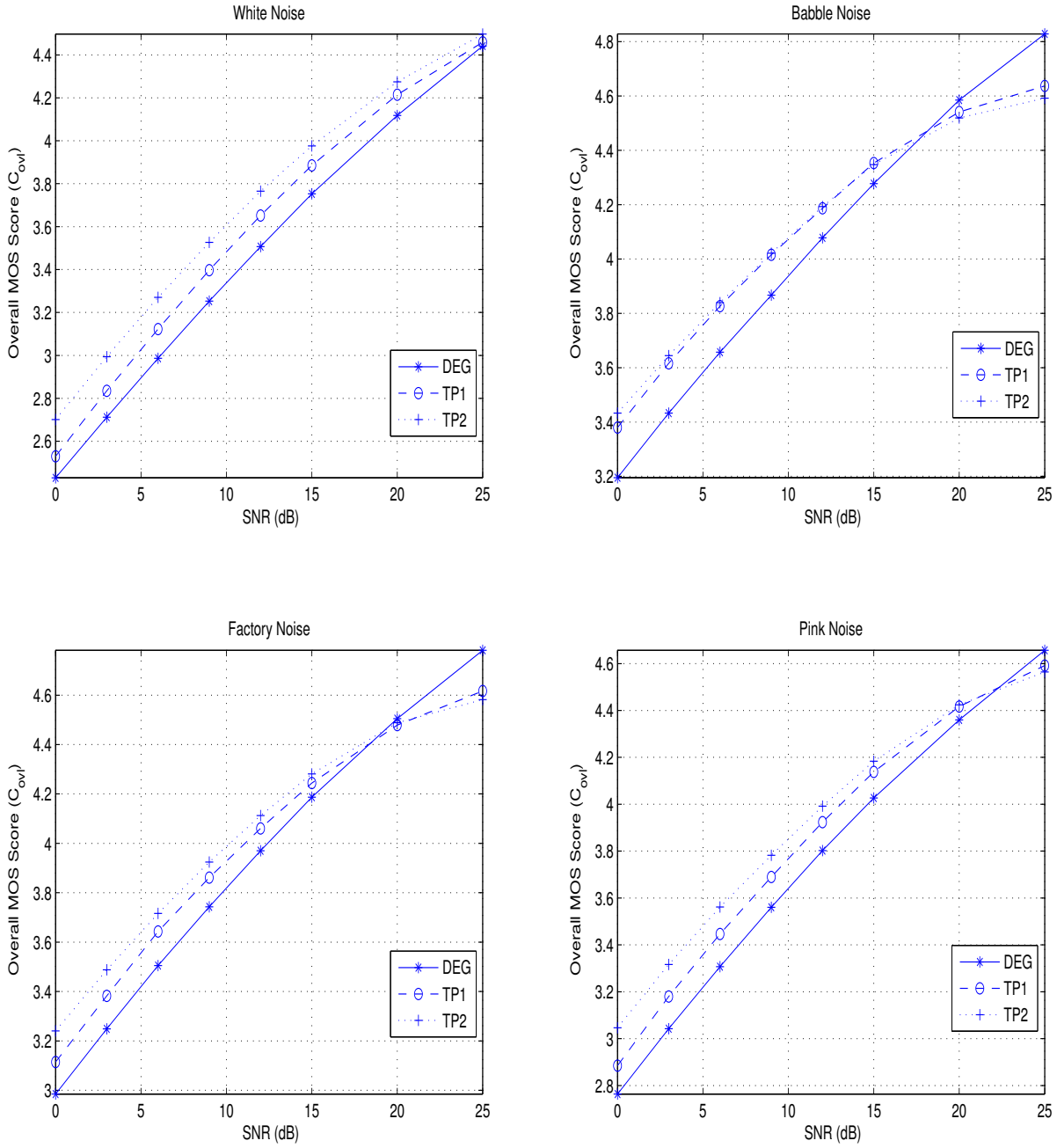


Figure 3.25: Overall MOS score (C_{ovl}) values for temporally processed speech signals. In figure, the abbreviations DEG, TP1 & TP2, respectively represent degraded speech, speech signal with only silence suppressed and speech processed by the combined weight function.

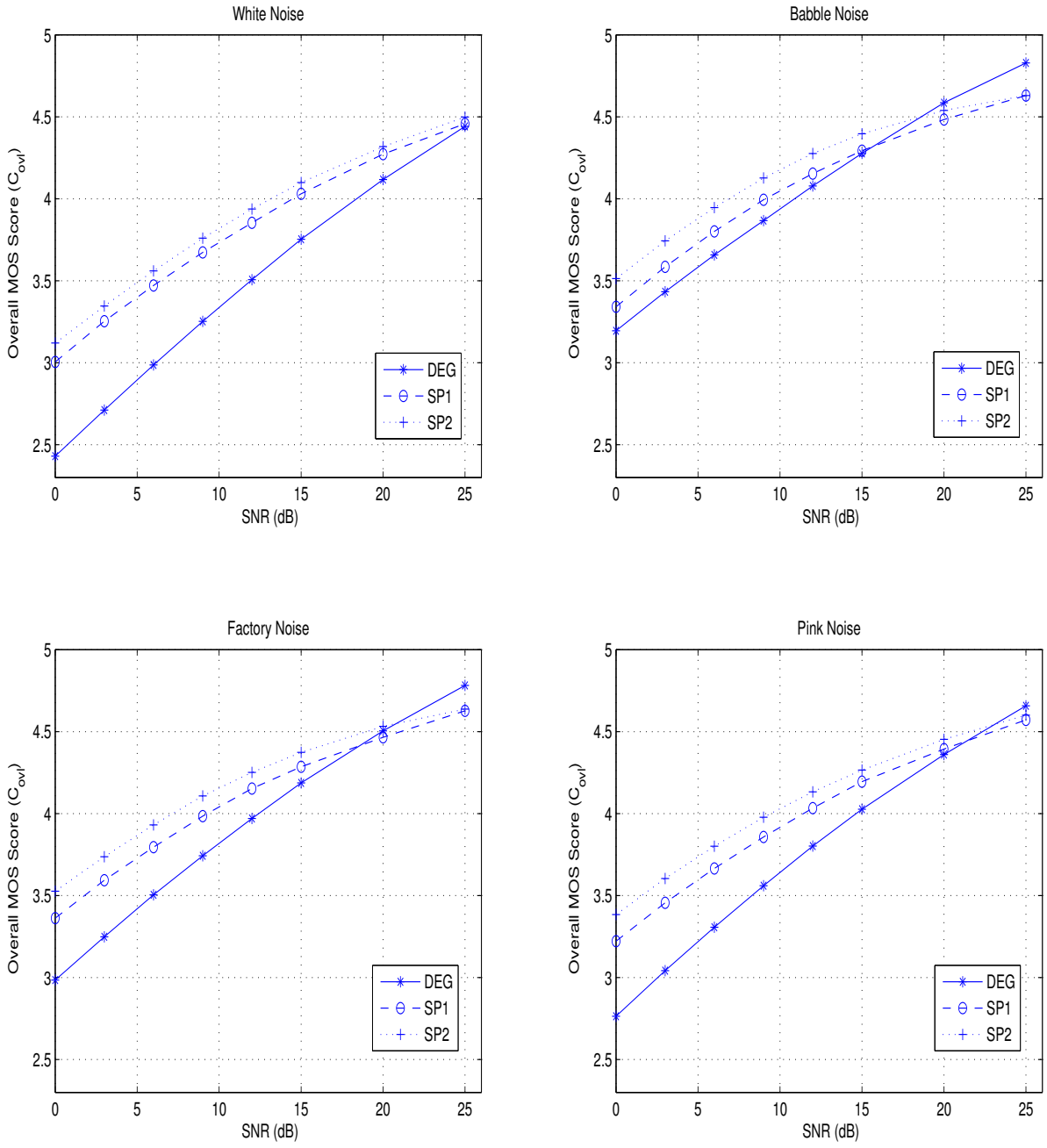


Figure 3.26: Overall MOS score (C_{ovl}) values for spectrally processed speech signals. In figure, the abbreviations DEG, SP1 & SP2 refer to degraded speech, speech processed by Ephraim and Malah noise suppression method, and speech processed by the Ephraim and Malah noise suppression combined with the proposed spectral enhancement method, respectively.

3.6 Summary

In this chapter a noisy speech enhancement method is proposed using combined TSP. The main objective of this study is to show that the combined TSP method gives relatively better performance compared to temporal or spectral processing alone in case of noisy speech enhancement. It is shown that the sum of the 10 largest peaks in the DFT spectrum, the smoothed HE of the LP residual and the modulation spectrum values of the noisy speech signal can be used as indicators for identifying high SNR and low SNR regions of noisy speech. The enhancement of noisy speech is achieved in two stages, namely, temporal enhancement followed by spectral enhancement. In temporal enhancement process, the HE of the LP residual is used to derive the information about strength of excitation. A fine weight function is derived using a FOGD to enhance the excitation source information around the glottal closure instants of speech. Because of high SNR nature of the regions around the GC events, the periodicity information is preserved even under high levels of degradation. A weight function for the LP residual is derived from the gross and fine weight functions. The enhanced speech is derived by exciting the time-varying all-pole filter with the LP residual modified by the weight function. In spectral enhancement, first enhancement is achieved by conventional spectral processing technique and an additional spectral enhancement is performed to further improve the perceptual quality of the speech. A performance evaluation is conducted using composite objective quality measures. The performance measures showed that the processed speech signals from the proposed method results better MOS compared to temporal or spectral processing alone. The next chapter presents a combined TSP method for the enhancement of reverberant speech.

4

Combined TSP for Reverberant Speech Enhancement

Contents

4.1 Objective of Combined TSP for Reverberant Speech Enhancement

This chapter presents an approach for the enhancement of reverberant speech by combined temporal and spectral processing. Temporal processing involves identification and enhancement of high signal to reverberation ratio (SRR) regions in the temporal domain. Spectral processing involves removal of late reverberant components in the spectral domain. First the spectral subtraction-based processing is performed to eliminate the late reverberant components. The spectrally processed speech is further subjected to the excitation source information based temporal processing to enhance the high SRR regions. The objective measures segmental SRR and log spectral distance are computed for different cases, namely, reverberant, spectral processed, temporal processed and combined temporal and spectral processed speech signals. The quality of the speech signal that is processed by the combined temporal and spectral processing is enhanced better compared to the reverberant speech as well as the signals that are processed by the individual temporal and spectral processing methods.

4.2 Introduction to Reverberant Speech Enhancement

Reverberation is one of the primary factors that degrades the quality of speech when collected by a distant microphone. Reverberation is caused by the fact that the microphone not only picks up the direct transmission of the signal, but also its reflections. The received signal generally consists of a direct sound, reflections that arrive shortly after the direct sound (early reverberation), and reflections that arrive after the early reverberation (late reverberation). The combination of the direct sound and early reverberation is sometimes referred to as the early sound component [19]. Early reverberation tends to perceptually reinforce the direct sound and are therefore considered harmless to speech intelligibility. This is often referred to as the precedence effect [147]. However, the frequency response of early reverberation is rarely flat: the speech spectrum is distorted and a change in speech quality can be perceived. This procedure is commonly known as coloration [144]. Reflections which arrive at larger delays with reference to the arrival of the direct sound are perceived either as separate echoes, or as reverberation; as such, late reflections impair speech intelligibility. The possible causes of reduction in speech intelligibly due to late reverberation are: (i) The energy overlap of a preceding phoneme on the following phoneme (overlap-masking) and (ii) the internal temporal smearing of energy within each phoneme (self-masking) [298]. As a result, speech becomes difficult to be understood in

the presence of room reverberation, especially for hearing-impaired and prevents computers from adequately extracting speech features. This gives rise to a need for reverberant speech enhancement algorithms.

Several studies of dereverberation have appeared in the literature. Most of these studies may be broadly classified into two categories: Temporal processing methods such as cepstral filtering, inverse filtering, temporal envelope filtering and LP residual processing, and spectral processing methods like spectral subtraction. Cepstrum-based techniques were reported first [161, 164, 175]. The underlying motivation is the fact that deconvolution in the time domain corresponds to subtraction in the cepstral domain. Since the cepstrum of speech is usually concentrated around the cepstral origin, while that of its echoes is composed of pulses away from the origin, dereverberation can be achieved by low time liftering in the cepstral domain. However, cepstral liftering relies on empirical evidence and uses a number of heuristics to determine the proper cutoff time for low time liftering [161,174]. Therefore, it has little use in practical speech dereverberation systems. Inverse filtering is frequently used to achieve speech dereverberation. If the acoustic impulse response is known (from calculations or measurements), reverberation can be removed by using the inverse filter or by minimum mean square error (MMSE) deconvolution [163]. However, typical acoustic impulse responses are non-minimum-phase and do not have stable causal inverses [142] and therefore inverse filtering based single-microphone techniques have limited scope in practice [163]. A method for recovering the temporal envelope of a signal from an observed reverberant signal by exploiting the characteristics of the modulation transfer function (MTF) [178] of source signal and a room impulse response is exploited in [185]. To realize this approach, it is assumed that the carrier signal of the speech and the impulse response are white noise. However, these assumptions are not accurate with regard to the real speech and reverberation. Therefore, this approach has not yet achieved high quality dereverberation [170].

Yegnanarayana *et al.* [22] proposed a reverberant speech enhancement system by manipulating LP residual signals based on the residual characteristics of clean and reverberation condition. The proposed method categorizes the peaks into speech events and reverberation based on the SRR values and attenuates the peaks in the LP residual that are derived from the reverberant speech. Later on in [23], the authors use Hilbert envelope (HE) of the LP residual to represent the strength of the peaks. The time-aligned HEs from the individual channels are summed and used as a weight vector which is applied to the LP residual of one of the channels to obtain the enhanced residual signal. In [186]

and [187], a spatiotemporal averaging of LP residual is proposed for enhancement of reverberant speech. The basis is that the waveform of the LP residual between adjacent larynx-cycles varies slowly, so that each such cycle can be replaced by an average of itself and its nearest neighboring cycle. The averaging results in the suppression of spurious peaks in the LP residual caused by room reverberation. The Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) algorithm [188] is employed for automatic estimation of glottal closure (GC) instants in voiced speech. However no attempt is made to eliminate spurious instants detected in the unvoiced and silence regions by DYPSA algorithm.

Recently, practically feasible single-microphone spectral subtraction based techniques have been developed that are able to suppress late reverberation [19, 21, 154, 156, 157, 159, 299]. The spectral subtraction methods estimate the power spectrum of the late reverberation and then subtract it from the power spectrum of reverberant speech. One of the main advantages of reverberation suppression method based on spectral subtraction is that these methods are not sensitive to the fluctuation of impulse responses. In [194] and [195], the authors proposed a reverberant speech enhancement algorithm using spatiotemporal and spectral processing. In the first stage, the spectral processing technique proposed in [19] is used to suppress late reverberation. In the second stage, the early reflections are suppressed by the LP residual processing in a similar way as in [186] and [187].

In speech processing tasks, removal of late reverberant components alone increases the quality of speech. In addition, the enhancement of high SRR regions by temporal processing may further increase the performance of the system, compared to the spectral processing alone. In general, reverberant speech enhancement methods should eliminate the late reverberant components and enhance the high SRR regions to improve speech intelligibility and feature extraction. From the literature it is found that, excitation source information-based temporal processing methods mainly enhance the speech-specific features of high SRR regions in the temporal domain. The main merit of temporal processing method is its effectiveness in the enhancement of high SRR regions. They are not used for removing weak reverberations in the tail of the impulse response, which unfortunately are also harmful to speech intelligibility [144]. However, spectral subtraction-based spectral processing methods reduce the late reverberation (i.e., tail of the impulse response) by estimating and subtracting the late reverberant spectrum from the degraded speech spectrum. On the other hand, these methods may not provide better enhancement in high SRR regions, compared to temporal processing methods. Both methods can therefore be combined effectively to enhance reverberant speech at high SRR regions and eliminate

late reverberation. In addition, both methods use different characteristics of the speech signal (in terms of the speech production mechanism) for enhancement. Therefore, they may provide better performance than the individual methods. This is the motivation of the work presented in this chapter.

In the proposed combined temporal and spectral processing method, spectral processing is performed first and the spectrally processed speech signal is then subjected to temporal processing. The main motivation behind this spectro-temporal processing is the identification of high SRR regions, primarily when the reverberation time is high. Due to the convolutive nature of reverberant speech, low SRR and reverberation-only regions (late reverberant regions) also look like speech signals which makes it difficult to separate low and high SRR regions. Therefore, first spectral processing is performed to eliminate the late reverberant regions and then temporal processing is performed. The attenuation of the late reverberant regions by spectral processing helps in two ways: First, it increases speech intelligibility and second, it increases the detection accuracy of low and high SRR regions. The contributions of the work presented in this chapter are, (i) A new temporal processing method, relatively more robust compared to the one reported in [22] - as will be demonstrated later, this is indicated by the improved performance, and (ii) A combined temporal and spectral processing method, in a similar way as in [194] and [195] using the proposed temporal processing method.

The rest of the chapter is organized as follows: Section 4.3 describes the reverberant signal model. Section 4.4 briefly reviews spectral processing of the reverberant speech signal. The temporal-domain processing of reverberant speech is described in Section 4.5. Experimental results of the proposed method and the objective measures performed on the experimental results are given in Section 4.6. Finally, the summary of work presented in this chapter is provided in Section 4.7.

4.3 Reverberation Signal Model

To justify the use of spectral subtraction for dereverberation, in this section we describe the characteristics of late reverberations and their relationship to direct-path response and early reflections. First, the impulse response of the room $h(n)$ can be written as [21]

$$h(n) = \begin{cases} h_a(n), & \text{for } 0 \leq n \leq N_1 \\ h_l(n), & \text{for } N_1 + 1 \leq n \leq N_i \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

where N_1 is the threshold, which is chosen such that $h_a(n)$ consists of the direct signal and few early reflections and $h_l(n)$ consists of all late reflections. The convolution of these segments with the desired signal results in the early reverberation, and late reverberation, respectively. This threshold N_1 is related to the sampling rate by $N_1 = F_s T_d$, where F_s is the sampling frequency in Hz and T_d is usually in the range of 40-80 ms. In particular, it is set to 50 ms for speech and 80 ms for music [144]. Then, the autocorrelation of the reverberant signal $z(n)$ is expressed as [157, 300]

$$r_{zz}(\tau) = E\left[z(n)z(n+\tau)\right] \tag{4.2}$$

$$= E\left[\sum_{m=-\infty}^{n} s(m)h(n-m) \sum_{q=-\infty}^{n+\tau} s(q)h(n+\tau-q)\right] \tag{4.3}$$

where $E[.]$ represents the expected value of random variable.

Given the nature of the speech signal and of the room impulse response, one can consider $s(n)$ and $h(n)$ as independent statistical processes [157, 300]. That is,

$$r_{zz}(\tau) = \sum_{m=-\infty}^{n} \sum_{q=-\infty}^{n+\tau} E\left[s(m)s(q)\right] E\left[h(n-m)h(n+\tau-q)\right]. \tag{4.4}$$

Assuming the room model as

$$h(n) = b(n)e^{-\overline{\delta}n}u(n) \tag{4.5}$$

where $b(n)$ represents a zero-mean white Gaussian noise, $u(n)$, the unit step function, and $\overline{\delta}$ is a damping constant related to the reverberation time T_{60}, obtained by [21]

$$\overline{\delta} = \frac{3\ln(10)}{T_{60}}. \tag{4.6}$$

Here, the reverberation time T_{60} is defined as the time needed for the sound energy to fall by 60 dB

after the original sound source is turned off [143]. From the room modeling given in Eqn. (4.5), the second expected value in Eqn. (4.3) is obtained as

$$E\left[h(n-m)h(n+\tau-q)\right] = e^{-2\overline{\delta}n}\sigma^2 e^{\delta(m+q-\tau)}\delta(m-q+\tau) \tag{4.7}$$

where $\delta(n)$ represents the unit sample sequence.

Thus,

$$r_{zz}(\tau) = e^{-2\overline{\delta}n}\sum_{m=-\infty}^{n} E\left[s(m)s(m+\tau)\right]\sigma^2 e^{2\overline{\delta}m}. \tag{4.8}$$

The speech signal has strong correlation within each local time region due to articulatory constraints, and it loses the correlation as a result of articulatory movements [159]. That is in practice the signals can be considered as stationary over periods of time that are short compared to the reverberation time T_{60}. This is justified by the fact that the exponential decay is very slow, and that speech is quasi-stationary [154]. Let T_s be the time span over which the speech signal can be considered stationary, which is usually around 20-40 ms [154]. If We consider $T_s \leq T_d \ll T_{60}$, it may be possible to assume that the autocorrelation of reverberant speech $z(n)$, $r_{zz}(\tau) = E[z(n)z(n+\tau)]$, has the following property [159]

$$r_{zz}(\tau) \simeq 0 \quad \text{iff} \quad \tau \geq T_d \tag{4.9}$$

where with reference to speech signal the value T_d can vary approximately from 30 to 100 ms depending on the phoneme of interest [159]. By considering the threshold T_d, we can split the summation in Eqn. (4.8) as

$$r_{zz}(\tau) = e^{-2\overline{\delta}n}\sum_{m=-\infty}^{n-T_d} E\left[s(m)s(m+\tau)\right]\sigma^2 e^{2\overline{\delta}m} + e^{-2\overline{\delta}n}\sum_{m=n-T_d+1}^{n} E\left[s(m)s(m+\tau)\right]\sigma^2 e^{2\overline{\delta}m} \tag{4.10}$$

Then, from Eqn. (4.10), the autocorrelation between the samples n and $n+\tau$ can be written as

$$r_{zz}(n-T_d, n-T_d+\tau) = e^{-2\overline{\delta}(n-T_d)}\sum_{m=-\infty}^{n-T_d} E\left[s(m)s(m+\tau)\right]\sigma^2 e^{2\overline{\delta}m} \tag{4.11}$$

where

$$r_{zz}(n, n+\tau) = r_{za}(n, n+\tau) + r_{zl}(n, n+\tau) \tag{4.12}$$

$$r_{zl}(n, n+\tau) = e^{-2\overline{\delta}T_d} r_{zz}(n-T_d, n-T_d+\tau). \tag{4.13}$$

The signal $z_l(n)$ is related to late reverberation, as a result of the convolution of $h_l(n)$ with $s(n)$, and $z_a(n)$ is associated with the direct signal and early reflections, being the result from the convolution

of $h_a(n)$ with $s(n)$. If we assume the condition of Eqn. (4.9), we can assume the late reverberations to be uncorrelated with the direct-path response [159]. Now, from Eqn. (4.11), the short-time power spectral density (PSD) of the reverberant speech $S_{zz}(n,k)$ is expressed as

$$S_{zz}(n,k) = S_{ze}(n,k) + S_{zl}(n,k). \tag{4.14}$$

The estimate of $S_{zl}(n,k)$ is obtained by attenuating and delaying the PSD of the acquired signal. That is,

$$S_{zl}(n,k) = e^{-2\overline{\delta}T_d} S_{zz}(n - T_d, k). \tag{4.15}$$

Therefore by assuming that late and early reflections are uncorrelated, the late reverberant signal can be treated as an additive noise, and the direct signal can be recovered through spectral subtraction [157, 300].

4.4 Spectral Processing of Reverberant Speech

As mentioned in reverberant speech enhancement methods, algorithms based on spectral subtraction can be utilized for reducing the late reverberation effects [19–21, 156–158, 193]. These methods model the impulse response as an outcome of a non-stationary random process using an exponential decay function to estimate the power of the reverberation. It is well known that speech signal is short-term stationary but long-term non-stationary. Late reflections of reverberation have delays that exceed the period during which speech can be reasonably considered stationary and, as a result, they smear speech spectra [20]. Early reflections, on the other hand, have delays within this period. Because of the short-term stationarity of speech, early reflections and the direct path signal have similar magnitude spectra. Consequently, early reflections cause coloration distortion and increase the intensity of reverberant speech. The time delay that separates early from late reflections is therefore not a property of the room impulse response; rather, it is a property of the source signal and indicates the boundary between short-term stationarity and long-term non-stationarity [20]. The delay separating early and late reflections is commonly set to 50 ms for speech and 80 ms for music [144]. It has also been shown that the early and late impulse components are approximately uncorrelated in the time domain. The late reverberant signal can be treated as an additive noise, and thus can be eliminated through spectral subtraction [21].

In general the power spectrum of the late reverberation is obtained by attenuating and delaying

the power spectrum of the acquired signal [19, 21, 156–158]. In this work, the PSD of late-impulse components is assumed to be a smoothed and shifted version of the PSD of the reverberant speech [20], i.e.,

$$\hat{S}_{zl}(l,k) = \chi w(l - N_1) * |Z(l,k)|^2 \tag{4.16}$$

where $Z(l,k)$ is the short time Fourier transform of reverberant speech $z(n)$. The short time speech spectrum is obtained using a Hamming window of 16 ms duration and 8 ms overlap between the frames. The symbol * denotes convolution in the time domain and $w(l)$ is a smoothing function. Finally, the scaling factor χ specifies the relative strength of the late-impulse components and is set to 0.32. It is already demonstrated by experimental results in [20] that system performance is not very sensitive to the specific values of χ. Considering the shape of the impulse response as an exponential decaying function, an asymmetrical smoothing function is chosen - the Rayleigh distribution [20]

$$w(l) = \begin{cases} \frac{l+a}{a^2} e^{\frac{-(l-a)^2}{2a^2}}, & l > -a \\ 0, & \text{otherwise} \end{cases} \tag{4.17}$$

where a controls the span of the smoothing function and is set to 5. Once the power spectrum of the late impulse components is estimated, the enhanced speech spectrum is obtained by subtracting the power spectrum of the late-impulse components from that of the reverberant speech. The spectral subtraction process can be described as a filtering operation in the frequency domain by

$$\hat{S}(l,k) = Z(l,k)H(l,k) \tag{4.18}$$

where

$$H(l,k) = \sqrt{\frac{|Z(l,k)|^2 - \hat{S}_{zl}(l,k)}{|Z(l,k)|^2}} \tag{4.19}$$

$$= 1 - \frac{1}{\sqrt{\gamma(l,k)}} \tag{4.20}$$

$$\gamma(l,k) = \frac{|Z(l,k)|^2}{\hat{S}_{zl}(l,k)} \tag{4.21}$$

where the term $\gamma(l,k)$ is interpreted as the *a posteriori SRR*.

The enhanced spectra obtained using the above relation may contain some negative values. The simplest solution is to half-wave rectifies these values to ensure a nonnegative magnitude spectrum. This nonlinear processing of negative values however creates small, isolated peaks in the spectrum

occurring at random frequency locations in each frame. Converted in the time domain, these peaks sound similar to tones with frequencies that change randomly from frame to frame, that is, tones that are turned on and off at the analysis frame rate. This type of noise is commonly referred as musical noise [33]. Many solutions have been proposed to tackle the problem of musical noise. In this work, two standard modifications are added to the algorithm to alleviate the problem of musical noise. The first modification consists of replacing the *a posteriori SRR* by the *a priori SRR* plus one [19, 157]. The modified spectral gain function becomes

$$H(l,k) = 1 - \frac{1}{\sqrt{1+\xi(l,k)}} \tag{4.22}$$

The *a priori SRR* $\xi(l,k)$ is calculated as (decision directed approach) [14]

$$\xi(l,k) = \eta \frac{\left|\hat{S}(l-1,k)\right|^2}{\hat{S}_{zz}(l-1,k)} + (1-\eta)\max\left\{\gamma(l,k) - 1, 0\right\} \tag{4.23}$$

Here, the value of η is chosen as 0.98 as suggested in [14, 19].

The second modification consists of using a spectral floor to prevent the gain from descending below a lower bound, as proposed in [33].

$$|\hat{S}(l,k)| = \begin{cases} |Z(l,k)|\, H(l,k), & \text{if} |Z(l,k)|H(l,k) > \beta |Z(l,k)| \\ \beta\, |Z(l,k)|\,, & \text{otherwise} \end{cases} \tag{4.24}$$

where β is the spectral floor factor and is chosen as 0.02. Finally, from the modified magnitude spectrum and original phase, the enhanced signal is recovered by the inverse discrete Fourier transform (IDFT) and overlap-add technique.

4.5 Temporal Processing of Reverberant Speech

As mentioned earlier, temporal processing refers to the processing of excitation source information in the temporal domain which mainly involves identification and enhancement of higher SRR regions. In case of reverberation, the residual signal is distorted by introducing random peaks to the periodic peaks representing the GC instants in clean speech [22]. The random peaks represent the reflected versions of earlier GC instants arriving at random time. By attenuating the random peaks, the difference between the residual signal and the non-reverberant residual signal can become smaller. This is the main objective of the temporal processing. In this work, the temporal processing is performed at two levels, namely the gross and fine temporal levels. The gross level processing helps to identify the high SRR regions of spectrally processed speech. Once the high SRR regions are identified, the epoch locations related to the direct components are enhanced relative to the random peaks with the help of fine level temporal processing.

4.5.1 Gross Level Temporal processing

The high SRR regions at the gross level are identified using the sum of the 10 largest peaks in the DFT spectrum, the smoothed HE of the LP residual and the modulation spectrum values of the spectrally-processed reverberant speech signal as described in Chapter-3. For illustration, Figs. 4.1 (c)-(e) show the normalized sum of peaks in the DFT spectrum, smoothed HE of the LP residual and modulation spectrum values of the spectrally processed reverberant signal given in Fig. 4.1(b). The indicators of the high SRR regions are further enhanced using FOD of the indicators obtained and then combined to identify the gross level features. The enhanced high SRR indicator plots for the sum of peaks in the DFT spectrum, the smoothed HE of the LP residual, and the modulation spectrum values of the speech signal shown in Figs. 4.1(c)-(e) are plotted in Figs. 4.2(b)-(d), respectively. The normalized sum and the nonlinearly mapped values (i.e., gross weight function) are given in Figs. 4.2(e) and (f), respectively.

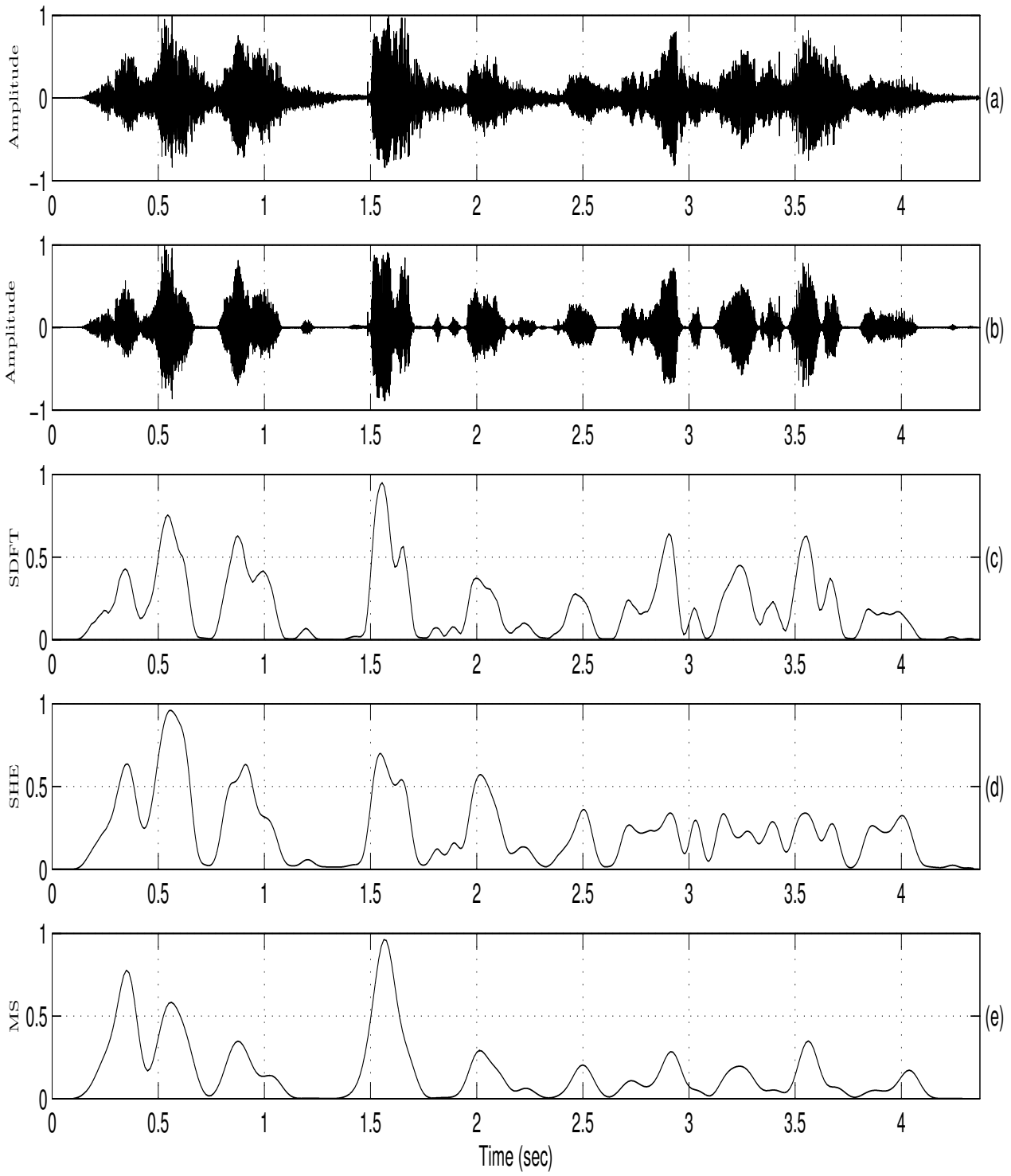


Figure 4.1: Gross level features: (a) Reverberant speech, (b) spectrally processed speech, (c) normalized sum of peaks in the DFT spectrum (SDFT), (d) normalized smoothed Hilbert envelope (SHE) of the LP residual and (e) normalized modulation spectrum (MS).

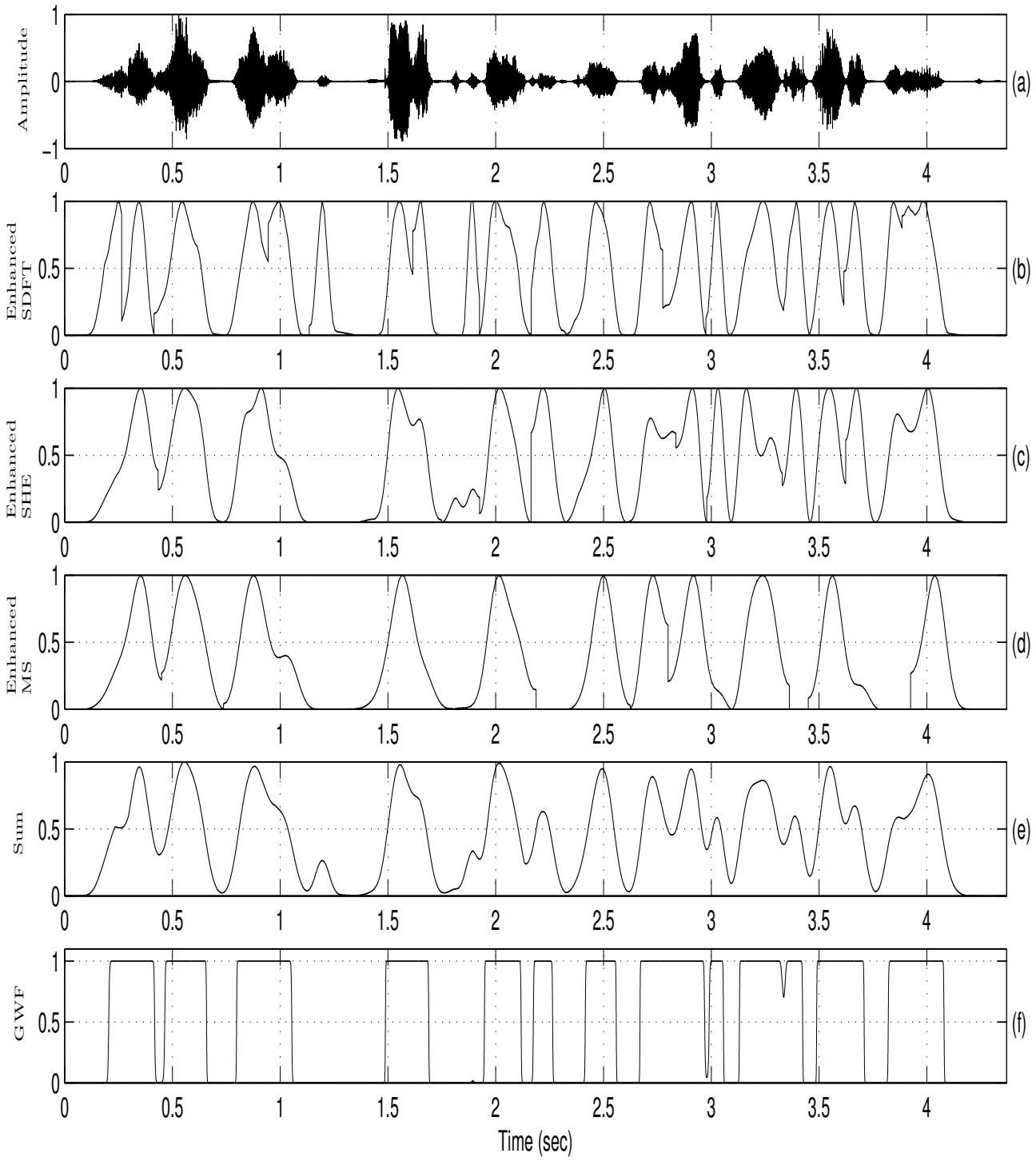


Figure 4.2: Gross level features identification: (a) spectrally processed speech, (b) enhanced sum of peaks in the DFT spectrum values, (c) enhanced smoothed HE values, (d) enhanced modulation spectrum values, (e) normalized sum and (f) nonlinearly mapped values.

4.5.2 Fine Level Temporal processing

The main objective of fine level processing is the identification and enhancement of regions around the instants of significant excitation in the LP residual. By locating these instants, it is possible to enhance speech in the region of the instants, relative to other regions [22, 23]. The sequence of these impulse-like excitations is robust to degradation in the sense that the relative spacing of these epochs due to the direct sound remains unchanged. On the other hand, the impulse-like excitations due to reflected sounds occur at random locations [23]. In high SRR regions, the direct component of the signal generally dominates over the reverberant component and enhancement can be achieved by identifying and enhancing the direct component locations. In the presence of reverberation, the HE of the LP residual can be viewed as the sum of periodic impulse-like excitation sequence due to the direct component and random impulse-like sequence due to the reverberant component [23].

In the proposed method, to determine the approximate locations of the instants of significant excitation, the LP residual (0 to 4 kHz components) of the speech signal is separated into four subbands equally spaced on a linear scale. The relative positions of these instants of significant excitation in the direct component of the speech signal remain unchanged in each band. In the next step, the HE of the LP residual is determined for each subband. From a frequency domain point of view, the HE has a frequency spectrum in the baseband for all cases. Also the relative spacing of the direct signal remains unchanged in each band; whereas the peaks related to the reverberant component occur at random positions because of the random noise-like spectrum in each band. If we sum the HE of all the bands, the peaks related to the direct signal are enhanced compared to the reverberant component, so that, the resultant summed signal can be used to identify the instants of the significant excitation.

The filter with passband of 1 kHz is chosen based on the assumption that the maximum fundamental frequency of human speech signal does not exceed 400 Hz and hence in the spectral domain each band of signal will have at least two peaks relative to the direct component. In the high SRR regions, the periodicity information is more affected in the high frequency region of the excitation source spectrum, compared to the low frequency region as shown in Figs. 4.3 and 4.4. In addition, the higher order harmonic components will have less energy than the lower order harmonics [16]. Fig. 4.3(a) shows a 100 ms portion of the LP residual of high voiced speech frame and Figs. 4.3(b)-(d) show the LP residual of the reverberant speech with the reverberation time of 0.25, 0.5, 0.75 and 1 sec, respectively. The corresponding spectra of the signals shown in Figs. 4.3(a)-(e) are given in Figs. 4.3(f)-(j),

respectively. Fig. 4.4 shows the HE and their spectra for the respective signals shown in Fig. 4.3. From these two figures it can be observed that in the high SRR regions the high frequency region of the excitation source spectrum is more affected than the low frequency region. Figs. 4.5(a)-(d) depict the spectra of the HE of clean speech LP residual obtained from each band (i.e., 0-1, 1-2, 2-3 and 3-4 kHz bands) and Fig. 4.5(e) shows the sum of all 4 bands HE spectra. Figs. 4.5(f)-(j) show the same with reference to spectrally processed reverberant speech with reverberation time of 1 sec. From Fig. 4.5(j) it can be observed that the spectra of the sum of the HE shows significant improvement in the direct component locations as compared to the reverberant component. Note that in all spectra related to the HE, the dc value is neglected when plotting, for better visualization.

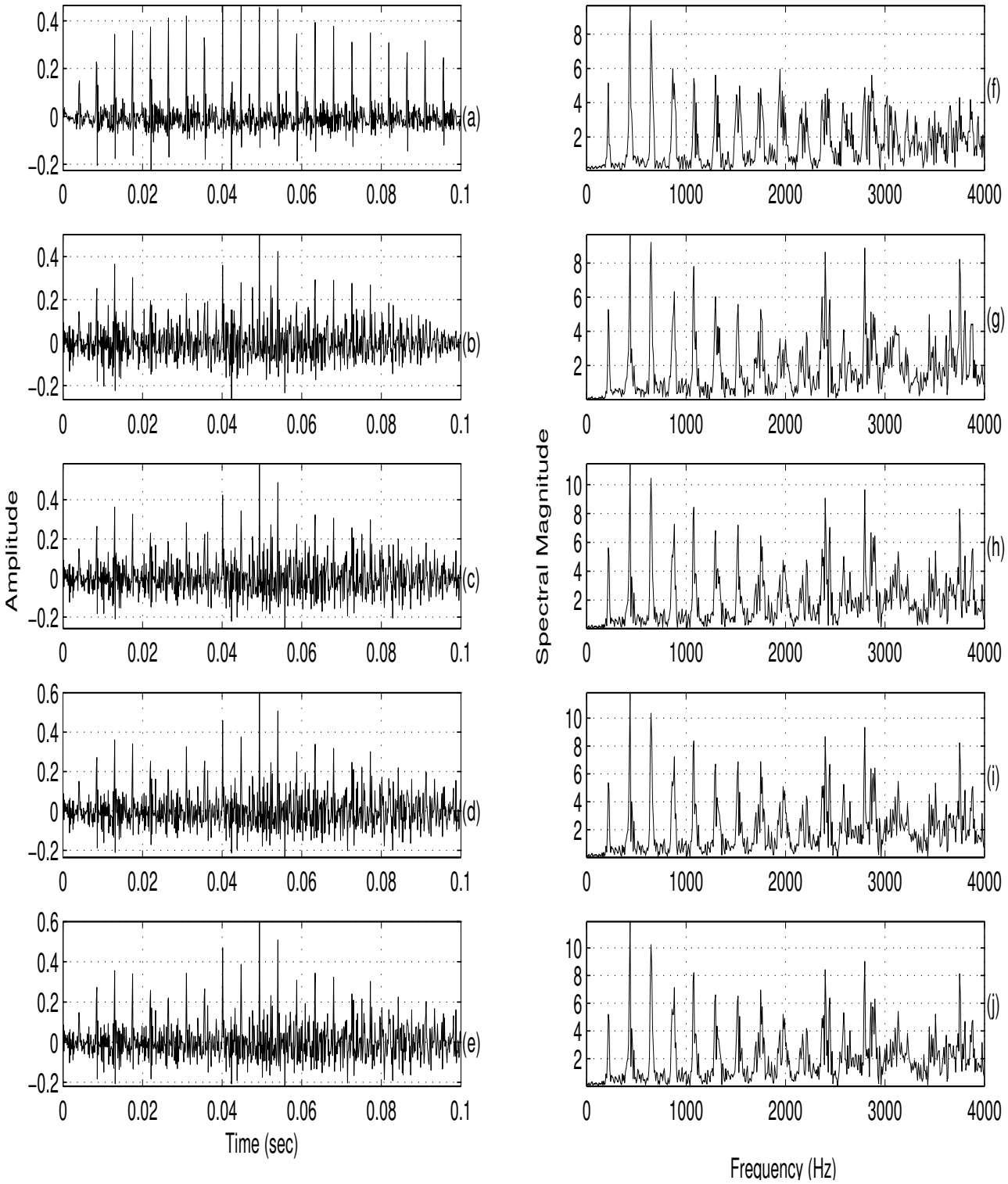


Figure 4.3: LP residual of (a) direct signal, (b)-(e) spectrally processed reverberant speech with T_{60}=0.25 sec, 0.5 sec, 0.75 sec and 1 sec, respectively. (f)-(j) Spectrum of the respective signals shown in (a)-(e).

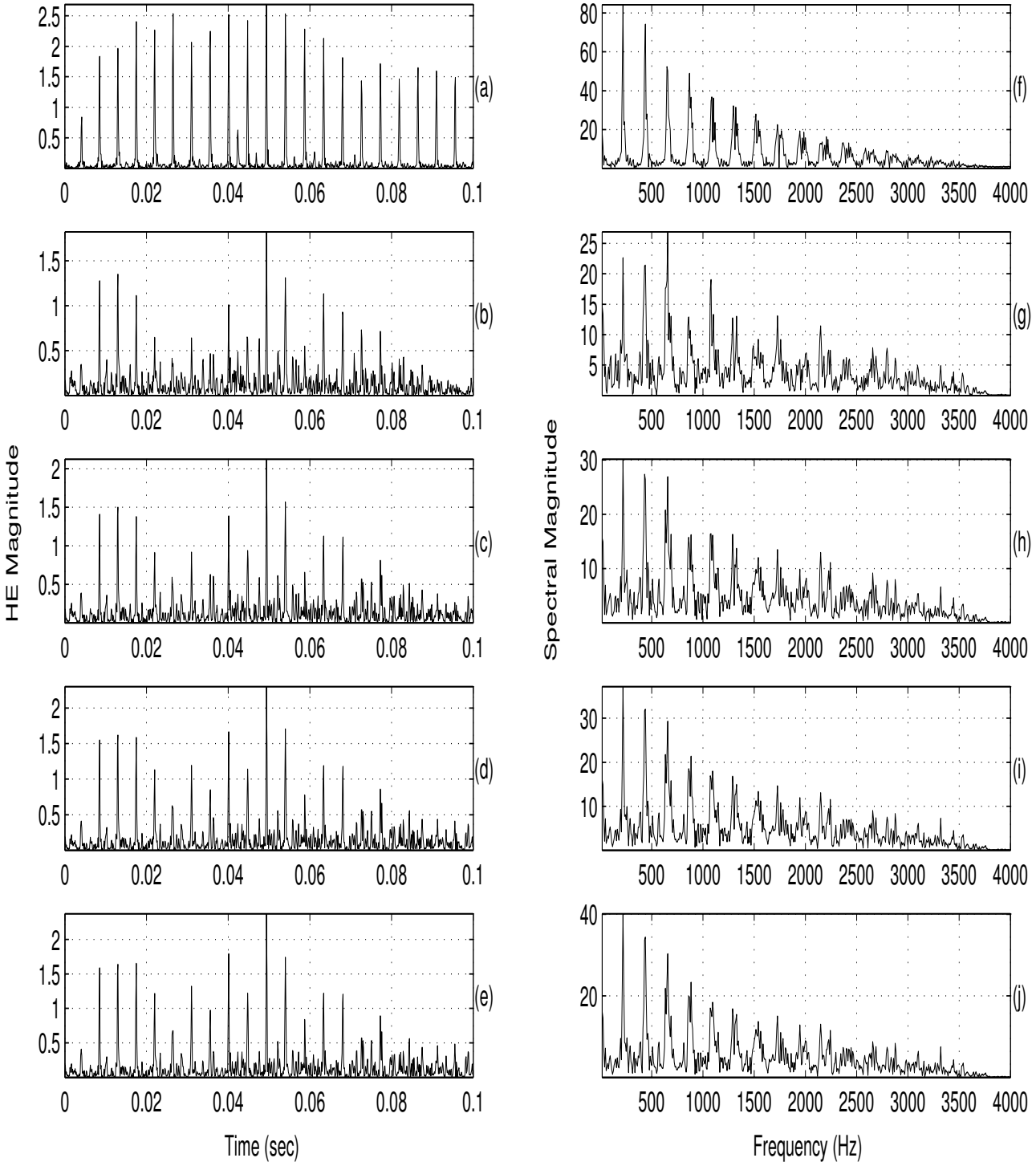


Figure 4.4: HE of the LP residual of (a) direct signal, (b)-(e) spectrally processed reverberant speech with T_{60}=0.25 sec, 0.5 sec, 0.75 sec and 1 sec, respectively. (f)-(j) Spectrum of the respective signals shown in (a)-(e).

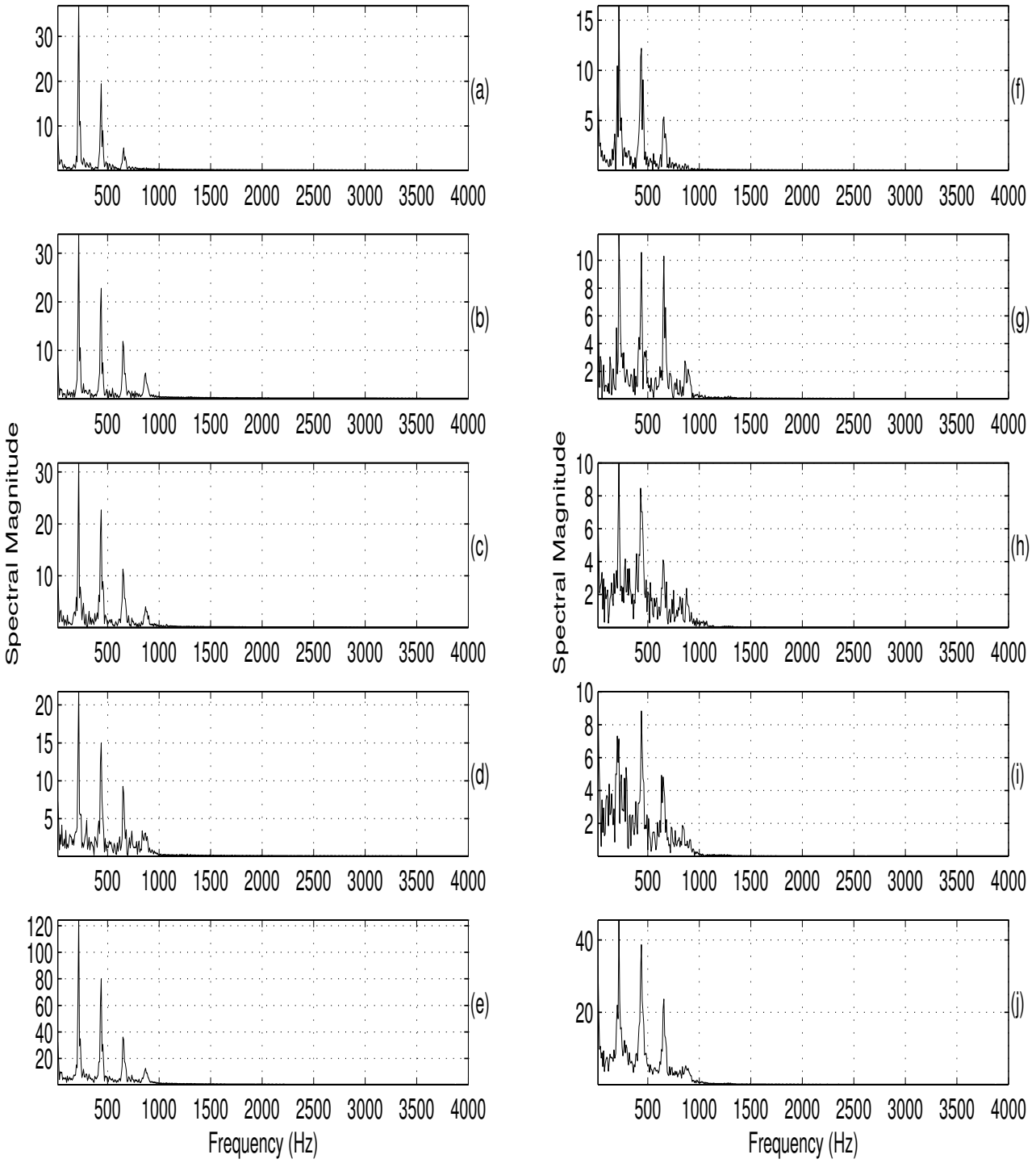


Figure 4.5: Spectrum of the HE of LP residual of direct signal in (a) 0-1, (b) 1-2, (c) 2-3, and (d) 3-4 kHz bands (e) sum of (a)-(d). Spectrum of Hilbert envelope of LP residual of spectral processed reverberant speech in (f) 0-1, (g) 1-2, (h) 2-3, and (i) 3-4 kHz bands, (j) sum of (f)-(i).

In the next step, the HE of all 4 bands is summed and the peak locations in the summed HE representing approximate locations of the instants of significant excitation are detected by convolving the HE with a FOGD operator. Some of the spurious instants that are detected from the small random peaks in the HE are eliminated by computing the measure called peak-to-sidelobe energy ratio which is similar to the measure called peak-to-sidelobe ratio (PSR) employed in [301]. The peak-to-sidelobe energy ratio (PSLER) of the signal $x(n)$ is defined as

$$PSLER = \frac{\sum_{n-l_1}^{n+l_1} x^2(n)}{\sum_{n-l_1-l_2}^{n-l_1-1} x^2(n) + \sum_{n+l_1+1}^{n+l_1+l_2} x^2(n)} \tag{4.25}$$

where l_1 and l_2 are set to values of 4 and 16, respectively. These values are chosen by assuming that the normal range of pitch period of human speakers is in the range of 2.5 - 10 ms. The PSLER measure gives the strength of the main peak in relation to the values around the peak. If there exists another major peak within 2.5 ms interval of the main peak being considered, the logarithm of PSLER value will become negative and the instant corresponding to the negative PSLER value is eliminated. This is illustrated in Fig. 4.6. Fig. 4.6(b) shows the approximate instant locations derived by convolving the HE of the LP residual shown in Fig. 4.6(a) with FOGD operator. Fig. 4.6(c) shows the logarithm of PSLER values at detected peak locations. Fig. 4.6(d) shows final approximate instant locations after discarding the instants with negative PSLER values.

In unvoiced segments also there may be epochs due to strong bursts of excitation, even though they may not occur at periodic intervals as in the voiced case. But their relative locations are unaffected by degradation [301]. Fig. 4.7 illustrates the computed fine weight function of an unvoiced speech region. As can be seen from the figure, even though most of the instants are correctly detected with reference to the clean speech HE, there are some spurious and undiscovered instants. Therefore it should be further investigated to improve the performance. A fine weight function is derived to enhance the region around the instants of significant excitation by convolving them with the Hamming window of 3 ms duration. Due to high strength of excitation, the SRR of speech is high around 3 ms as compared to other regions [8]. The minimum value of the fine weight function is set to 0.4 to reduce perceptual distortion. The final weight function for the LP residual of spectral processed speech is derived by multiplying the gross weight function with the fine weight function.

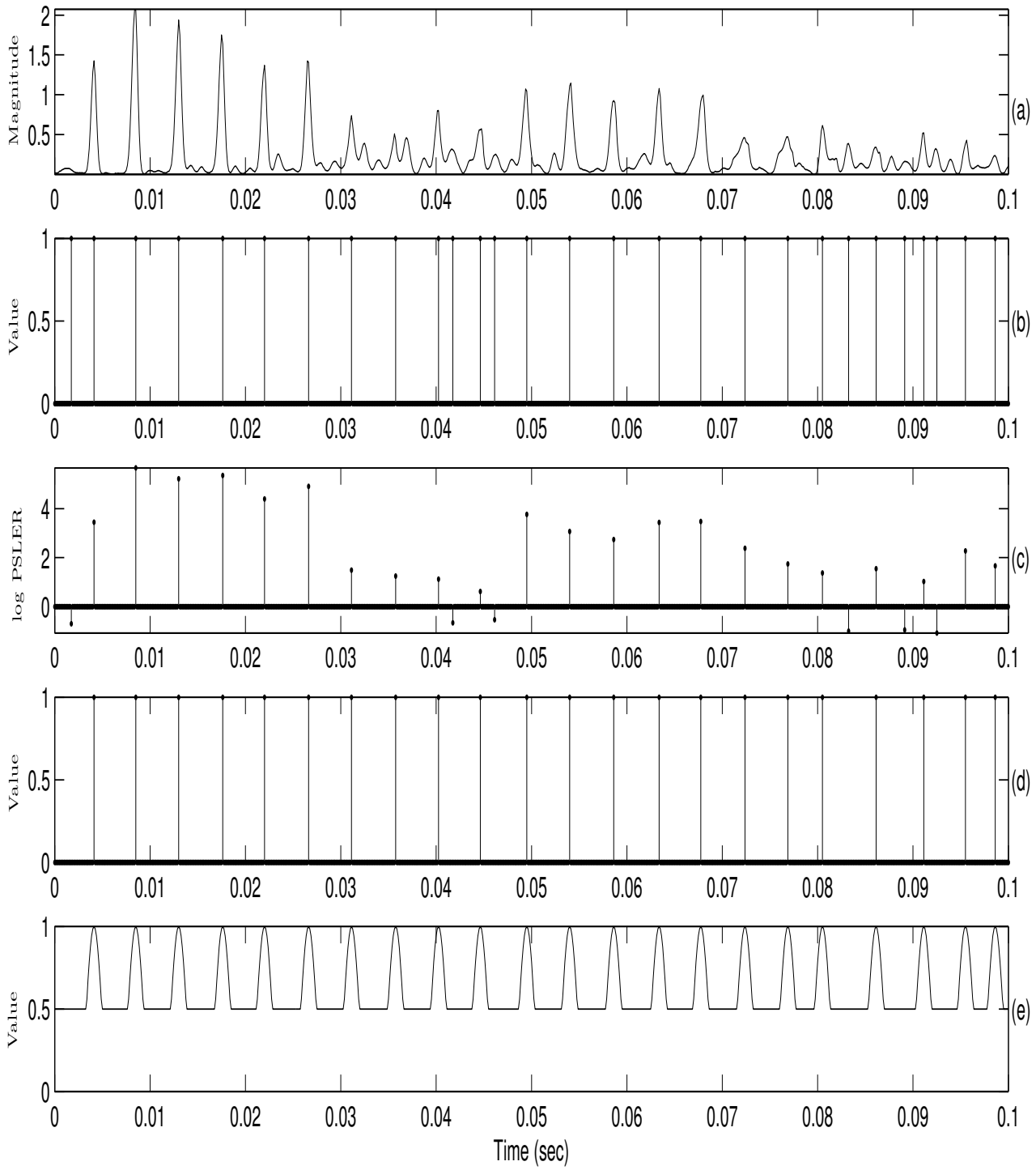


Figure 4.6: Fine weight function determination: (a) HE of the LP residual, (b) GC instants obtained from FOGD operator, (c) log peak to sidelobe energy ratio (log PSLER), (d) approximate GC instants, and (e) fine weight function.

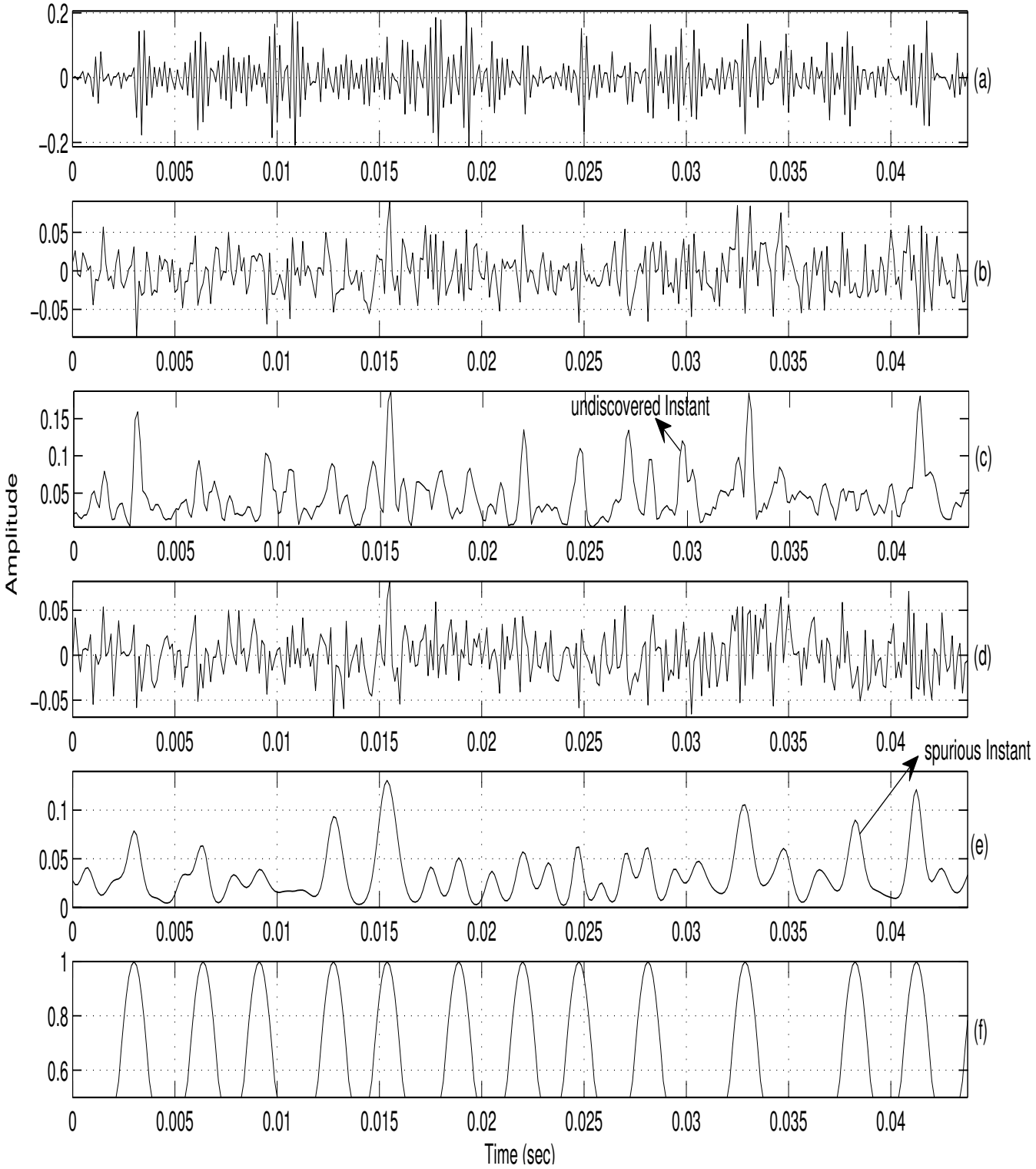


Figure 4.7: Fine weight function determination for an unvoiced speech: (a) portion of unvoiced direct signal and its (b) LP residual, (c) HE of the LP residual, (d) LP residual of the reverberant speech, (e) HE of the LP residual obtained by the proposed method for the signal shown in (d), and (f) fine weight function.

The complete set of temporal processing steps are illustrated in Fig. 4.8. Fig. 4.8(a) shows a small portion of the LP residual of spectrally processed speech shown in Fig. 4.1(b), the respective gross weight function and the combined weight function are given in Figs. 4.8(b) and (c). The LP residual of spectrally processed speech is then multiplied with the combined weight function to obtain enhanced residual. The time-varying all-pole filter derived from the spectrally processed speech is excited by the enhanced residual to obtain the temporally processed speech. The residual samples are weighted rather than the speech samples, mainly because (i) the residual samples are less correlated and (ii) weighting the residual samples may lead to less perceptual distortion [8]. The block diagram shown in Fig. 4.9 illustrates the various steps involved in the combined temporal and spectral processing method.

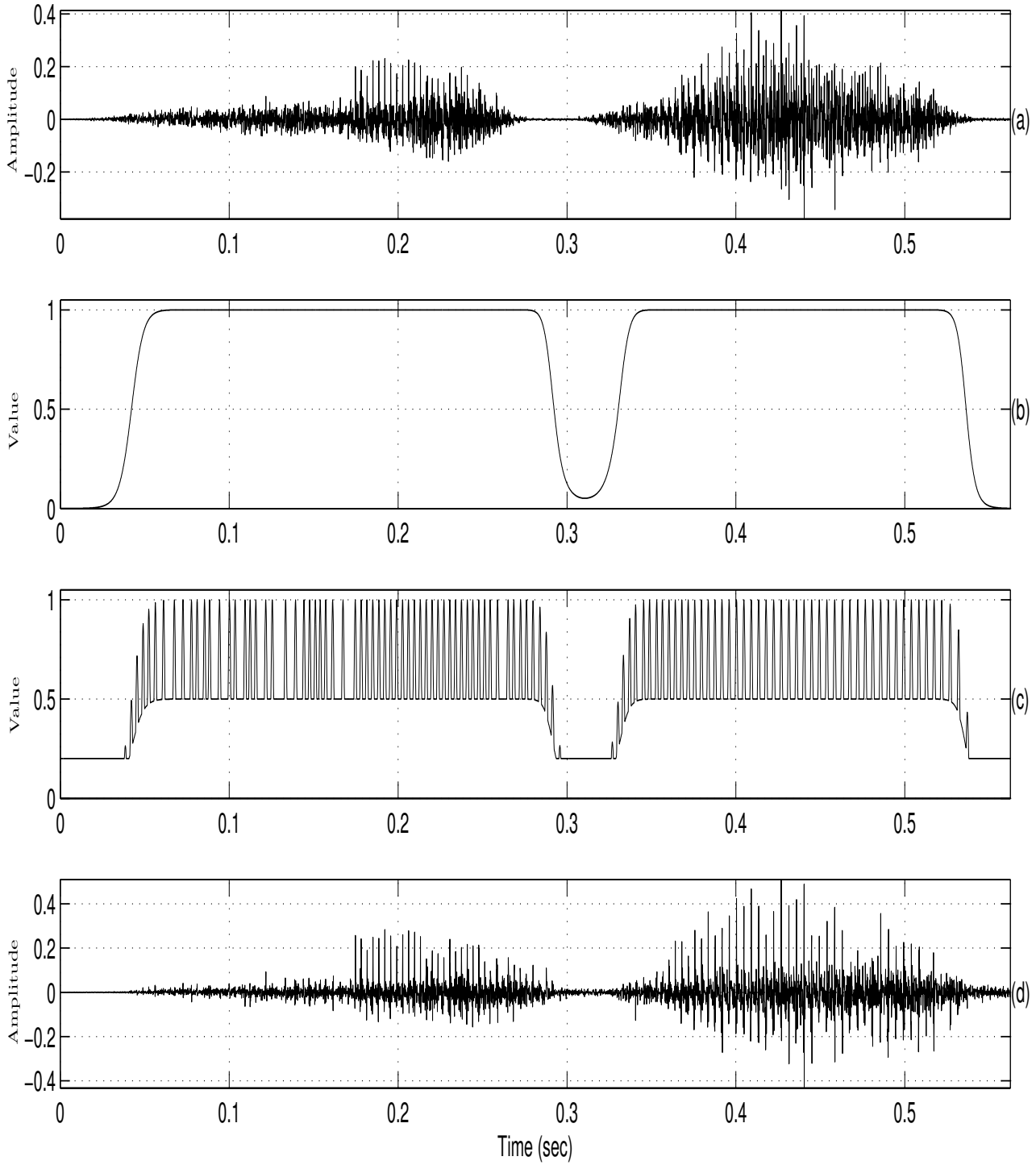


Figure 4.8: Temporal Processing: (a) LP residual of spectral processed speech, (b) gross weight function, (c) combined weight function, and (d) enhanced residual obtained by weighting.

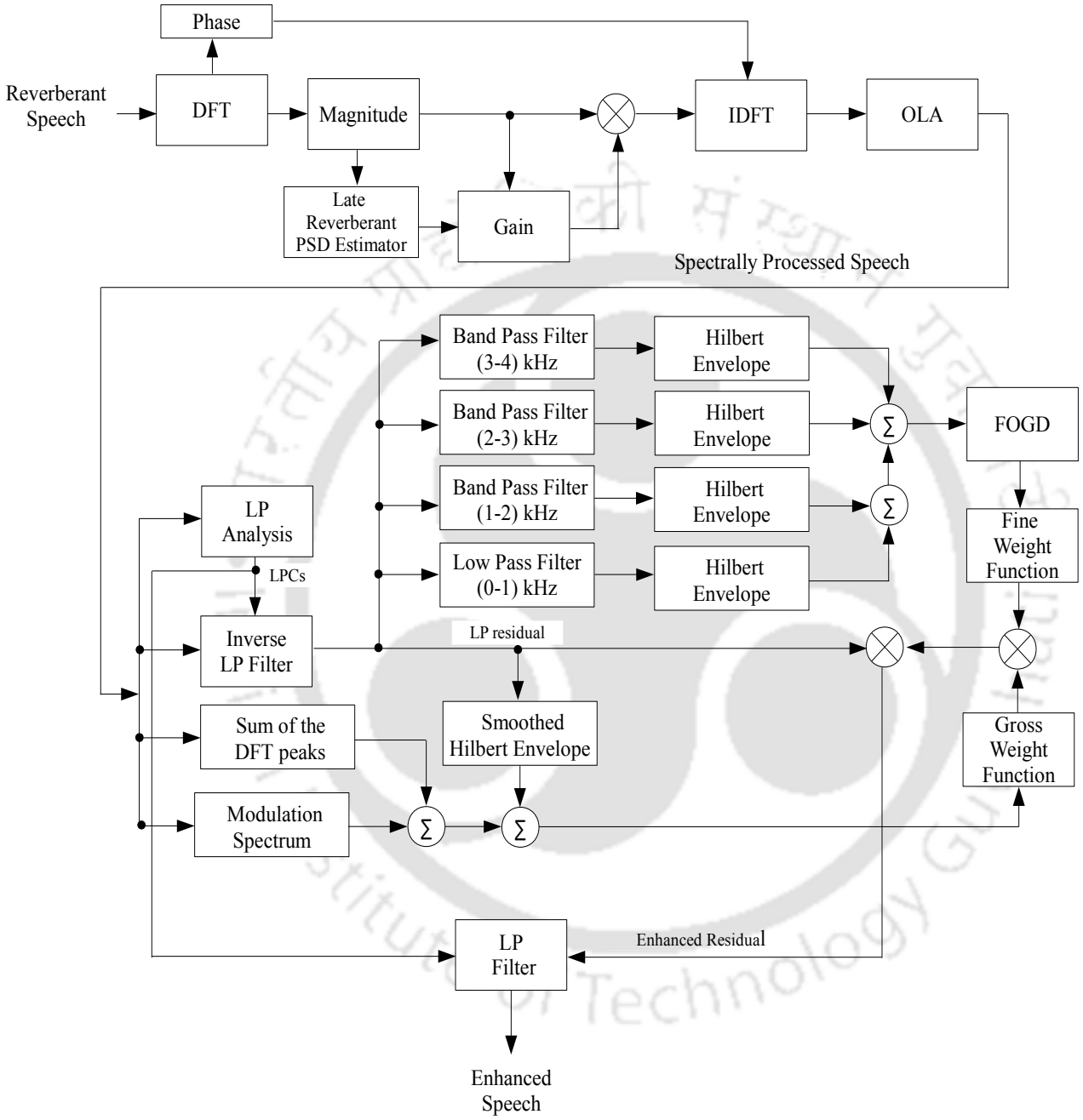


Figure 4.9: Block diagram of the proposed combined temporal and spectral processing method for reverberant speech enhancement.

4.6 Experimental Results and Performance Evaluation

4.6.1 Time Domain Waveforms and Spectrograms

An experiment is carried out to illustrate the objective of the proposed method using simulated room impulse response. The image-source model is a well-known technique that can be used in order to simulate the synthetic room impulse response, i.e., a transfer function between a sound source and an acoustic sensor, in a given environment [302–304]. Once room impulse response is available, reverberant speech data can be obtained by convolving the room impulse response with the given speech signal. The MATLAB implementation of image method can be found in [305].

For illustration, speech example is taken from the TIMIT database [289, 290]. A non-minimum phase room impulse response is synthesized using the image method to model an office-size room with dimensions $6 \times 4 \times 3$ m in length $\times$ width $\times$ height. Figs. 4.10(a) and (b) show the clean speech and the corresponding reverberant speech obtained by convolving the obtained impulse response with a reverberation time of about 1 sec. The speech processed by the temporal processing, spectral processing and combined temporal and spectral processing are given in Figs. 4.10(c)-(e), respectively. The spectrograms of the signals shown in Figs. 4.10(a)-(e) are given in Figs. 4.10(f)-(j), respectively. From Fig. 4.10 it can be inferred that the spectral processing removes the late reverberant speech components whereas the temporal processing fails to remove the late reverberation components in some portions. Figs. 4.11(a)-(e) show frames of high SRR region of the LP residual of direct, reverberant, temporal processed, spectral processed and the combined temporal and spectral processed speech signals, respectively. From Fig. 4.11 it can be seen that, with reference to the direct component instant locations (pointed by the down arrow symbol in the direct speech LP residual), the LP residual of spectral processed speech does not show any noticeable improvement in reference to the degraded one. On the other hand, the temporal processing shows an improvement around the major instant locations. The combined temporal and spectral processing method therefore gives better enhancement in high SRR regions and also suppression of late reverberation.

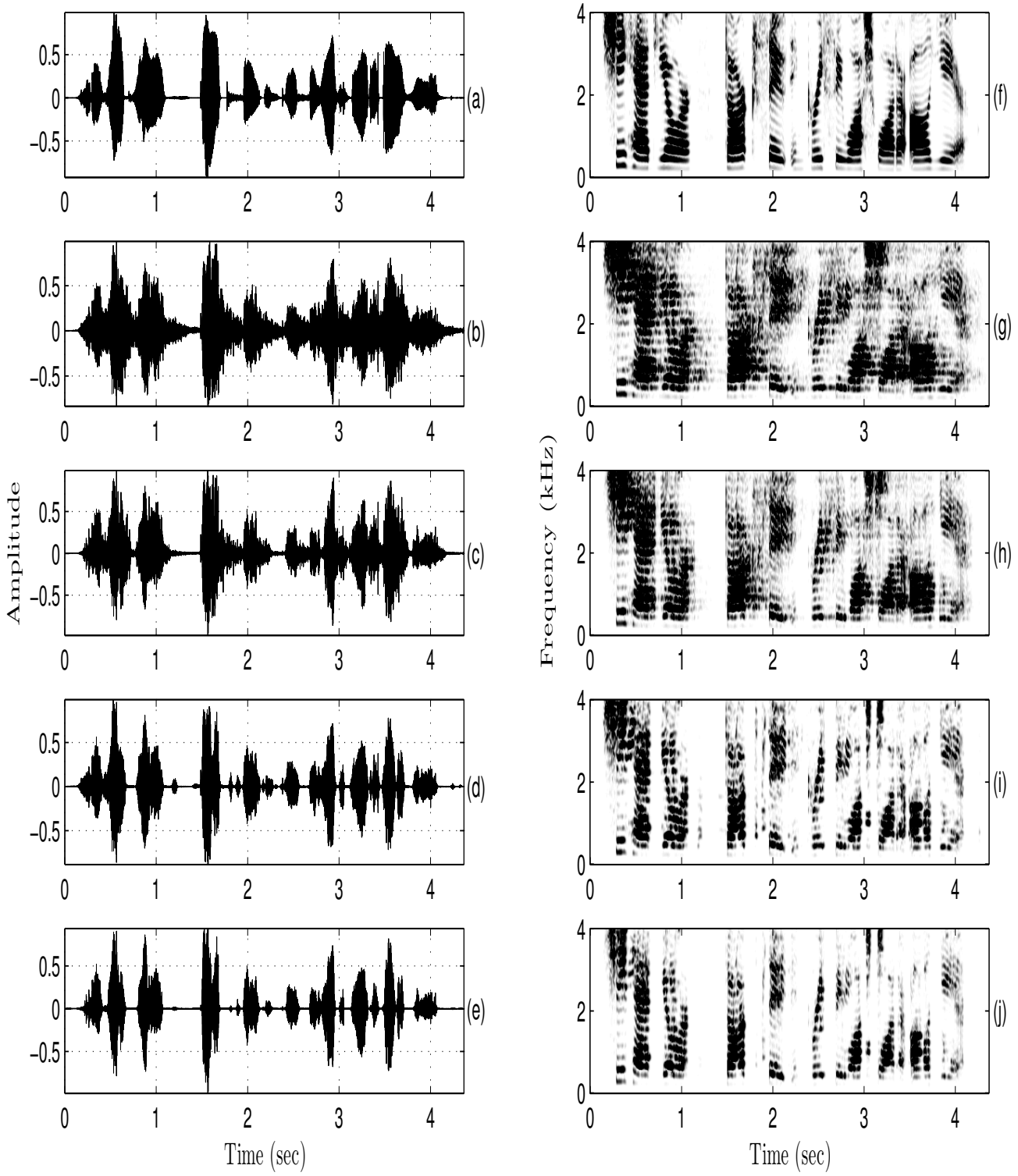


Figure 4.10: Results of enhancement of reverberant speech of a female voice: (a) clean speech, (b) degraded speech, (c) speech processed by temporal processing, (d) speech processed by spectral processing, (e) speech processed by spectral and temporal processing and (f)-(j) spectrograms of the respective signals shown in (a)-(e).

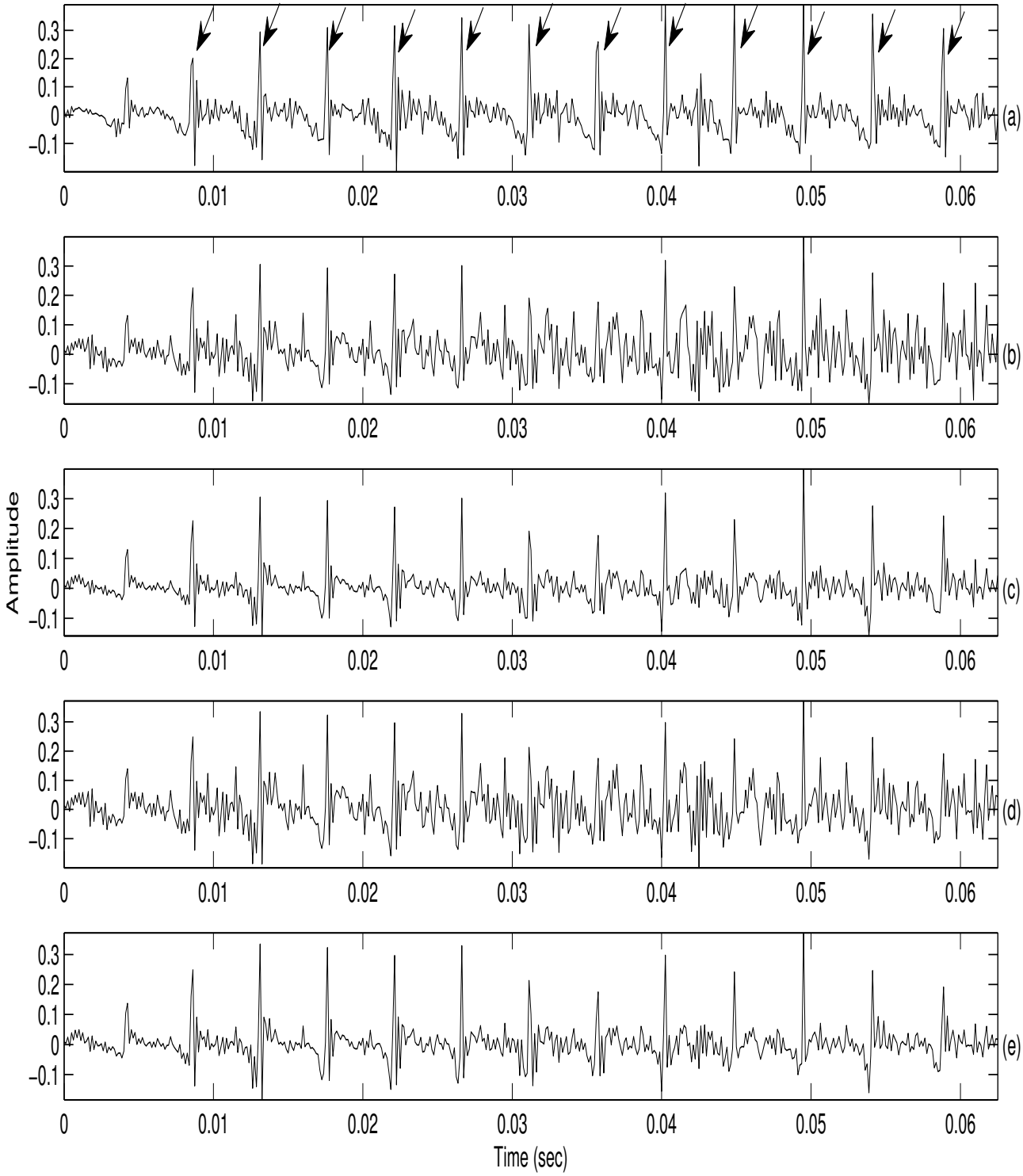


Figure 4.11: LP residual of a frame of: (a) direct signal, (b) reverberant speech, (c) speech processed by temporal processing, (d) speech processed by spectral processing, and (e) speech processed by temporal and spectral processing.

4.6.2 Gross Weight Function Performance

Various experiments are carried to evaluate the performance of gross and fine level processing algorithms by considering five male and five female speech examples from the TIMIT database. First, the performance of the gross weight function detection algorithm is evaluated using the manually marked high SRR and low SRR regions. For comparison, the gross weight function is computed for each of the parameters independently and also for the combined one. Table 4.1 shows the percentage correct detection accuracy (P_c) under different reverberation times (0.2 - 1.0 sec) by considering source microphone distances of 1, 1.5 and 2 m.

As mentioned earlier, the reverberant tail portions of the degraded speech make it difficult to identify high and low SRR regions. To evaluate the improvement in the accuracy of the gross level detection using spectral processing, the same gross level processing method is applied to both reverberant and spectrally processed speech. The probability of the false detection rate P_f of the combined gross level detection method for reverberant and spectrally processed speech are computed. The results are shown in Fig. 4.12. As can be seen in Fig. 4.12, the spectral processing significantly reduces the P_f values as compared to the reverberant speech.

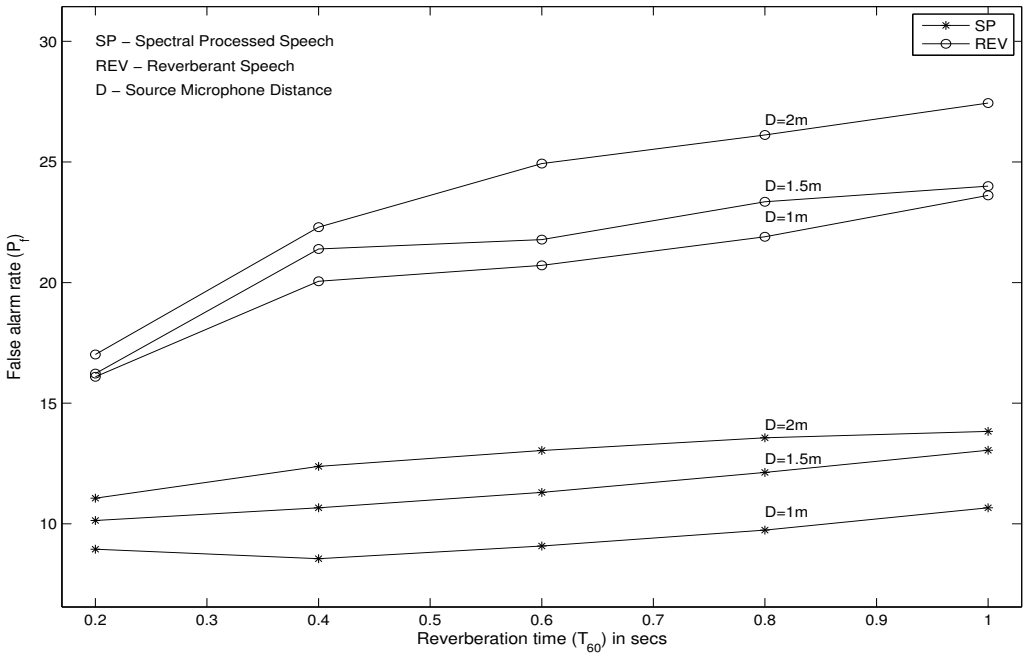


Figure 4.12: False detection rate (P_f) of gross level features.

Table 4.1: Gross weight function performance. In the table, abbreviations SDFT, SHE, MS and COMB refer to sum of peaks in the DFT spectrum, smoothed HE of the LP residual, modulation spectrum and combination of all three parameters, respectively.

T_{60}(sec)	SDFT	SHE	MS	COMB
	Source Microphone Distance=1 m			
0.2	86.00	80.71	79.76	86.77
0.4	85.45	80.43	79.60	87.05
0.6	85.31	81.27	79.04	86.49
0.8	85.03	81.75	78.69	85.38
1.0	85.58	81.69	78.62	86.42
	Source Microphone Distance=1.5 m			
0.2	86.56	81.82	79.11	87.60
0.4	86.42	83.15	78.97	87.88
0.6	85.65	83.64	77.92	87.12
0.8	85.45	82.66	78.48	86.35
1.0	85.24	83.43	78.13	87.05
	Source Microphone Distance=2 m			
0.2	86.70	82.38	78.90	87.81
0.4	85.65	83.98	78.34	87.81
0.6	84.47	83.57	77.30	86.70
0.8	81.27	83.36	77.02	84.75
1.0	80.92	85.65	76.88	84.12

4.6.3 Fine Weight Function Performance

The performance of the fine weight function detection method is evaluated by computing the deviation in the approximate instants location of the spectrally processed speech with respect to the instants of direct signal. The result of the percentage of accuracy analysis is given in Table 4.2. The entries in the Table 4.2 show the percentage of approximate instants and their deviation with respect to the direct signal instants location. Most of the spectrally processed speech instants detected from the HE of the LP residual lie within 2 ms of the instants of direct signal.

A further experimental evaluation is performed to analyze the fine weight function performance by considering the sum of HEs over subbands and the HE of the fullband signal. Table 4.3 shows the percentage of approximate instants and their deviation with respect to the direct signal instants location (for a source microphone distance of 2 m). The sum of HE over subbands gives relatively higher performance than the full band signal HE. In particular, for lower deviations (0.5 ms and 1 ms) more instants are detected with higher accuracy (i.e., in terms of smaller deviation with reference to

the direct signal instants), when we consider the sum of HEs over subbands. Fig. 4.13 illustrates the nature of the HE for these two cases with reference to the direct signal HE of the LP residual.

Another experiment is carried out to analysis the performance by considering only lower band(s) and sum of all four bands. The results are tabulated in Table4.4. From the table it can be inferred that, the HE of all four bands provides relatively better performance compared to considering lowest band (0-1 kHz) alone. Similar to the previous observation, here also it can seen that if we consider the deviation of 0.5 ms and 1 ms, the results show the improved performance in terms of number of instants that are detected with less deviation with reference to direct signal instants. This shows even though higher bands are more affected by the reverberation, the periodicity information (related to direct component) present in the higher bands HE may further increase the relative amplitude of the direct components compared to considering the lowest band alone. Fig. 4.14 illustrates the nature of the HE by considering only the lowest band (i.e., 0-1 kHz).

Table 4.2: Percentage of approximate instants and their deviation with respect to the direct signal instants location.

T_{60}**(sec)**	Deviation in time			
	0.5 ms	**1.0 ms**	**1.5 ms**	**2.0 ms**
	Source Microphone Distance=1 m			
0.2	69.21	82.70	92.23	96.71
0.4	68.87	82.56	91.49	96.66
0.6	67.37	82.36	91.28	96.39
0.8	65.94	81.20	91.21	96.19
1.0	64.10	79.70	90.26	95.37
	Source Microphone Distance=1.5 m			
0.2	64.71	78.27	91.35	95.78
0.4	61.58	77.25	90.05	95.50
0.6	58.65	77.04	89.31	95.10
0.8	56.61	74.52	89.24	94.89
1.0	56.20	72.34	88.22	94.62
	Source Microphone Distance=2 m			
0.2	57.33	75.82	89.65	95.78
0.4	56.03	73.30	89.31	95.16
0.6	55.52	71.46	89.03	94.37
0.8	54.22	71.19	87.06	93.66
1.0	52.79	69.55	86.15	93.04

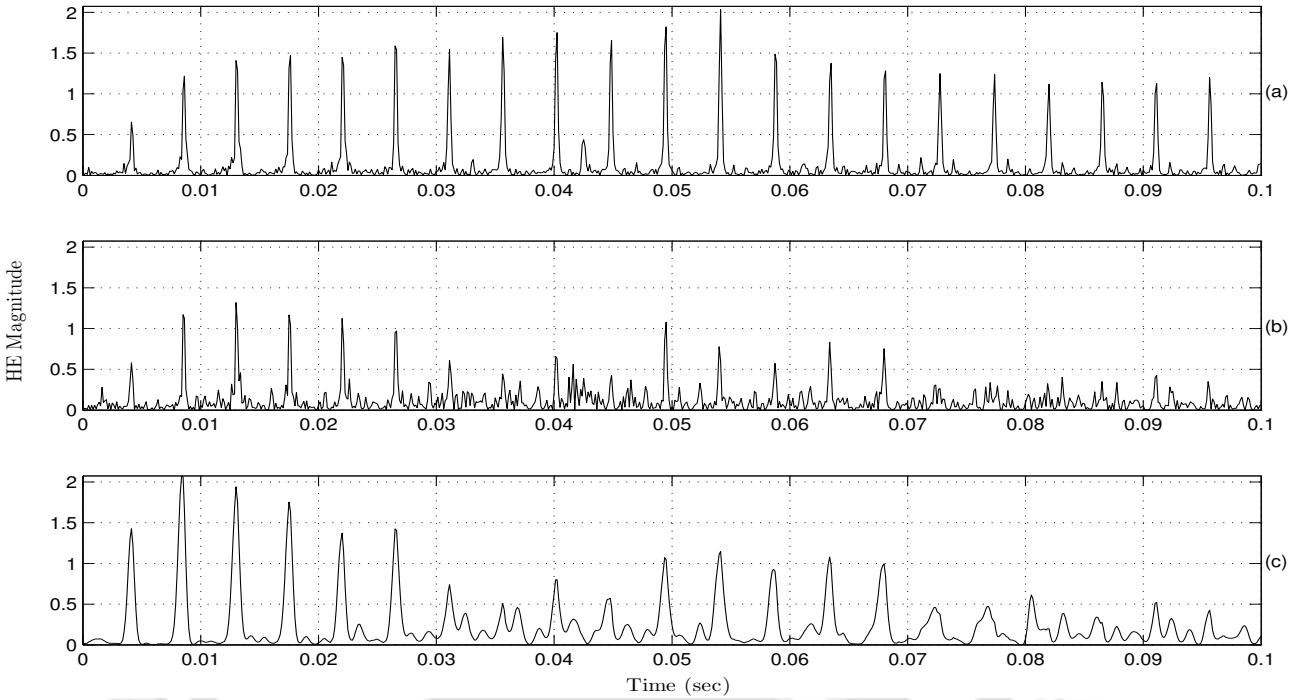


Figure 4.13: HE of the LP residual of (a) direct signal, (b) full band reverberant signal and (c) sum computed over subbands.

Table 4.3: Percentage of approximate instants and their deviation with respect to the direct signal instants location for a source microphone distance of 2 m.

	Deviation in time				Deviation in time			
T_{60}**(sec)**	**0.5 ms**	**1.0 ms**	**1.5 ms**	**2.0 ms**	**0.5 ms**	**1.0 ms**	**1.5 ms**	**2.0 ms**
	Sum of the HE of subbands				HE of fullband signal			
0.2	57.33	75.82	89.65	95.78	48.81	69.68	85.12	92.87
0.4	56.03	73.30	89.31	95.16	48.40	68.72	85.03	92.32
0.6	55.52	71.46	89.03	94.37	48.20	67.09	84.51	92.12
0.8	54.22	71.19	87.06	93.66	47.11	66.71	83.53	91.37
1.0	52.79	69.55	86.15	93.04	45.95	63.16	82.10	90.07

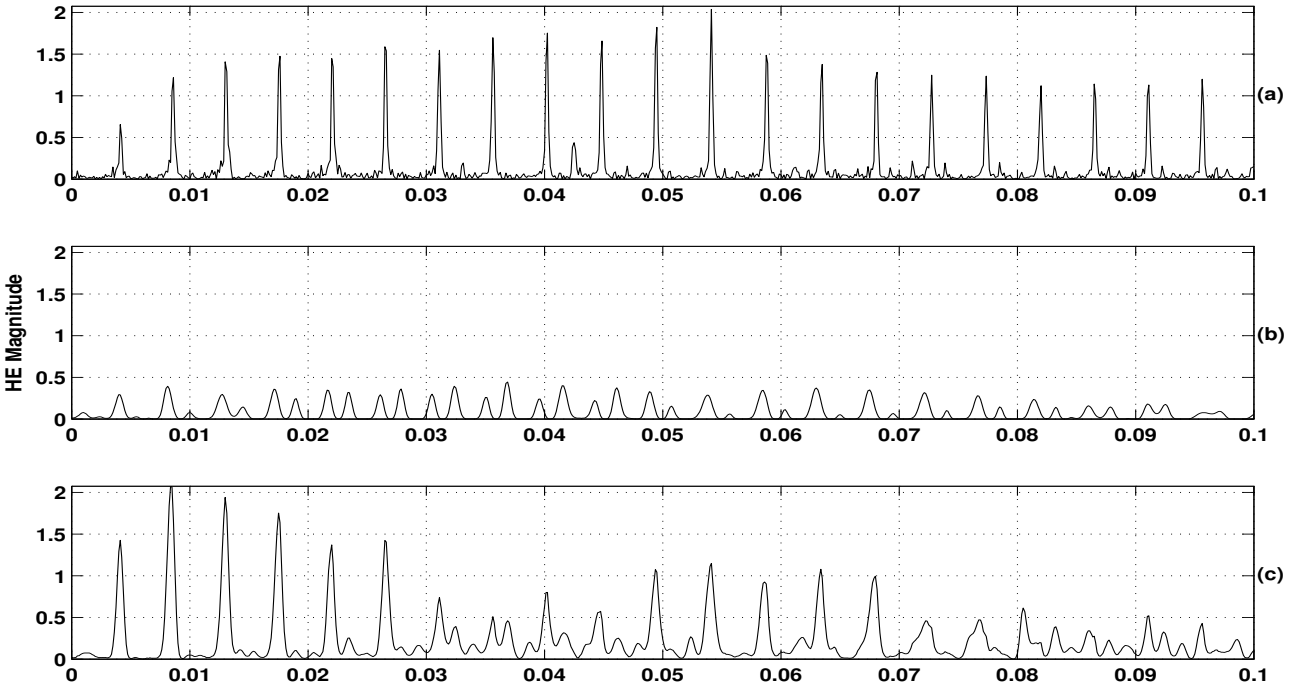


Figure 4.14: Illustartion of lowpass filtered HE: (a) HE of the LP residual of direct signal, (b) HE of the LP residual of the full band reverberant signal and (c) sum of HE computed over subbands

Table 4.4: Percentage of approximate instants and their deviation with respect to the direct signal instants location for a source microphone distance of 2 m.

T_{60}	**0.5 ms**	**1 ms**	**1.5ms**	**2.ms**	**T60**	**0.5 ms**	**1 ms**	**1.5ms**	**2.ms**
	1 Band [(0-1) kHz]					**2 bands [(0-1) &(1-2) kHz)]**			
0.2	39.85	67.71	82.97	92.03	**0.2**	48.68	70.14	87.83	94.84
0.4	39.65	67.03	82.76	92.03	**0.4**	48.09	68.27	86.03	93.25
0.6	39.99	66.21	82.08	91.89	**0.6**	47.09	64.86	85.9	93.73
0.8	37.74	61.58	79.83	90.8	**0.8**	46.93	63.91	84.1	91.62
1	36.04	58.24	76.22	89.1	**1**	45.43	62.46	84.06	91.28
	3 bands [(0-1),(1-2) & (2-3) kHz)]					**4 Bands**			
0.2	51.95	72.57	88.12	95.73	**0.2**	57.33	75.82	89.65	95.78
0.4	50.79	71.86	87.71	95.12	**0.4**	56.03	73.3	89.31	95.16
0.6	50.22	66.39	87.17	94.12	**0.6**	55.52	71.46	89.03	94.37
0.8	49.66	65.14	86.15	92.82	**0.8**	54.22	71.19	87.06	93.66
1	48.54	64.12	84.08	92.35	**1**	52.79	69.55	86.15	93.04

4.6.4 Objective Quality Measures

The proposed combined temporal and spectral processing method is evaluated using the segmental signal-to-reverberation ratio and log spectral distance (LSD) objective quality measures. The segmental SRR (SegSRR) of the l^{th} frame is defined as

$$SegSRR(l) = 10\log_{10}\left[\frac{\sum_{n=lL_r}^{lL_r+L-1} s_d^2(n)}{\sum_{n=lL_r}^{lL_r+L-1}\left(s_d(n)-\hat{s}(n)\right)^2}\right] \tag{4.26}$$

where $s_d(n)$ is the direct signal, which is the delayed version of the clean speech signal (i.e., $s_d(n) = s(n) * h_d(n)$, $h_d(n)$ is obtained from the known impulse response) and $\hat{s}(n)$ is the reverberant speech or enhanced speech signal. L is the number of samples per frame and L_r is the frame rate in samples. The mean segmental SRR is then obtained by averaging Eqn. (4.26) over all frames.

The LSD between the direct signal and the degraded/enhanced signal is obtained using the following expression

$$LSD(l) = \frac{2}{N}\sum_{k=0}^{\frac{N}{2}-1}\left|\pounds\left\{S_d(l,k)\right\} - \pounds\left\{\hat{S}(l,k)\right\}\right| \tag{4.27}$$

where $\pounds\left\{X(l,k)\right\} = 20\log_{10}\left(|X(l,k)|\right)$ and N is the number of points used for computing the FFT. The mean LSD is obtained by averaging Eqn. (4.27) over all the frames containing speech. The results of the segSRR and LSD for different reverberation times and source microphone distances are shown in Table 4.5 and 4.6, respectively. They are the average of measurements for five male and five female speech examples from the TIMIT database. In the table, the abbreviations T_{60}, REV, SP, TP, TSP1, TSP2, YM and TSA refer to reverberation time, reverberant speech, spectral processing, temporal processing, temporal and spectral processing (only with gross weight function), combined TSP (with overall weight function), conventional LP residual method [22] and two stage algorithm [20], respectively. The following observations can be made from the contents of Table 4.5 and 4.6:

(i) The combined TSP method shows improved performance, compared to the individual processing methods.

(ii) The proposed temporal processing method performs better, compared to the conventional single microphone-based temporal processing method proposed in [22]. The main difference between these two methods lies in the determination of weight function. In [22], the weight function is

derived mainly using information in the LP residual alone (gross weight function is derived from the entropy of the LP residual signal and fine weight function is derived from the normalized LP error), whereas in the proposed method, the weight function is derived using the information at different levels in the whole speech signal. Fig. 4.15 clearly illustrates the difference between these two weight functions.

(iii) The efficacy of the proposed method is also compared with the two stage algorithm proposed in [20]. The results show that the proposed method provides higher segmental SRR and lower spectral distortion scores, compared to the two stage algorithm, in most of the cases.

(iv) In the case of lower reverberation time (T_{60}=0.2 sec) and for smaller source microphone distance, the strength of the late reverberant echo is very low and the combined method results in slightly worse performance, compared to the individual processing methods. This may be due to overestimating the late reverberant spectrum. As a result of this, some of the low SRR regions get suppressed more and further weighting with the gross weight function reduces the signal level in the low SRR regions. This results in a lower objective score, compared to the individual processing. However, for this case also, the second level of the fine weight function increases the segmental SRR value as compared to the combined method with gross level processing alone (TSP1), by further enhancing the high SRR regions of the spectrally processed speech.

From the perception point of view, for lower reverberation times (0.2 - 0.6 sec), the enhanced speech does not exhibit any unbearable distortion. However, for higher reverberation time, especially for the source microphone distance of 2 m, the processed speech by the spectral subtraction and the combined method results some nonlinear distortion such as musical noise. Note that, for a typical room in home or office, the reverberation time ranges from 0.1 sec (slightly reverberant) to 0.6 sec (highly reverberant) [144]. Lastly, to study the selection of the smoothing factor (a) in the Rayleigh smoothing function of Eqn. (4.17), the segmental SRR and LSD values are computed for different values of a (a= 3 to 10, such that $a < N_1$) in different reverberant conditions and the results are given in Table 4.7. The results show only little improvement, compared to the original value of $a = 5$.

Table 4.5: Segmental SRR measure. In the table, abbreviations T_{60}, REV, SP, TP, TSP1, TSP2, YM and TSA refer to reverberation time, reverberant speech, spectral processing, temporal processing, temporal and spectral processing (only with gross weight function), temporal and spectral processing (with overall weight function), conventional Yegnanarayana and Murthy (YM) LP residual method and two stage algorithm, respectively.

T_{60}(sec)	REV	SP	TP	TSP1	TSP2	YM	TSA
	Source Microphone Distance=1 m						
0.2	1.65	3.02	3.54	3.00	3.46	2.42	3.49
0.4	-1.66	0.11	0.24	0.14	0.65	-0.92	0.62
0.6	-3.02	-0.72	-1.12	-0.64	-0.08	-2.31	-0.28
0.8	-3.77	-1.12	-1.91	-0.99	-0.46	-3.14	-1.57
1.0	-4.27	-1.44	-2.45	-1.24	-0.73	-3.71	-2.07
	Source Microphone Distance=1.5 m						
0.2	0.33	1.54	2.02	1.55	2.01	1.23	1.90
0.4	-2.64	-0.40	-0.67	-0.33	0.11	-1.68	-0.13
0.6	-3.87	-1.07	-1.94	-0.99	-0.61	-3.05	-0.82
0.8	-4.52	-1.39	-2.60	-1.19	-0.86	-3.76	-2.17
1.0	-4.98	-1.64	-3.12	-1.40	-1.06	-4.28	-2.36
	Source Microphone Distance=2 m						
0.2	-0.89	0.59	0.80	0.68	1.19	0.03	0.99
0.4	-3.99	-1.39	-2.20	-1.21	-0.86	-3.08	-1.01
0.6	-5.05	-2.17	-3.35	-2.00	-1.59	-4.30	-1.71
0.8	-5.59	-2.70	-3.95	-2.51	-2.07	-5.03	-3.16
1.0	-5.94	-3.05	-4.33	-2.96	-2.38	-5.50	-3.45

Table 4.6: LSD measure. In the table, abbreviations T_{60}, REV, SP, TP, TSP1, TSP2, YM and TSA refer to reverberation time, reverberant speech, spectral processing, temporal processing, temporal and spectral processing (only with gross weight function), temporal and spectral processing (with overall weight function), conventional Yegnanarayana and Murthy (YM) LP residual method and two stage algorithm, respectively.

T_{60}(sec)	REV	SP	TP	TSP1	TSP2	YM	TSA
	Source Microphone Distance=1 m						
0.2	5.83	5.68	4.88	5.71	5.29	5.67	5.27
0.4	9.32	7.56	7.42	7.49	6.90	8.35	6.91
0.6	11.04	8.22	9.39	8.03	7.50	10.04	8.24
0.8	12.03	8.63	10.69	8.42	7.81	11.27	8.81
1.0	12.84	9.04	11.45	8.92	8.22	12.09	9.28
	Source Microphone Distance=1.5 m						
0.2	5.96	5.29	4.99	5.34	5.32	6.04	5.39
0.4	10.12	7.66	8.11	7.32	6.78	9.42	7.35
0.6	12.71	8.45	9.94	8.22	7.60	11.35	8.58
0.8	14.11	9.45	11.33	9.32	8.23	12.97	8.92
1.0	15.13	10.04	12.53	9.87	8.87	14.24	9.42
	Source Microphone Distance=2 m						
0.2	7.22	6.09	5.96	5.94	5.45	7.00	5.91
0.4	11.46	8.18	10.09	7.97	7.33	10.38	7.52
0.6	13.37	9.66	12.18	9.42	8.76	12.51	9.06
0.8	14.47	10.44	13.36	10.11	9.51	13.85	10.23
1.0	15.40	11.26	14.11	10.98	10.22	14.72	11.76

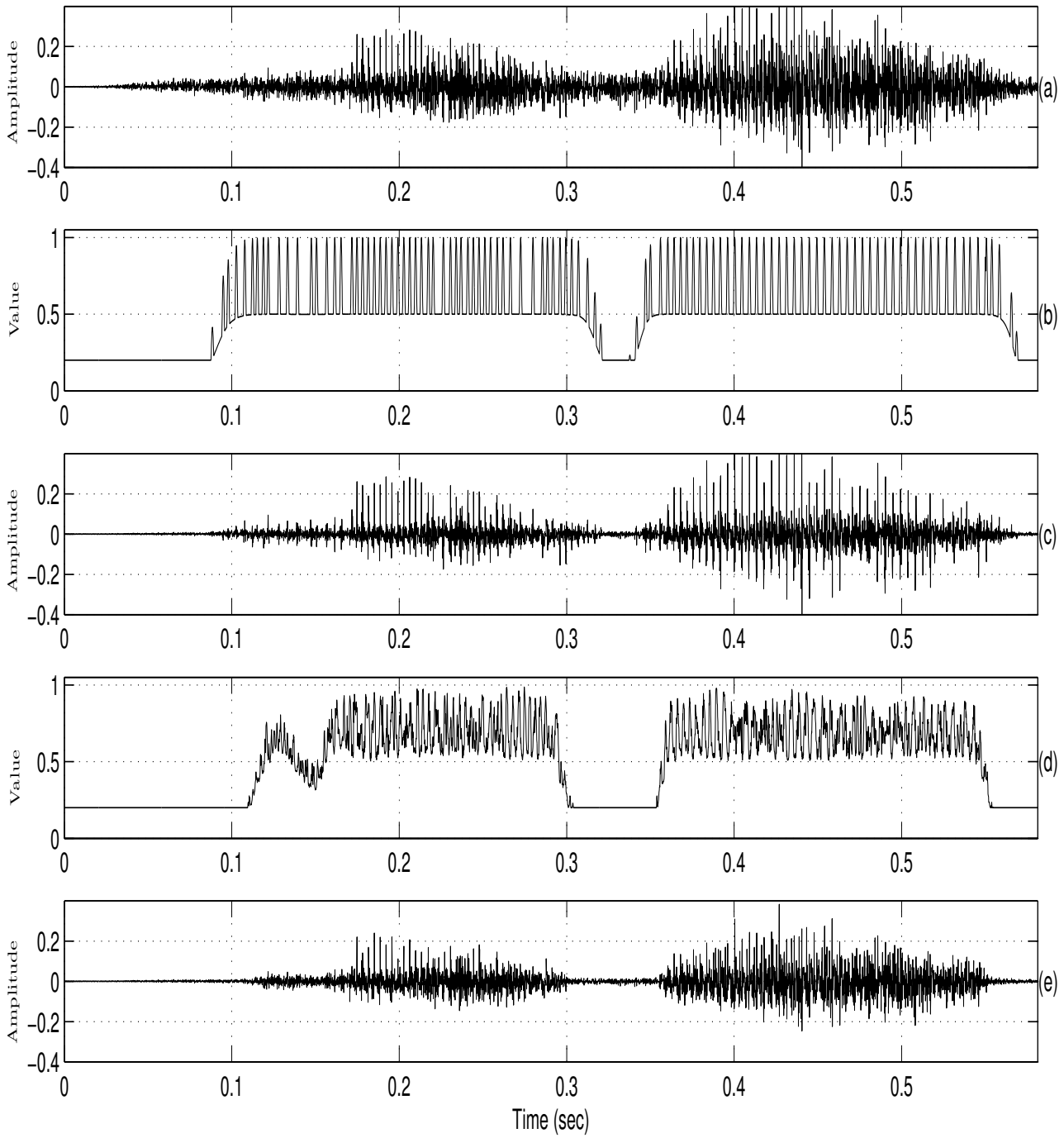


Figure 4.15: Weight function determination by different methods: (a) LP residual of the reverberant speech, (b) weight function obtained by the proposed method, (c) weighted residual by the weight function shown in (b), (d) weight function obtained by the conventional Yegnanarayana and Murthy (YM) LP residual method and (e) weighted residual by the weight function shown in (d).

Table 4.7: Effect of smoothing factor (a) in the Rayleigh smoothing function for a source microphone distance of 2 m.

a	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
T_{60}**(sec)**	Segmental SRR (SegSRR)							
0.2	0.49	0.56	**0.59**	**0.59**	0.58	0.56	0.54	0.54
0.4	-1.43	**-1.39**	**-1.39**	-1.43	-1.47	-1.52	-1.56	-1.59
0.6	-2.23	-2.18	**-2.17**	-2.18	-2.23	-2.28	-2.33	-2.38
0.8	-2.84	-2.75	-2.70	**-2.68**	-2.69	-2.72	-2.76	-2.81
1.0	-3.24	-3.12	-3.05	**-3.00**	-3.02	-3.05	-3.07	-3.19
	Log Spectral Distance (LSD)							
0.2	**5.64**	5.87	6.09	6.33	6.53	6.71	6.88	7.02
0.4	**7.82**	8.00	8.18	8.41	8.64	8.92	9.12	9.29
0.6	9.62	9.56	**9.66**	9.80	9.97	10.15	10.27	10.41
0.8	10.42	10.35	**10.44**	10.46	10.51	10.44	10.52	10.66
1.0	11.46	11.31	11.26	**11.25**	11.28	11.37	11.44	11.5

4.7 Summary

This chapter proposed a combined TSP method for the enhancement of reverberant speech. The proposed approach utilizes the main merits of both methods, removal of late reverberation by spectral subtraction and enhancement of high SRR regions by temporal processing. Spectral subtraction is performed to eliminate the late reverberant components. The signal is further subjected to temporal processing to enhance the high SRR regions. In the temporal processing, the high SRR regions are identified by computing the sum of the 10 largest peaks in the DFT spectrum, the smoothed HE of the LP residual and the modulation spectrum values of spectrally processed speech. In the next level of temporal processing, a fine weight function is derived to enhance the excitation source information around the instants related to direct components. Performance evaluation is carried out with segmental SRR and log spectral distance objective quality measures. The results show that the combined method gives better performance than temporal or spectral processing alone.

5

Combined TSP for Two Speaker Speech Separation

Contents

5.1 Objective of Combined TSP for Two Speaker Speech Separation

This chapter presents a combined temporal and spectral processing method for separating speech of individual speakers from the mixed speech of two speakers. Speech in a two speaker environment is simultaneously collected over two spatially separated microphones. The speech signals are first subjected to temporal processing which involves processing of excitation source information. This results in the identification and enhancement of high signal to noise ratio (SNR) regions of desired speaker. Temporally processed signals are then subjected to spectral processing. Spectral processing involves processing of short time spectra. This results in the identification and enhancement of speech specific features present around the pitch and harmonics of desired speaker. As indicated by different objective and subjective quality measures, the proposed combined temporal and spectral processing method achieves better separation of each speaker compared to either temporal or spectral processing alone.

5.2 Introduction to Two Speaker Speech Separation

In natural environments, speech typically occurs simultaneously with other sounds or interference. This creates the problem of speech separation, i.e. separating target speech from interference. A source of degradation which is more difficult to handle is due to the speech of a competing speaker. This is popularly known as cocktail party effect. This case is difficult for enhancement because the degrading signal too has the characteristics of speech, and hence difficult to distinguish it from the desired signal. Several approaches have been proposed in the literature to enhance the speech degraded by the speech of competing speaker. Most of these methods may be broadly grouped into three categories, namely, blind source separation (BSS) using independent component analysis (ICA), computational auditory scene analysis (CASA) and speech-specific approaches (SSA). Depending on the number of microphones used for collecting speech, these methods may be further classified into single and multi-channel cases. In single channel case, speech is collected over a single microphone, and the objective is to process multi-speaker speech to emphasize desired speaker's speech. This approach is more commonly termed as co-channel speaker separation [24]. In multi-channel case, speech is collected simultaneously over several (two or more) spatially distributed microphones. Signals from all the microphones are processed to enhance speech of one or more speakers. Separation of speech signals can be done effectively, if the speech signals are collected simultaneously over two or more

microphones. This is mainly because multi-channel methods exploit the spatial diversity resulting from the fact that desired and undesired speakers are in practice located at different points in space. The speech signal of each speaker arrives in any pair of microphones at slightly different times (one is a delayed version of other). The estimate of this delay can be used for separating desired speaker's speech.

Psychophysical research shows that periodicity, or pitch, is one of the most effective cues employed by human listeners to separate sounds [306]. Based on this concept initial work in co-channel speaker separation evolved from speech enhancement algorithms designed for separating voiced speech from background noise given a pitch estimate from the target talker [307]. Co-channel speaker separation algorithms have attempted to first estimate the pitch of at least one of the talkers, and then to exploit the pitch and harmonics to separate the two talkers [24, 25]. It is assumed that the pitch contours of the desired and competing speaker speech are sufficiently separated, so that the different pitch harmonics are resolvable. This algorithm is reported to yield good results when both the desired and interfering speech signals are strong. A minimum cross entropy spectral analysis method is proposed for co-channel speaker separation that seems to recover even the weak signals with significantly reduced interference [217]. Sinusoidal modeling of speech is also suggested to obtain the co-channel speaker separation [220].

Recently, BSS by ICA has received lot of attention. Blind separation of instantaneous mixture is achieved by the ICA which aims at decomposing the multivariate data into linear sum of independent components [245]. The goal in BSS is to recover set of independent sources given only a set of sensor observations that are generated from the individual source signals through an unknown linear mixing process. [241]. The BSS relies on two main assumptions: The signal sources must be statistically independent and mixing process must be linear. Many CASA systems have been proposed for speech separation according to the principles of auditory scene analysis introduced by Bregman [223, 232]. In CASA methods, first mixed signal is segmented into time frequency cells using either short-time Fourier transform (STFT) or gammatone filter bank. Then, based on some criteria, namely fundamental frequency, amplitude modulation, onset, offset, position, and continuity, the cells that are believed to belong to one source are grouped and speech is synthesized [234].

A method for processing multi-speaker speech using excitation source information is also proposed by the authors in [27]. The speech of each speaker is enhanced with respect to the other by performing

relative emphasis of speech around the instants of significant excitation of desired speaker. The relative emphasis is achieved by giving larger weight to the linear prediction (LP) residual samples around the instants of significant excitation and lower weight to samples in other regions [27].

In this chapter, we present a multi-channel approach for separation of two speaker speech by combined temporal and spectral processing using speech-specific knowledge. The proposed method is demonstrated using speech collected over two microphones. In the proposed temporal and spectral processing, temporal processing is based on the characteristics of excitation source of speech production and also on the spatial information available in the multi-channel case [27]. The major source of excitation in speech can be approximated as a sequence of impulse-like instants termed as instants of significant excitation [269]. The knowledge of these impulse-like excitations is used for separating speech of individual speakers. Clearly, it is necessary to know the instants of significant excitation of each speaker. However, instants extracted from the degraded speech will correspond to both speakers. Instants of each speaker may be separated by employing multi-channel case. In multi-channel case, there is a time-delay in the arrival of speech of each speaker at a pair of microphones. This delay is different for different speakers as no two speakers can be at the same position and time. That is the relative positions of the instants of significant excitation in the direct component of the speech signal remain unchanged at each of the microphones for a given speaker. These sequences differ only by a fixed delay corresponding to the relative distances of the microphones from the speaker. Therefore time-delays are exploited to separate the instants of significant of excitation of each speaker. Once the instants of desired speaker are identified, then a temporal weight function is derived to emphasize speech components around these instants. The LP residual of the multi-speaker speech is weighted by the weight function. The speech is synthesized from this residual and LPCs of mixed speech. As a result, the desired speaker's speech is perceptually emphasized in the temporal domain. To further improve the separation, the temporally processed speech is subjected to spectral domain processing. The spectral energy of voiced speech is concentrated around the pitch and harmonic frequencies. Therefore spectral processing involves enhancing the regions around the pitch and harmonic peaks of short time spectra computed from the temporally processed speech.

The next issue is the order of processing. In the present work temporal processing precedes spectral processing. This is because determining and tracking pitch of desired speaker is one of the challenging tasks in spectral processing. By performing temporal processing first, speech of one speaker is enhanced

and other speaker is suppressed. Due to this, pitch estimation will be easy in the temporally processed speech. Further, in case of non-overlapping speech regions (i.e., desired and undesired speakers are non-overlapping), the temporal weight function derived deemphasizes the periodicity information of undesired speaker and hence pitch tracking is not required in the proposed spectral processing method.

As will be discussed in the following sections, the novelty of the work presented in this chapter may be summarized as follows: (1) Identifying the potential of combining temporal and spectral processing for speech separation from multi-speaker speech. (2) Even though the principle of temporal processing is same as in [27], the approach for deriving weight function is novel. (3) A simple pitch extraction method for multi-speaker speech. (4) Identifying potential for using pitch information available from temporal processing for spectral processing. (5) Developing and demonstrating that combined temporal and spectral processing indeed provides better performance compared to existing temporal or spectral processing methods alone.

The rest of the chapter is organized as follows: The significance of excitation source information and proposed temporal processing method for two speaker speech separation are discussed in Section 5.3. Section 5.4 discusses the proposed spectral processing method for speech separation. Experimental results and various objective and subjective measures performed on them are given in Section 5.5. Finally, the summary of the work presented in this chapter is mentioned in Section 5.6.

5.3 Speech Separation by Temporal Processing

The main issues in the proposed temporal processing are, (i) identification of instants of significant excitation, (ii) time delay estimation of speakers, and (iii) enhancing speech of individual speakers. First, the instants of significant excitation are determined from the Hilbert envelope (HE) of LP residual as described in Chapter-3.

5.3.1 Time-Delay Estimation

In the present work, for time-delay estimation, it is assumed that speakers are stationary and are not positioned along the perpendicular bisector of the line joining the microphones. Several methods, mainly based on spectral features, have been proposed for time-delay estimation [25, 77, 308–313]. In this study a method based on the excitation source information is used [313]. Accordingly, the time-delay between speech signals at a pair of microphones is estimated by computing the normalized cross-correlation of the HEs of LP residuals. The normalized cross-correlation sequence of the two

HEs $h_1(n)$ and $h_2(n)$ is computed as [288]

$$\rho_{12}(\tau) = \frac{r_{12}(\tau)}{\sqrt{r_{11}(0)r_{22}(0)}}; \quad \tau = 0, \pm1, \pm2, ..., \pm L-1 \tag{5.1}$$

$$= \frac{\sum_{n=i}^{L-|p|-1} h_1(n)h_2(n-\tau)}{\sqrt{\sum_{n=0}^{L-1} h_1^2(n) \sum_{n=0}^{L-1} h_2^2(n)}} \tag{5.2}$$

where $i = \tau$, $p = 0$ for $\tau \geq 0$ and $i = 0$, $p = \tau$ for $\tau < 0$ and L is the length of segments of HE (= 400 for F_s = 8 kHz). In the normalized cross-correlation sequence, the displacement of major peak with respect to the center sample is considered as the time delay in samples. Note that, in the vicinity of the instants of significant excitations, the speech signal exhibits a high SNR relative to the other regions, due to damping of the impulse response of the vocal tract system. While the reflected components and noise may also contribute to some high SNR regions, their relative positions will be different in the signals collected at the two microphones [313]. Hence, the time-delay estimate obtained is robust to noise and reverberation.

The cross-correlation of two microphone signals is shown in Fig. 5.1. We cannot make a decision based on single frame time delay value. This is because there could be spurious peak in the cross-correlation sequence. To overcome this, delay is computed for successive frames of 50 ms duration shifted by 5 ms. The choice of frame size depends on the accuracy of tracking. Smaller frame size will yield better tracking. But larger frame size will yield accurate delay estimation. The plot of time delays, estimated for the speech signals collected over two microphones using frame size of 50 ms and shift of 5 ms is shown in Fig. 5.2(a). Since speech of two speakers is present in the microphone data, two horizontal lines can be observed representing two delays from Fig. 5.2(a). The random values in the plot are mostly due to the non-speech regions. The percentage of frames for each delay value is shown in Fig. 5.2(b). In the histogram shown in Fig. 5.2(b), the locations of peaks correspond to the time delays due to different speakers.

Note that, the term advancement of speaker refers to the speaker which is closest to the microphone. For example, let us assume that speaker 1 and speaker 2 speech reaches the microphone-1 at the time delays of 5 ms and 10 ms and with the time delays of 15 ms and 4 ms to microphone-2 (Refer to Fig. 5.3). In this case, if time delay is estimated with reference to microphone-1 mixture, the time-delays will be d_1 = 10 ms (delay) & d_2 = -6 ms (Advancement). This shows microphone-2 is closest

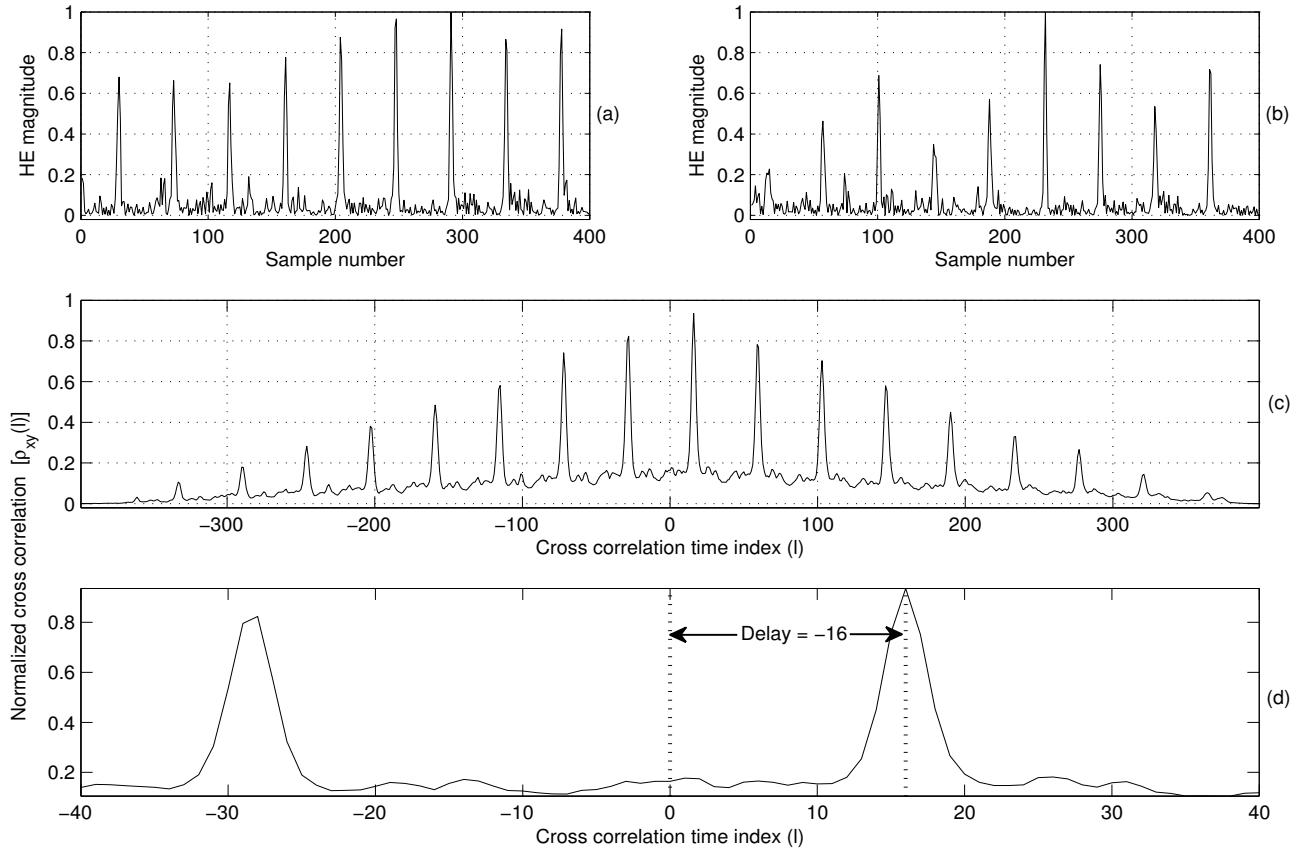


Figure 5.1: Time delay estimation: (a) HE of mic-1 speech signal, (b) HE of mic-2 speech signal, (c) cross-correlation of two microphone signals and (d) only few samples around the center value are shown to indicate the delay between two microphones.

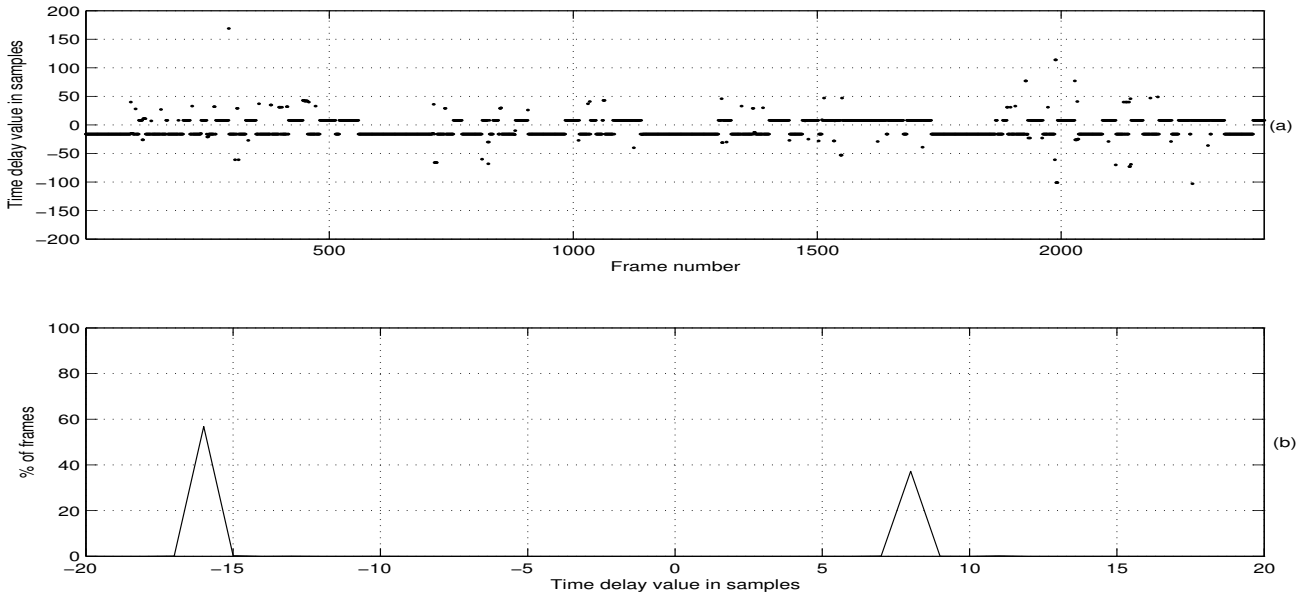


Figure 5.2: Time delay estimation: (a) time-delays estimated for speech signals collected over two microphones, using frame size of 50 ms and frame shift of 5 ms and (b) Histogram of samples showing percentage of frames for each delay value.

to any one of the speaker.

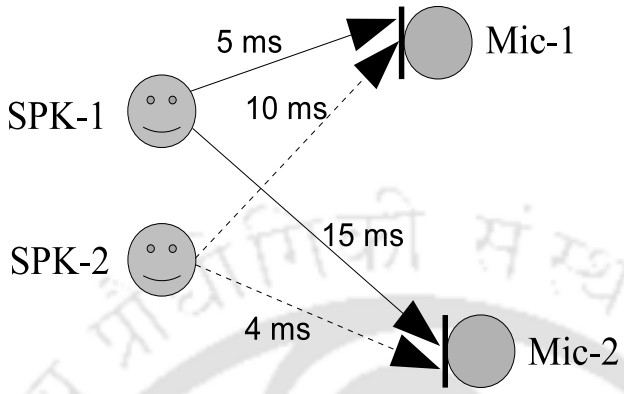


Figure 5.3: Illustration of time delay.

5.3.2 Basis for Temporal Processing

Let $h_1(n)$ and $h_2(n)$ be the normalized HE sequences of speech signals collected at mic-1 and mic-2, respectively. The time-delays d_1 and d_2 between the two microphone signals are obtained from their HEs $h_1(n)$ and $h_2(n)$. Here, positive delay corresponds to advancement of speaker information in $h_1(n)$ compared to $h_2(n)$ and vice versa. To enhance the excitation characteristics of one speaker relative to other speaker, the normalized HEs are combined by shifting $h_2(n)$ by d_1 values, and then taking the minimum of $h_1(n)$ and $h_2(n-d_1)$ for speaker-1. That is,

$$h_{s1}(n) = \min\left(h_1(n), h_2(n-d_1)\right). \tag{5.3}$$

The signal $h_{s1}(n)$ contains emphasized component of $h_1(n)$ corresponding to speaker-1. Due to coherence of speaker-1 in $h_1(n)$ and $h_2(n-d_1)$, $h_{s1}(n)$ retains the peaks around the instants of significant excitation for speaker-1. Due to lack of coherence of speaker-2, $h_{s1}(n)$ suppresses the peaks corresponding to speaker-2.

For illustration, plots of normalized HE of mic-1 and mic-2 signals are shown in Figs. 5.4(a) and (b), respectively. Time aligned HE of mic-2 signal is plotted in Fig. 5.4(c). The sample minima of the HEs computed in Eqn. (5.3) help in reducing the effect of spurious peaks, while preserving genuine

instants of respective speaker as shown in Fig. 5.4(d). Similarly, minimum of $h_1(n)$ and $h_2(n-d_2)$ is taken for speaker-2. That is,

$$h_{s2}(n) = \min\left(h_1(n), h_2(n-d_2)\right). \tag{5.4}$$

Here $h_{s2}(n)$ contains emphasized component of $h_2(n)$ corresponding to speaker-2. Thus the instants of significant excitation of individual speakers are separated.

5.3.3 Speech Separation

The separated instants of significant excitation $h_{s1}(n)$ and $h_{s2}(n)$ can be used to separate the speech of individual speakers. The separation is achieved by deriving a weight function for each speaker, which can be used to suitably modify the LP residual of mixed speech signal. At the first level, separation is achieved by computing the difference between the HEs of $h_{s1}(n)$ and $h_{s2}(n)$. That is,

$$h_{12}(n) = h_{s1}(n) - h_{s2}(n). \tag{5.5}$$

This difference $h_{12}(n)$ shows the instants of desired speaker as positive peaks and the instants of undesired speaker as negative peaks. To enhance desired speaker, the regions around the positive peaks of $h_{12}(n)$ needs to be emphasized and regions around large negative peaks need to be de-emphasized. For instance, separated HEs of individual speakers $h_{s1}(n)$, $h_{s2}(n)$ and the difference of these two values are plotted in Figs. 5.4(d), (e) and (f), respectively. Similarly, difference between the HEs of $h_{s2}(n)$ and $h_{s1}(n)$ emphasizes the instants of excitation of speaker-2 relative to the instants of speaker-1.

5.3.3.1 Gross Weight Function

There will be changes in the excitation characteristics both at the fine and gross levels during speech production. The fine level changes may be from closed phase to open phase in a pitch period and the gross level changes may be from silence to voiced excitation. The weight function for the LP residual to enhance the desired speaker speech is derived at two different levels, namely, gross and fine levels. The gross level weight function is derived to identify the speech and non-speech regions of degraded speech signal. It is obtained by smoothing the absolute value of error function by 50 ms Hamming window and nonlinearly mapping the smoothed sequence by sigmoidal nonlinear function.

The smoothed sequence is obtained as

$$h_{sm}(n) = |h_{12}(n)| * h_{w1}(n) \tag{5.6}$$

where $*$ denotes convolution operation and $h_{w1}(n)$ is Hamming window of 50 ms duration. The smoothed sequence is further subjected to nonlinear mapping operation to emphasize high values towards unity and suppress remaining values to certain minimum value. The gross weight function is obtained as

$$w_g(n) = (1 - w_{gm}) \frac{1}{1 + e^{-\lambda(h_{sm}(n) - T)}} + w_{gm} \tag{5.7}$$

where λ is the slope parameter, T is threshold and w_{gm} is minimum value of the gross weight function. For illustration, Fig. 5.5(a) shows a degraded speech collected from mic-1. The difference values $h_{12}(n)$ and corresponding smoothed sequence $h_{sm}(n)$ are plotted in Figs. 5.5(b) and (c), respectively. The gross weight function derived using the mapping function is shown in Fig. 5.5(d). The parameters chosen to compute $w_g(n)$ are $\lambda = 20$, $T = 0.2$ times the average value of $h_{sm}(n)$ and $w_{gm} = 0.05$. An experimental evaluation on different values of λ and T is given in section 5.5.

5.3.3.2 Fine Weight Function

The fine weight function for the residual is derived by identifying the instants of desired and undesired speakers. To detect the instants of significant excitation of both speakers, first smaller fluctuations in the difference values $h_{12}(n)$ are minimized by smoothing $h_{12}(n)$ with Hamming window of 3 ms duration. The smoothed signal is obtained as

$$h_s(n) = h_{12}(n) * h_{w2}(n) \tag{5.8}$$

where $h_{w2}(n)$ is the Hamming window of 3 ms duration. Fig. 5.6(b) shows smoothed sequence for the difference values plotted in Fig. 5.6(a). As mentioned, major peaks in the smoothed difference values $h_s(n)$ indicate approximate locations of instants of significant excitation. In particular, positive peaks correspond to the instants of desired speaker and negative peaks correspond to the instants of undesired speaker. The instants of significant excitation for desired speaker are detected by convolving the positive values with the first order Gaussian differentiator (FOGD). Similarly, instants of the undesired speaker are detected by convolving absolute of negative values with FOGD.

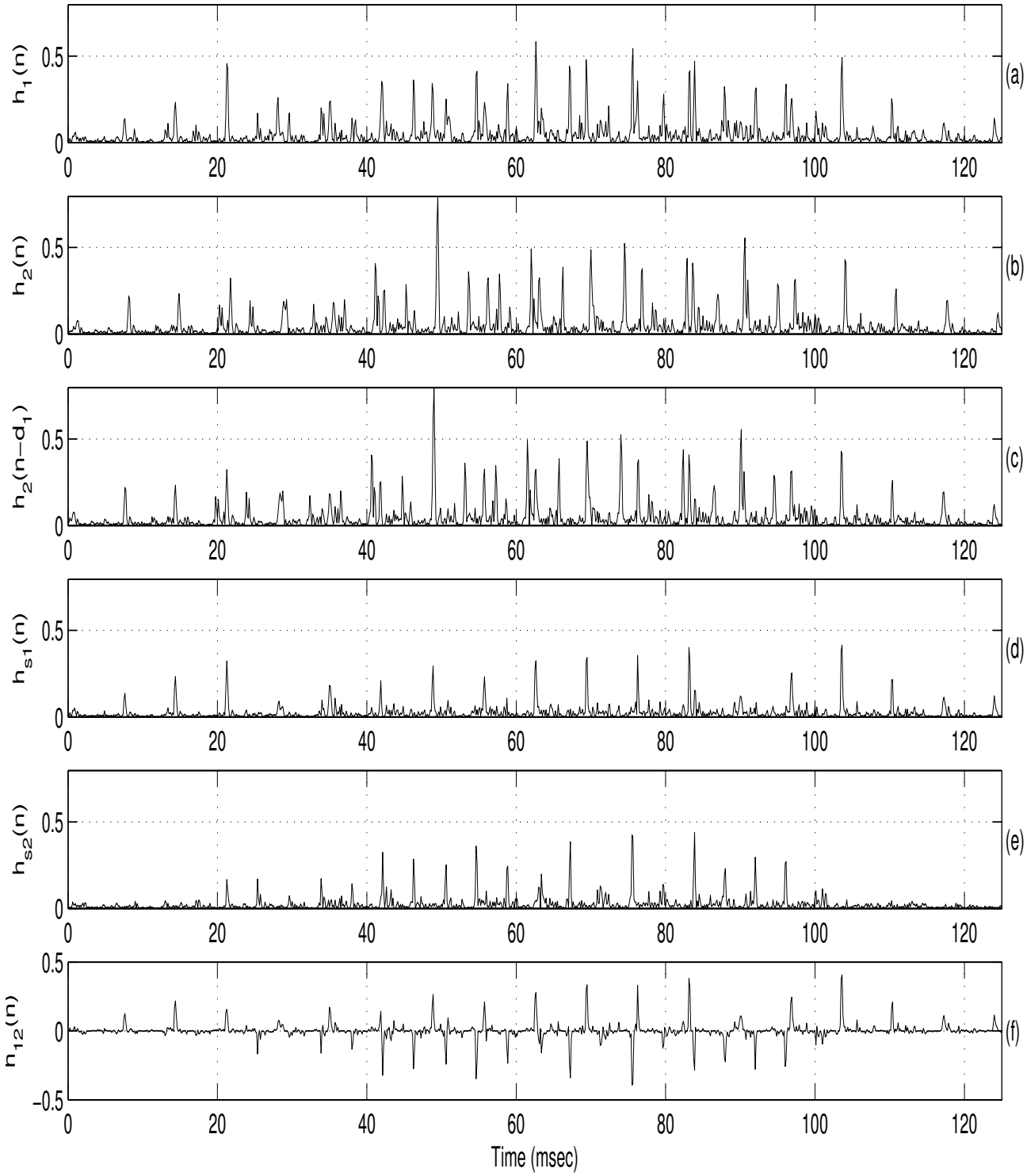


Figure 5.4: Basis for temporal processing: (a) HE of mic-1 signal, (b) HE of mic-2 signal, (c) time aligned HE of mic-2 signal, (d) separated HE of speaker-1, (e) separated HE of speaker-2 and (f) difference values of separated HEs of speaker-1 and speaker-2.

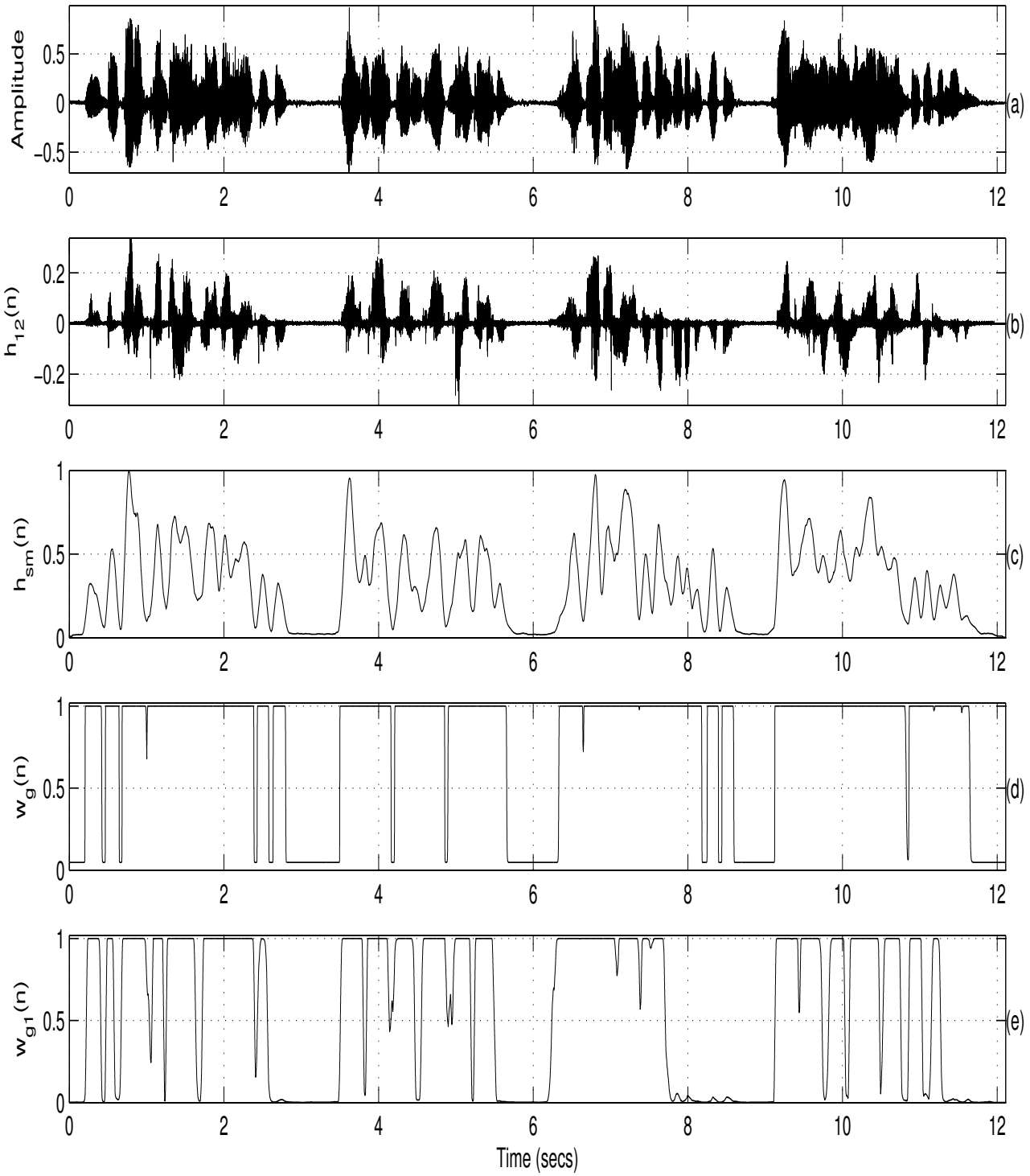


Figure 5.5: Gross weight function determination: (a) degraded speech signal collected from mic-1, (b) difference values of separated HEs of speaker-1 and speaker-2, (c) smoothed difference values, (d) gross weight function and (e) gross weight function for desired speaker.

The fine weight function for the LP residual signal to emphasize desired speaker and deemphasize undesired speaker is derived by convolving the detected instants with Hamming window of 3 ms duration. As mentioned in earlier chapters, this is because due to impulse-like excitation and damped sinusoid like impulse response of the vocal tract system, a short (1-3) ms segment in the voiced speech signal around the instants of significant excitation corresponds to high SNR portion of speech.

Let a_i and b_i be the approximate locations of instants of significant excitation of desired and undesired speaker, respectively. Then the fine weight function $w_f(n)$ is derived according to the following relations

$$w_f(n) = [w_{\min} + (1 + w_{\min})w_a(n)] - w_{\min}w_b(n) \tag{5.9}$$

where $w_{\min} = 0.3$ and

$$w_a(n) = I_a(n) * h_a(n) \quad \& \quad w_b(n) = I_b(n) * h_b(n) \tag{5.10}$$

where $*$ denotes the convolution operation and

$$I_a(n) = \sum_{i=1}^{N_a} \delta\left(n - a_i\right) \quad \& \quad I_b(n) = \sum_{i=1}^{N_b} \delta\left(n - b_i\right) \tag{5.11}$$

$$h_a(n) = h_b(n) = 0.54 + 0.46\ \cos\left(\frac{2\pi n}{N_w}\right); -\frac{N_w}{2} \leq n \leq \frac{N_w}{2} \tag{5.12}$$

where N_a and N_b respectively, represent total number of detected instants of desired and undesired speaker and $N_w = (3 \times F_s)/1000$. For instance, the detected instants of desired and undesired speaker ($I_a(n)$ and $I_b(n)$) are plotted in Fig. 5.6(c). However, it may happen that undesired speaker instant location may occur with in 3 ms duration of desired speaker instant location as indicated by down arrow symbol in Fig. 5.6(c). In such situation importance is given for emphasizing the region around the desired speaker instant rather than deemphasizing the undesired speaker. The resulting fine weight function is given in Fig. 5.6(d).

The choice of $w_{\min}$ depends on the level of distortion tolerable and amount of separation to be achieved. Higher weighting of LP residual introduces distortion but better separation in the processed signal. That is, lower value of $w_{\min}$ introduces perceptual distortion but results better separation. The value of $w_{\min}$ is chosen as 0.3 which produces negligible distortion. This may leave out some amount of undesired speaker in the temporally processed signal. This can be further reduced in the subsequent spectral processing.

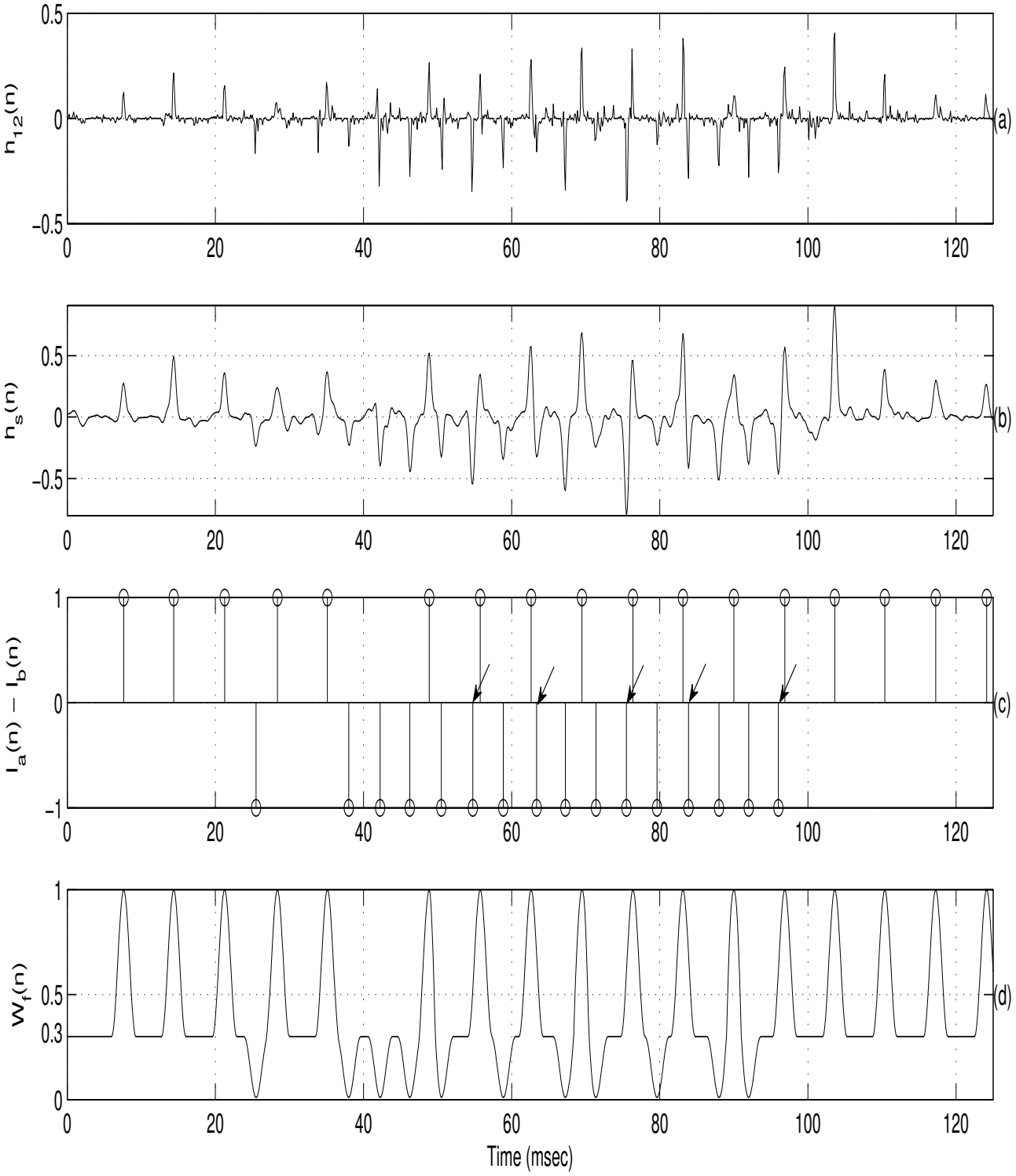


Figure 5.6: Fine weight function determination: (a) difference values of separated HEs of speaker-1 and speaker-2, (b) smoothed difference values, (c) instants of desired and undesired speaker and (d) fine weight function.

5.3.3.3 Combined Weight Function

The combined weight function for temporal processing is derived by multiplying gross weight function with fine weight function.

5.3.3.4 Speech Separation

The LP residual of degraded speech is multiplied with the combined weight function to generate the enhanced residual. The enhanced residual is used to excite the time-varying all-pole filter derived from the degraded speech to obtain the temporally processed speech. The different steps involved in the temporal processing are illustrated in Fig. 5.7. For clarity, only small portion of about 2 secs has been chosen for illustration. Fig. 5.7(a) shows the LP residual of multi-speaker speech collected from mic-1. The respective gross and the combined weight functions are given in Figs. 5.7(b) and (c). Fig. 5.7(d) shows the enhanced residual obtained by weighting.

5.4 Speech Separation by Spectral Processing

Temporal processing method enhances the speech-specific features (i.e., excitation features) of the desired speaker at the temporal level. However, degradation at the spectral level still persists due to the use of all-pole filters derived from the degraded speech. To further improve the characteristics at spectral level and to provide better separation of desired speaker, the spectral processing is performed on the temporally processed speech. Spectral processing mainly depends on accurate estimation of pitch frequency of individual speakers. From the estimated pitch, spectral level enhancement is obtained by sampling and enhancing pitch and harmonics of temporally processed short-time speech spectra.

At the first level, the gross regions of the desired speaker are identified from the positive difference values obtained from Eqn. (5.5). As mentioned, the positive difference values represent the instants of desired speaker and hence these values are smoothed using 50 ms Hamming window and non-linearly mapped as described in Section 5.3.3.1 to identify the gross regions of desired speaker. Fig. 5.5(e) shows non-linearly mapped values ($w_{g1}(n)$) of smoothed positive differences. This approach will work only for non-overlapping speech regions. For overlapping speech regions, the fine weight function derived from the instants of significant excitation will deemphasize undesired speaker. This will be demonstrated in the following description.

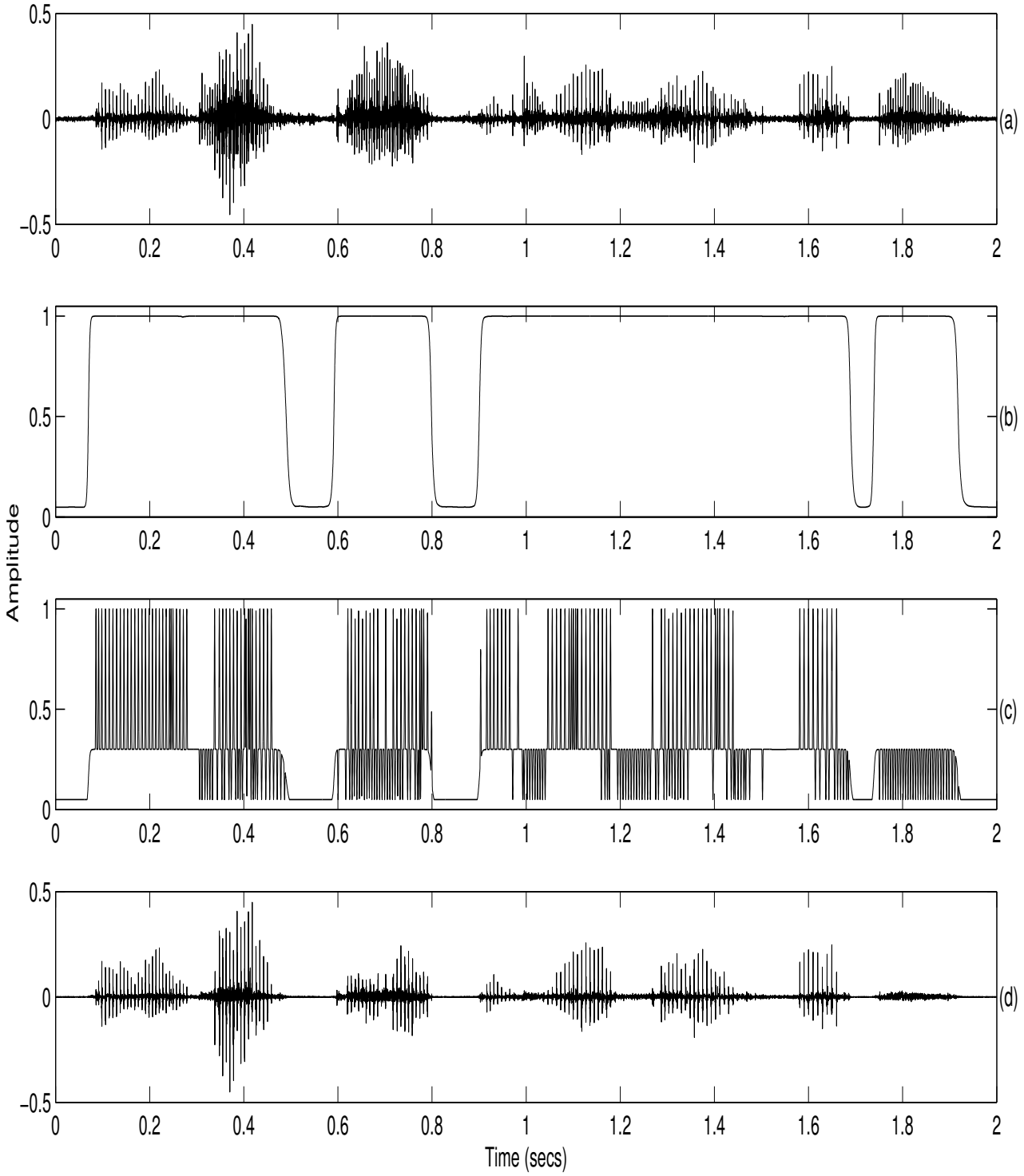


Figure 5.7: LP residual enhancement: (a) LP residual of multi-speaker speech collected from mic-1, (b) gross weight function, (c) combined weight function and (d) enhanced LP residual signal.

5.4.1 Pitch Estimation

When two speakers simultaneously speak, the absolute spectrum of the resulting speech shows two trains of equally spaced harmonics imposed one on the other. The main task here is to separate required harmonic train by finding pitch of the desired speaker. By using the separated harmonics, the desired spectrum can be reconstructed. Estimation of pitch plays an important role in the spectral separation process. Various pitch estimation and tracking algorithms based on different approaches have been developed for the estimation of pitch frequency. In this work, the pitch estimate of desired speaker is obtained from the normalized autocorrelation of mean subtracted HE of LP residual of temporally processed speech [211].

Let $sp1(n)$ be the enhanced speaker signal by temporal processing method and $h(n)$ be the HE of LP residual of $sp1(n)$. For each block of 40 ms with shift of 10 ms, the normalized autocorrelation is obtained as [288]

$$R(\tau) = \frac{\sum_{n=0}^{L-1-l} h_m(n) h_m(n+\tau)}{\sum_{n=0}^{L-1} h_m^2(n)}; \;\; \tau = 0, 1, 2, ..., L-1 \tag{5.13}$$

where $L = 320$ for $F_s = 8$ kHz and

$$h_m(n) = h(n) - E\{h(n)\} \tag{5.14}$$

where $E\{.\}$ denotes the expected value operator. The first major peak with reference to zero time lag is considered as pitch period of speaker. The autocorrelation methods need at least two pitch periods to detect pitch and hence frame size of 40 ms is chosen. The voicing decision is also made by computing the following features from the normalized autocorrelation sequence. The features are

(i) The magnitude of first major peak (R_p) [203]

(ii) The similarity behavior of samples around the first major peak of the normalized autocorrelation sequence $R(l)$ [211]

The similarity is measured by comparing samples in a region of 2 ms on either side of the first major peak of present frame with samples from previous and next frame. The similarity measure is computed as

$$C_s = \max\left\{\frac{COV(R_i, R_{i-1})}{\sigma_{R_i}\sigma_{R_{i-1}}}, \frac{COV(R_i, R_{i+1})}{\sigma_{R_i}\sigma_{R_{i+1}}}\right\} \tag{5.15}$$

where $COV(X,Y) = E(XY) - E(X)E(Y)$ and $\sigma_X = \sqrt{E(X) - E^2(X)}$. R_i, R_{i-1} and R_{i+1} represent samples around the first major peak in the current, previous and next frame, respectively. The frame of speech subjected to autocorrelation is considered as voiced frame only when the values of $R_p \geq 0.4$ and $C_s \geq 0.7$ [203, 211].

For instance, the nature of normalized autocorrelation sequence (along with values of normalized peak strength R_p and similarity measure C_s) for three different cases are shown in Fig. 5.8. The three different cases include:

(i) Frame of degraded speech signal consists of both desired and competing speaker (Figs. 5.8(a)-(d))

(ii) Frame of degraded speech signal consists of only desired speaker (Figs. 5.8(e)-(h))

(iii) Frame of degraded speech signal consists of only competing speaker (Figs. 5.8(i)-(l)) and Figs. 5.8(m)-(p)).

It can be observed from Fig. 5.8 that for the first case the weight function derived from temporal processing method enhances the instants of desired speaker and deemphasizes other speaker's instants. Therefore one speaker is present at high level and other speaker at reduced level and hence the pitch frequency obtained will be of enhanced speaker. For the last case, from Fig. 5.8(l) it can be observed that, the weight function deemphasize instants of undesired speaker and the resulting autocorrelation sequence does not have any periodicity information. This may not be true in all cases. It can be noted from Fig. 5.8(p), even though the energy level of excitation source signal is reduced significantly, the periodicity information of undesired speaker still remains in the deemphasized signal. This aspect will also be demonstrated in Section 5.5 with some quantitative measure.

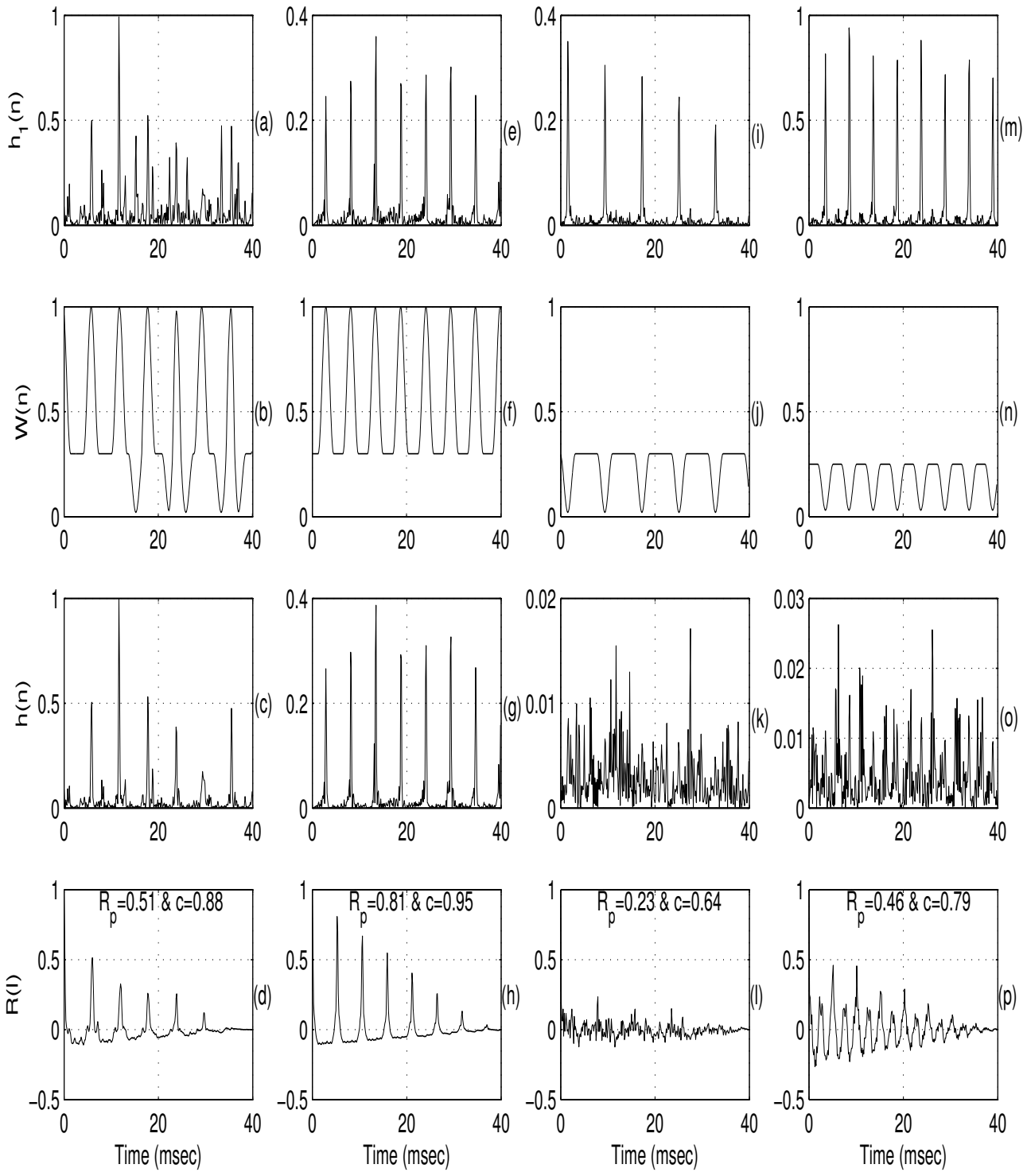


Figure 5.8: Pitch determination: (a), (e), (i) & (m) HE of LP residual of degraded speech, (b), (f), (j) & (n) final weight function, (c), (g), (k) & (o) HE of LP residual of temporally processed speech signal and (d), (h), (l) & (p) normalized autocorrelation ($R(\tau)$) of mean subtracted HE of LP residual.

5.4.2 Spectral Processing

The spectral processing starts from segmenting speech signal into frames of 40 ms with shift of 10 ms. Each frame is weighted by Hamming window. We first perform 1024 point DFT of windowed speech termed $X(k)$. Then frequency indices of pitch and harmonics are determined from the estimated pitch. Let the obtained ideal pitch and harmonic indices of voiced frame be l_i. Then we select the index of pitch and harmonics termed p_i, by examining $X(k)$ in the range of $l_i - 2 \leq p_i \leq l_i + 2$ according to the following criterion.

$$|X(p_i)| \geq |X(p_i + 1)| \text{ and } |X(p_i)| \geq |X(p_i - 1)|\,;\ 1 \leq i \leq N_p \tag{5.16}$$

where N_p represents total number of harmonics in the particular frame. The value of N_p can be obtained from the pitch frequency estimate. The above specified criterion is employed to pick peaks in the spectrum nearest to harmonics. In case if above criterion is not satisfied in the range of $l_i - 2 \leq p_i \leq l_i + 2$, then p_i is taken as

$$p_i = \arg\max_{k_i} \{X(k_i)\}\,;\ \ l_i - 2 \leq k_i \leq l_i + 2. \tag{5.17}$$

Once the frequency index of pitch and harmonics are found, then window function for sampling magnitudes of pitch and harmonics is obtained as

$$W(k) = P(k) * h_r(k) \tag{5.18}$$

where $*$ denotes convolution operation and

$$P(k) = \sum_{i=1}^{N_p} \delta\left(k - p_i\right) \tag{5.19}$$

$$h_r(k) = \begin{cases} 1, & -2 \leq k \leq 2 \\ 0, & \text{otherwise} \end{cases} \tag{5.20}$$

Finally, the enhanced speech spectrum is obtained as follows

- **Case 1:** $W_{g1} \geq 0.2$, $R_p \geq 0.4$ and $c \geq 0.7$ (Desired speaker and voiced region)

$$X_s(k) = A_f \times X(k) \times W(k). \tag{5.21}$$

In this equation the multiplication factor A_f is used to further enhance pitch and harmonics of desired speaker with reference to undesired speaker (Harmonic sampling with spectral enhance-

ment of desired speaker). In the present study A_f is chosen as 2.

- **Case 2:** $W_{g1} < 0.2$, $R_p \geq 0.4$ and $C_s \geq 0.7$ (Undesired speaker and voiced region)

$$X_s(k) = X(k) \times (1 - W(k)). \tag{5.22}$$

- **Case 3:** $R_p < 0.4$ or $C_s < 0.7$: (Non-speech or unvoiced region)

$$X_s(k) = X(k). \tag{5.23}$$

There is no special attempt made for processing unvoiced regions. However in the temporal processing step, the burst and frication type of excitation representing unvoiced regions are identified and enhanced with reference to undesired speaker. This is illustrated in Fig. 5.9, which shows the computed weight function for speech mixture consisting of unvoiced (desired speaker) and voiced region (undesired speaker).

Finally, enhanced speech spectrum is obtained as

$$\hat{X}(k) = \begin{cases} X_s(k), & X_s(k) > \beta X(k) \\ \beta X(k), & \text{otherwise} \end{cases} \tag{5.24}$$

where β is the spectral floor factor and is chosen as 0.02 [33].

The temporal and spectral processed signal is synthesized using IDFT and overlap-add technique. In order to obtain the two speakers speech from single microphone, we have used respective pitch values of speaker-1 and speaker-2. By using obtained pitch frequencies, the two speaker signals are enhanced by the spectral processing method. The various steps involved in the proposed combined temporal and spectral processing for two speaker separation are illustrated in the block diagram shown in Fig. 5.10. In the block diagram, the processing delay block is mainly included to account for the side chain involving the computation of product of the gross and fine weight functions.

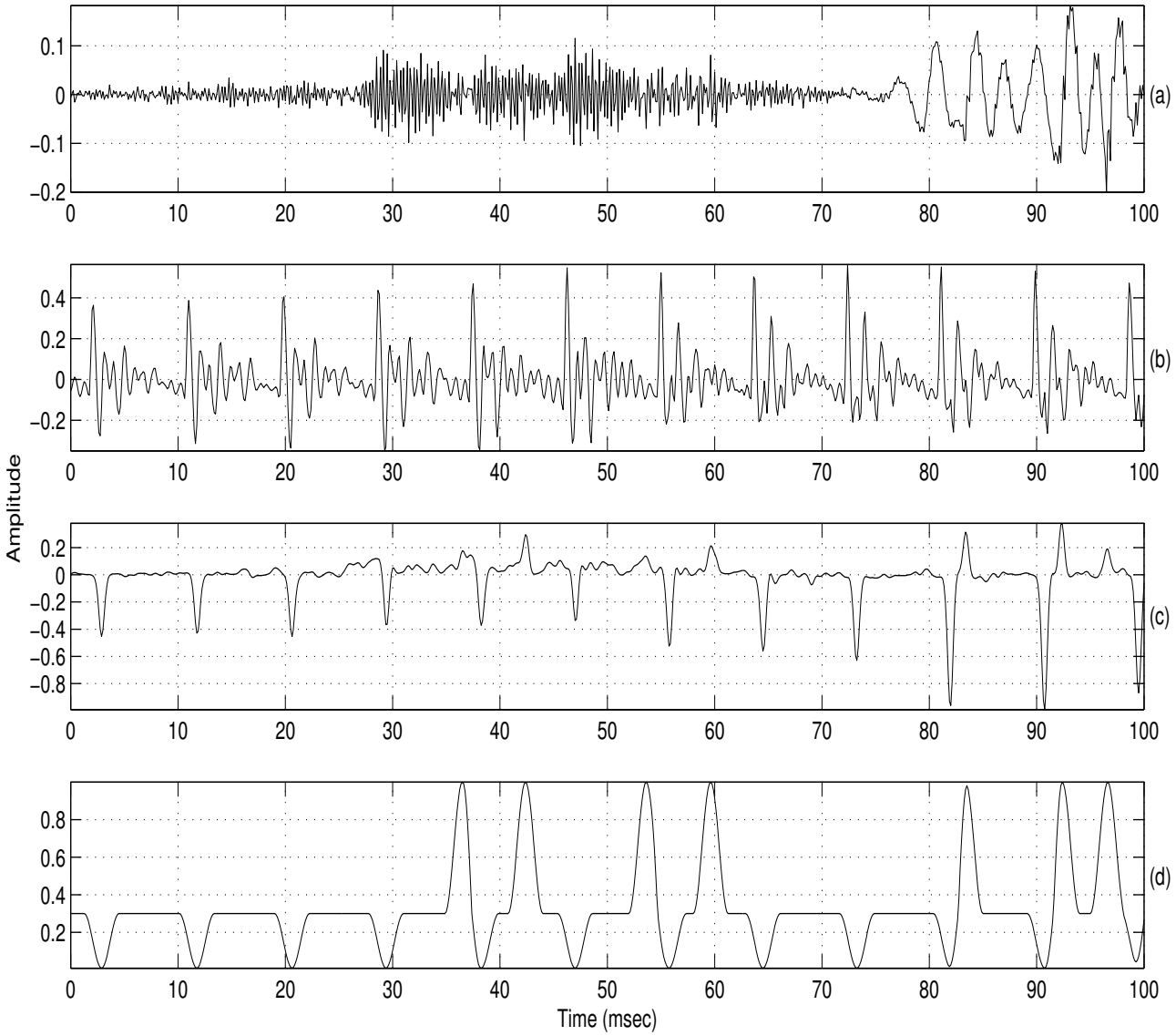


Figure 5.9: LP residual enhancement for unvoiced regions: (a) Clean speech of speaker-1 (unvoiced), (b) Clean speech of speaker-2 (voiced), (c) smoothed difference values ($h_s(n)$) and (d) combined weight function.

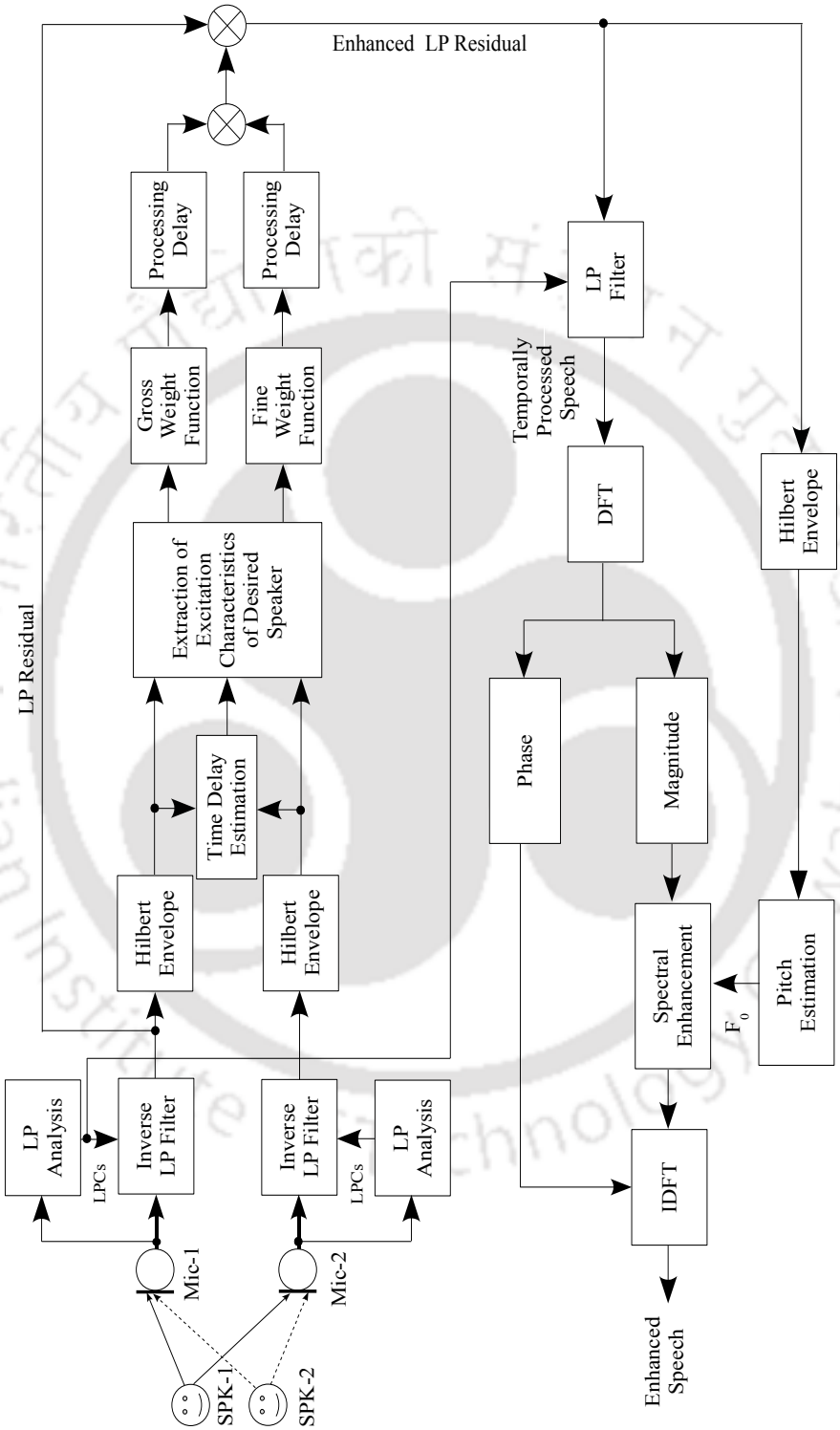


Figure 5.10: Block diagram of proposed combined temporal and spectral processing method for two speaker separation.

5.5 Experimental Results and Performance Evaluation

Different experimental studies are conducted to evaluate the performance of individual and combined processing methods. For comparison purpose independent spectral processing method is also developed. The spectral processed speech is obtained by sampling pitch and harmonics of degraded speech. It should be noted that the pitch frequency for this spectral processing is estimated from the temporally processed speech. The estimation of pitch will therefore be robust compared to other pitch extraction methods designed for multi-speaker environment.

The speech examples for the experimental studies include (i) 10 speech mixtures synthesized from five male and five female speech examples of TIMIT database, (ii) 10 speech mixtures synthesized from five male and five female speech examples of IIT Guwahati (IITG) speech database and (iii) two speech mixtures collected from the real laboratory environment. All speech mixtures are sampled at 8 kHz and stored with 16 bit/sample resolution. TIMIT data taken for study are resampled to 8 kHz. In IITG database, speech signals are collected in clean laboratory environment have Indian English accent and the average duration of each signal is about 10 secs. The speech mixtures consist all possible combinations of gender. All speech examples have approximately same loudness. All the objective evaluation results given in this section are average of measurements of 10 speech mixtures taken from the TIMIT and IITG database.

For generating synthetic two-microphone mixtures the following relations are used. Suppose assume that speaker-1 speech (s_1) and speaker-2 speech (s_2) are delayed/advanced by d_1 and d_2 samples with reference to any one of the microphones. Then the two microphone signals are generated by

$$S_{m1} = s_1 + s_2 \tag{5.25}$$

$$S_{m2} = s_1(n - d_1) + s_2(n - d_2) \tag{5.26}$$

The target-to-masker ratio (TMR) of the speech mixture is maintained in the range of -2 dB to +2 dB with an average of -0.73 dB for TIMIT database signals and -0.04 dB for IITG database signals. The TMR values of various speech mixtures considered in this work are given in Table 5.1.

Table 5.1: Target-to-masker ratio values (TMR in dB) of an individual speech mixtures

Signal	Speaker-1	Speaker-2	TMR	Signal	Speaker-1	Speaker-2	TMR
TIMIT1	Male	Female	-0.85	IITG1	Male	Female	0.32
TIMIT2	Male	Female	1.44	IITG2	Male	Female	0.49
TIMIT3	Male	Female	-1.23	IITG3	Male	Female	-0.79
TIMIT4	Male	Female	0.25	IITG4	Male	Female	-0.89
TIMIT5	Male	Female	-0.49	IITG5	Male	Female	1.32
TIMIT6	Male	Male	-1.65	IITG6	Male	Male	-0.59
TIMIT7	Male	Male	-0.84	IITG7	Male	Male	0.34
TIMIT8	Female	Female	-0.95	IITG8	Male	Male	-0.46
TIMIT9	Female	Female	-1.15	IITG9	Female	Female	-0.25
TIMIT10	Female	Female	-1.86	IITG10	Female	Female	0.08

5.5.1 Time Domain Waveforms and Spectrograms

The data for study in this section is obtained by adding speech of male and female speakers taken from TIMIT database. Figs. 5.11(a) and (b) show clean speech signals of two speakers. These two speech signals are mixed with delays of $d_1 = 8$ and $d_2 = -16$ samples to get mixed speech signals for a pair of microphones. The corresponding degraded signals are plotted in Figs. 5.11(c) and (d). The enhanced speech of speaker-1 obtained by temporal, spectral and combined temporal and spectral processing are given in Figs. 5.11(e)-(f), respectively. Similarly, Figs. 5.11(h)-(j) show the enhanced speech signals of speaker-2. The separation of speech of individual speakers can be observed from the figure. In particular, the general temporal characteristics of desired speaker speech are better preserved in the combined method compared to individual processing methods. The spectrograms of respective speech signals shown in Fig. 5.11 are given in Fig. 5.12. All the spectrograms are constructed using Hamming window of 128 samples with shift of 64 samples. The spectrogram of combined method is visually more similar to that of original spectrogram than the spectrograms of individual processing methods.

To show the effectiveness of combined method in more clear way, Figs. 5.13(a)-(e) show the HE of LP residual of clean (speaker-1), degraded, temporal, spectral and combined temporal and spectral processed speech, respectively. Similarly, Figs. 5.13(f)-(j) show the short time magnitude spectrum of

clean, degraded, temporal, spectral and combined temporal and spectral processed speech, respectively. The combined processing shows improvement in both the excitation source signal and short time spectrum, whereas individual processing methods show major improvement either at the excitation source signal or at the short time spectrum only.

Note that for the case of two speakers with almost similar pitch characteristics, if time-delays between the two speakers are greater than pitch period then the proposed temporal processing method will enhance the excitation characteristics of both the speakers. On the other hand, if the time delay is less than the pitch period, the separation will be poor.

5.5.2 Instants Detection Accuracy

The performance of two speaker speech separation methods may be evaluated by computing the deviation in approximate instants location of clean speech of desired speaker with respect to enhanced signal. First, the instants of significant excitation corresponding to clean speech of two speakers are extracted. Then, the percentage of accuracy in determining the instants of significant excitation is found. The results of this analysis are given in Table 5.2. The entries in Table 5.2 show the percentage of approximate instants and their deviation with respect to the instants of clean speech of desired speaker for individual and combined processing methods. Table also shows the total number of instants derived from clean and enhanced speech signals. All the enhanced speech signals are properly time aligned with reference to clean speech of desired speaker before computing deviation. Temporal processing gives significantly higher performance than spectral processing. From the Figs. 5.13(a)-(e) it can be observed that temporal processing results significant improvement in the excitation source signal than spectral processing and hence it provides improved performance than spectral processing. In particular from Fig. 5.13(d) in the region from 0.08 to 0.1 sec (indicated by down arrow symbol), nature of HE seems to be more affected in the spectrally processed speech.

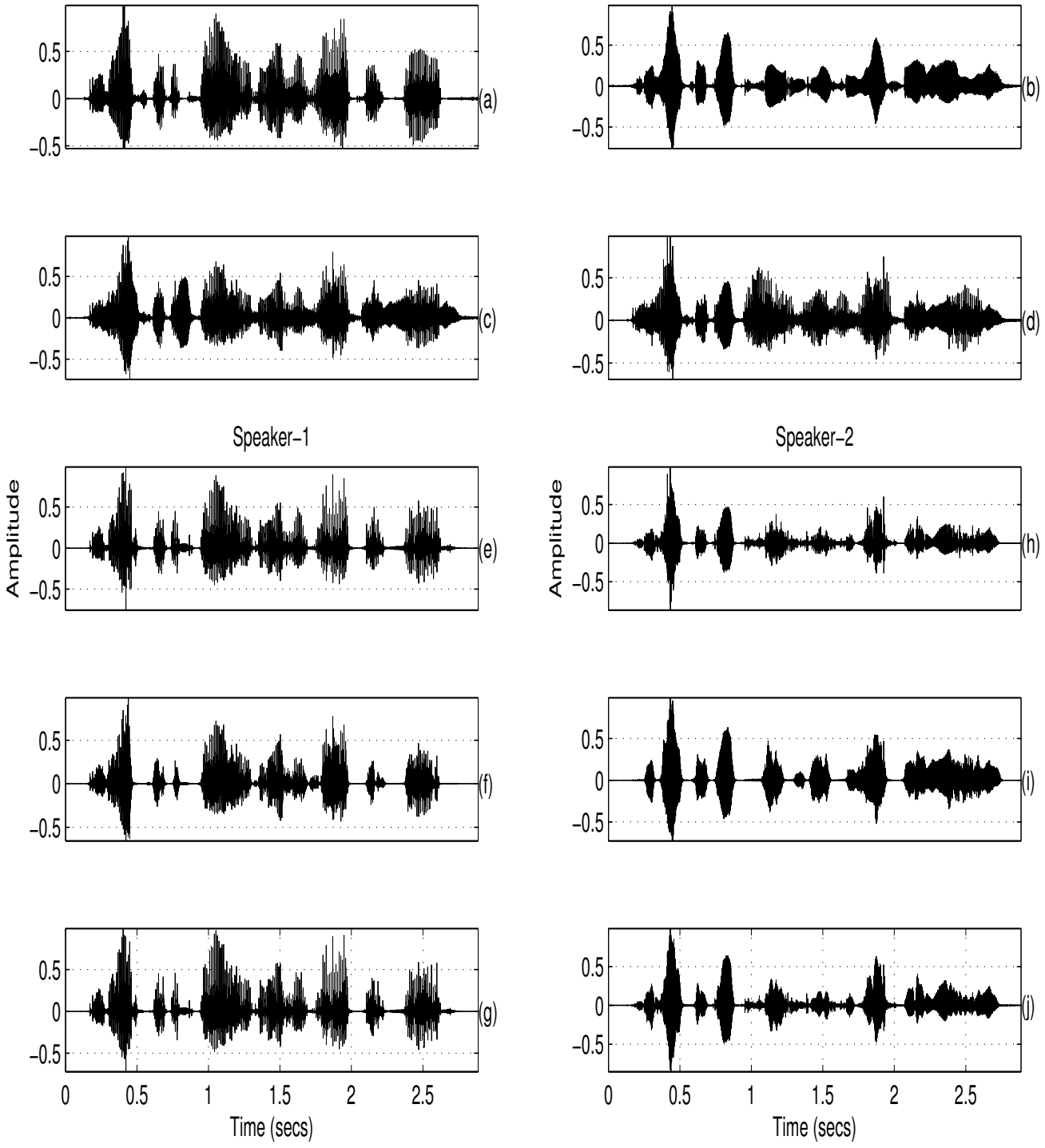


Figure 5.11: Time domain representation of (a) clean speech of speaker-1, (b) clean speech of speaker-2, (c) degraded speech collected from mic-1, (d) degraded speech collected from mic-2, (e) enhanced speaker-1 obtained by temporal processing, (f) enhanced speaker-1 obtained by spectral processing, (g) enhanced speaker-1 obtained by the combined method, (h) enhanced speaker-2 obtained by temporal processing, (i) enhanced speaker-2 obtained by spectral processing and (j) enhanced speaker-2 obtained by the combined method.

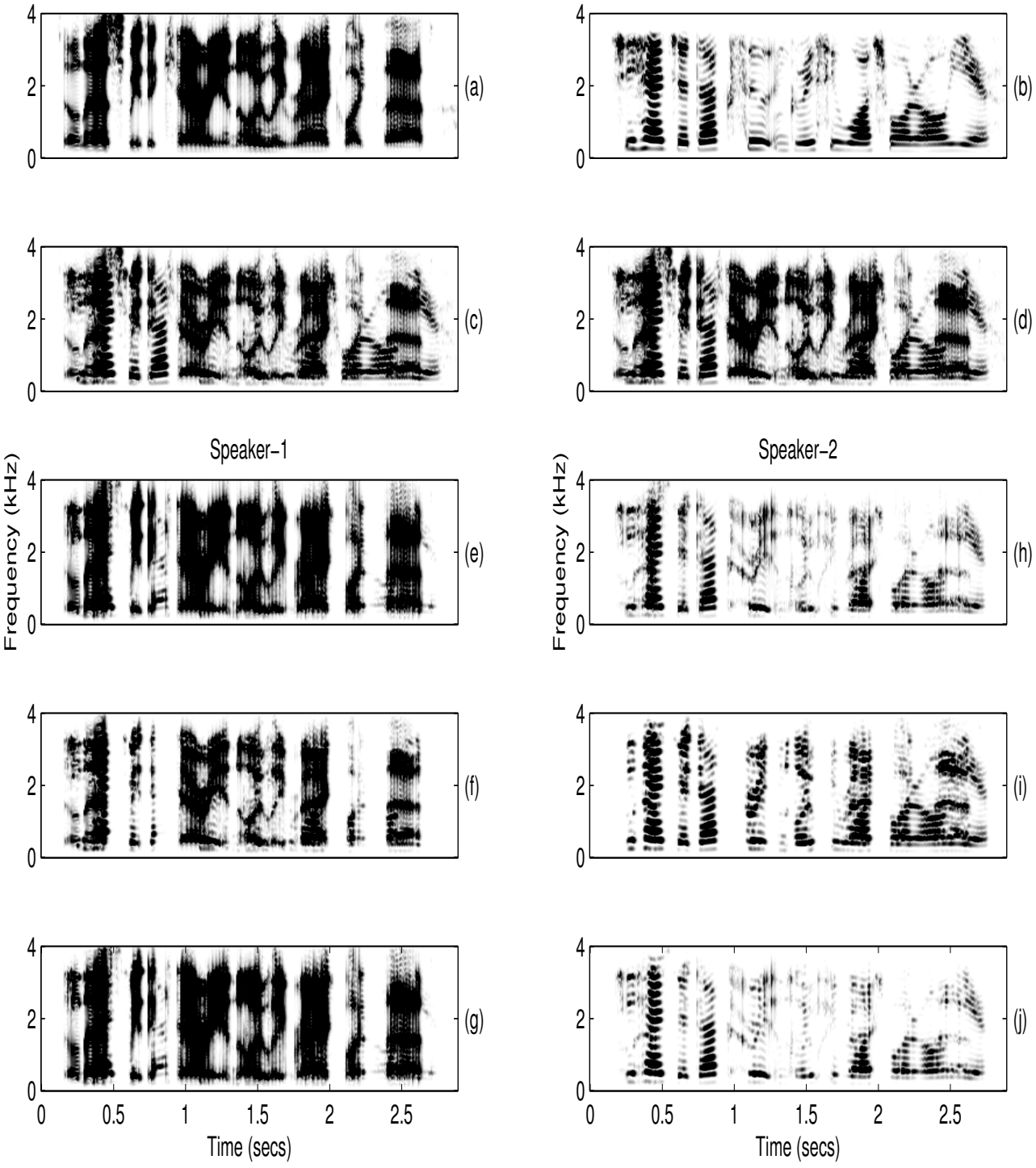


Figure 5.12: Spectrogram representation of (a) clean speech of speaker-1, (b) clean speech of speaker-2, (c) degraded speech collected from mic-1, (d) degraded speech collected from mic-2, (e) enhanced speaker-1 obtained by temporal processing, (f) enhanced speaker-1 obtained by spectral processing, (g) enhanced speaker-1 obtained by the combined method, (h) enhanced speaker-2 obtained by temporal processing, (i) enhanced speaker-2 obtained by spectral processing and (j) enhanced speaker-2 obtained by the combined method.

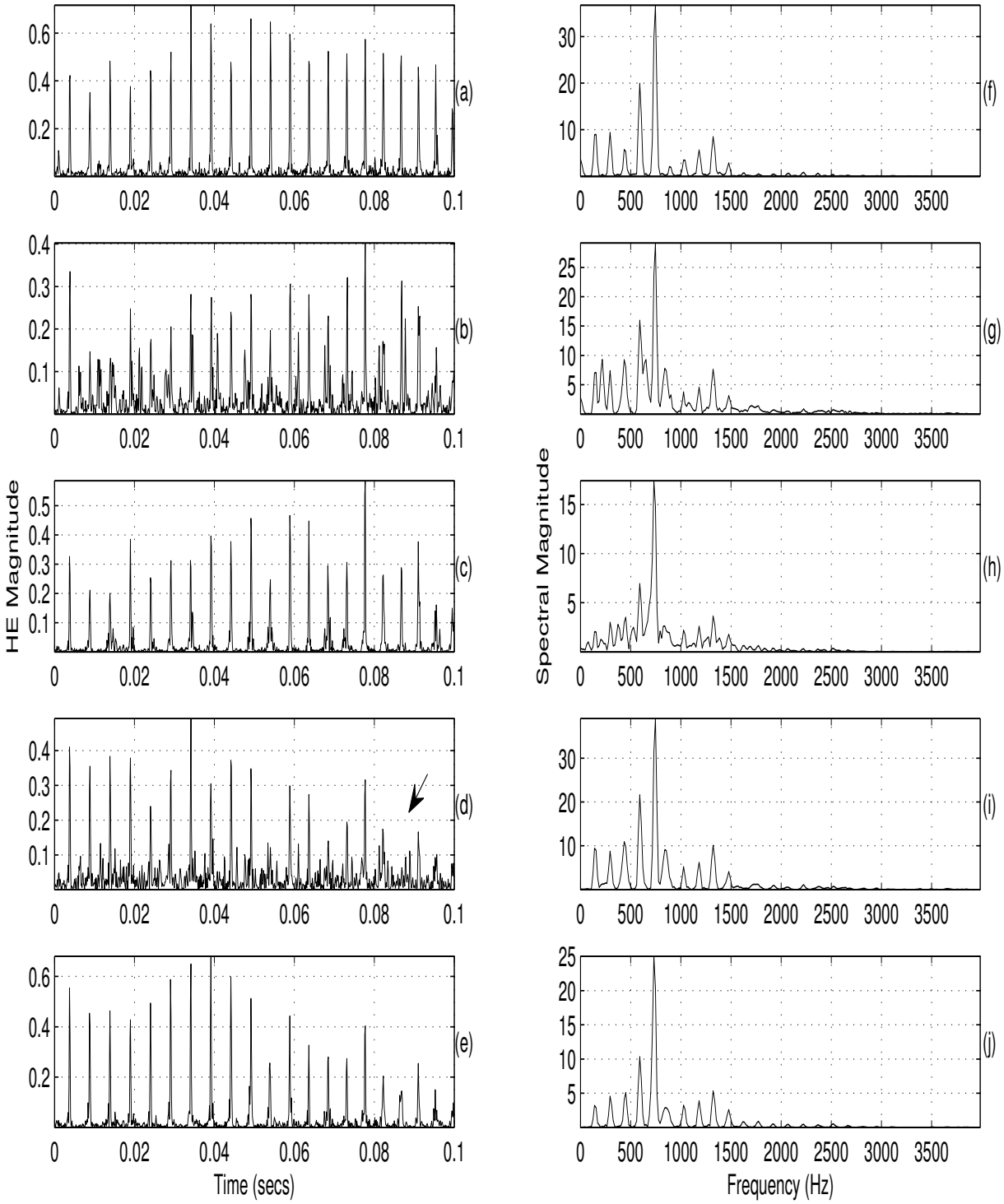


Figure 5.13: HE of LP residual of: (a) clean speech of speaker-1, (b) degraded speech collected from mic-1, (c) speech processed by temporal processing, (d) speech processed by spectral processing, (e) speech processed by temporal and spectral processing, and short time magnitude spectrum of (f) clean speech of speaker-1, (g) degraded speech collected from mic-1, (h) speech processed by temporal processing, (i) speech processed by spectral processing, and (j) speech processed by temporal and spectral processing.

Table 5.2: Percentage of approximate instants derived for different deviations with respect to desired speaker's speech instant locations. The abbreviations S1TP, S1SP and S1TSP refer to speaker-1 separated by temporal, spectral and combined temporal and spectral processing, respectively. Similarly, S2TP/S2SP/S2TSP corresponds to speaker-2. N_{cs} and N_{es} represent total number of instants derived from the clean and the enhanced speech signal, respectively.

Speech Signal	N_{cs}	N_{es}	**Deviation in time**			
			0.5 ms	**1 ms**	**1.5 ms**	**2 ms**
TIMIT Database Examples						
S1TP	3616	3654	87.28	91.01	92.87	94.33
S1SP	3616	3247	73.4	76.5	78.41	79.95
S1TSP	3616	3638	88.34	91.98	93.83	94.91
S2TP	2966	3210	82.21	86.60	88.31	90.31
S2SP	2966	2655	65.75	71.11	74.5	77.78
S2TSP	2966	3195	83.10	86.68	88.10	89.90
IITG Database Examples						
S1TP	10469	10887	80.34	85.17	88.65	90.99
S1SP	10469	9035	65.02	68.84	71.91	74.28
S1TSP	10469	10738	79.42	84.16	87.26	89.37
S2TP	10348	10609	83.24	87.53	90.71	92.93
S2SP	10348	8951	65.12	68.36	71.03	73.33
S2TSP	10348	10557	82.94	87.07	90.24	92.56

5.5.3 Pitch Estimation Performance

The performance of pitch estimation is evaluated in terms of deviation between pitch frequency of clean speech and temporally processed speech of desired speaker. The pitch frequency of clean speech is estimated from the autocorrelation of the HE of LP residual as described in Section 5.4.1. Then the accuracy of pitch estimation is calculated as

$$P_e = \frac{N_{tp}}{N_{cs}} \times 100 \tag{5.27}$$

where N_{cs} is the total number of frames in the clean speech of desired speaker and N_{tp} is the total number of frames having $F_{cs} > 0$ and $|F_{cs} - F_{tp}| \leq F_r$. The abbreviations F_{cs} and F_{tp} represent pitch frequency (in Hz) of clean and temporally processed speech, respectively. F_r is frequency deviation considered for the evaluation. The results of this evaluation are given in Table 5.3 for different values of F_r (F_r= $\pm$5 and $\pm$10 Hz). A small reduction in the performance may be due to:

(i) During the transition periods such as speech to unvoiced or speech to non-speech, the periodicity information of desired speaker frames may get affected by the weight function (i.e., $F_{cs} > 0$ and $F_{tp} = 0$). The percentage value of this error can be computed as

$$P_{e1} = \frac{N_{e1}}{N_{cs}} \times 100 \tag{5.28}$$

where N_{e1} is the number of frames having $F_{cs} > 0$ and $F_{tp} = 0$

(ii) It may also happen for some frames that the periodicity information of undesired speaker may not be completely deemphasized by the weight function (i.e., $F_{cs} = 0$ and $F_{tp} > 0$). Quantitatively this can be assessed as

$$P_{e2} = \frac{N_{e2}}{N_{cs}} \times 100 \tag{5.29}$$

where N_{e2} is the number of frames having $F_{cs} = 0$ and $F_{tp} > 0$. However from speech separation point of view for this case the proposed spectral processing method attenuates the pitch and harmonics of undesired speaker as described by Eqn. (5.22).

(iii) There may be some random pitch estimates due to the pitch of undesired speaker or periodicity information of desired speaker may get affected (i.e., $F_{cs} > 0$ and $|F_{cs} - F_{tp}| \geq 10$). Quantitatively it can be measured by

$$P_{e3} = P_e - P_{e1} - P_{e2} \tag{5.30}$$

The quantitative values of P_{e1}, P_{e2} and P_{e3} are given in Table 5.3.

Table 5.3: Percentage of accuracy of the pitch estimation with respect to clean speech of the desired speaker's speech pitch frequency. The abbreviation S1TP refers to speech separated by the temporal processing. Similarly, S2TP corresponding to speaker-2.

	P_d		P_{e1}	P_{e2}	P_{e3}
Speech Signal	$F_r = \pm 5$ **Hz**	$F_r = \pm 10$ **Hz**			
	TIMIT Database Examples				
S1TP	84.27	88.57	3.18	5.12	3.13
S2TP	83.99	87.13	3.79	6.01	3.07
	IITG Database Examples				
S1TP	88.69	90.39	2.38	4.81	2.42
S2TP	88.92	89.31	1.99	5.26	3.44

5.5.4 Objective Quality Measures

The performance evaluation of combined TSP is done using ideal binary mask [314]. The ideal binary mask for a target signal is found for each time-frequency (T-F) unit by comparing the energy of desired signal to the energy of undesired signal. Whenever the desired signal energy is higher within a T-F unit, the T-F unit is assigned the value 1 and whenever the competing speech signal have more energy, the T-F unit is assigned the value 0. Then for each of the separated signals, the percentage of energy loss P_{EL} and percentage of noise residue P_{NR} are calculated as [251]

$$P_{EL} = \frac{\sum\limits_{n} e_1^2(n)}{\sum\limits_{n} I^2(n)} \tag{5.31}$$

$$P_{NR} = \frac{\sum\limits_{n} e_2^2(n)}{\sum\limits_{n} O^2(n)} \tag{5.32}$$

where $O(n)$ is the estimated signal, and $I(n)$ is the recorded mixture synthesized after applying the ideal binary mask. $e_1(n)$ denotes the signal present in $I(n)$ but absent in $O(n)$ and $e_2(n)$ denotes the signal present in $O(n)$ but absent in $I(n)$. P_{EL} indicates the percentage of target speech excluded from segregated speech, and P_{NR} the percentage of intrusion included. To obtain, $e_1(n)$ a mask is constructed as follows. A T-F unit is assigned 1 if and only if it is 1 in the ideal binary mask but 0 in the segregated target stream. $e_1(n)$ is then obtained by synthesizing the input mixture from the obtained mask. $e_2(n)$ is obtained in a similar way.

Also the output signal to noise ratio (SNR_o) can be measured. Here the SNR_o is defined using the synthesized speech from the ideal binary mask as the ground truth [251].

$$SNR_o = 10\log_{10}\left[\frac{\sum\limits_{n} I^2(n)}{\sum\limits_{n} (I(n) - O(n))^2}\right] \tag{5.33}$$

From the SNR_o, the SNR gain is estimated as

$$\Delta SNR = SNR_o - SNR_i \tag{5.34}$$

where SNR_i is the input signal-to-noise ratio before separation. It is estimated as the ratio of desired signal energy to the competing speaker speech signal energy.

The quantitative values of different objective quality measures described earlier are given in Table 5.4 for individual and combined processing methods. It is known fact that in the proposed method,

second level of spectral processing may further slightly reduce energy level of the desired speaker and hence percentage of energy loss is slightly higher for combined method compared to individual processing methods. However from P_{NR} measure, the amount of undesired speaker is reduced significantly in combined processing method. Table values also show SNR improvement in dB for individual and the combined processing methods. Combined one shows relatively higher improvement than individual processing methods. Even though P_{EL} is high for combined method, significant reduction in the undesired speaker results in high SNR improvement.

Further experiments have been conducted on TIMIT database examples to analyze the performance of temporal and combined processing method under different values of slope parameter (λ) and threshold (T) in Eqn. (5.7). First, to study the effect λ, the SNR gain (ΔSNR) was computed for different values of λ (λ= 5, 10, 15, 20, 25 and 30) and the results are given in Table 5.5. The contents of the table show that, value of λ is not very critical, as long as it is in a particular range. However, lower values of λ, for example λ=5 has low rate of change and therefore during the transitions of non-speech to speech region some of the initial portions of the voiced segment may not get detected, which results slightly lower SNR gain. On the other hand, higher values of λ may results some spurious detections. That is, some of non-speech regions may be misclassified as voiced. Hence, the value of λ is chosen as 20 in the present work.

The next experiment was conducted to evaluate the performance for various values of threshold (T). Here T is assumed to be $T_h \times \overline{h_{sm}}$ (average value of $h_{sm}(n)$). The results obtained for five different values of T_h (T_h= 0.1, 0.2, 0.3, 0.4 and 0.5.) are given in Table 5.6. For this case also, the improvement in SNR is negligible as long as T_h is in a particular range. It is known fact that higher values of threshold does not detect the weak voiced regions which are of short durations and hence it shows the lower SNR gain for T_h=0.5. At the same time, lower value of T_h (T_h= 0.1) results in more non-speech segments being included. Therefore, we have chosen the value of T_h as 0.2 in this study.

Table 5.4: Objective quality measures. The abbreviations P_{EL}, P_{NR}, SNR_i, SNR_o and ΔSNR refer to percentage of energy loss, percentage of noise residue, input signal to noise ratio, output signal to noise ratio and SNR improvement, respectively. S1TP, S1SP and S1TSP refer to speaker-1 separated by temporal, spectral and combined temporal and spectral processing, respectively. Similarly, S2TP/S2SP/S2TSP corresponding to speaker-2.

	Objective Measures				
Condition	P_{EL}	P_{NR}	SNR_i	SNR_o	ΔSNR
	TIMIT Database Examples				
S1TP	8.41	6.63	-0.73	7.70	8.42
S1SP	18.04	9.68	-0.73	7.01	7.74
S1TSP	20.97	3.65	-0.73	8.54	9.26
S2TP	9.50	4.02	0.73	8.50	7.77
S2SP	17.03	6.68	0.73	7.93	7.20
S2TSP	18.83	1.27	0.73	9.20	8.47
	IITG Database Examples				
S1TP	9.02	3.32	-0.04	12.23	12.28
S1SP	13.79	6.94	-0.04	9.95	10.00
S1TSP	14.28	2.95	-0.04	13.64	13.69
S2TP	12.14	2.64	0.04	11.76	11.71
S2SP	14.97	5.69	0.04	7.87	7.83
S2TSP	16.11	1.55	0.04	12.04	12.00

Table 5.5: SNR Gain (ΔSNR) for different values of slope parameters (λ) in Eqn. (5.7). The abbreviations S1TP and S1TSP refer to speaker-1 separated by temporal and combined temporal and spectral processing methods, respectively. Similarly, S2TP and S2TSP corresponding to speaker-2.

Condition / λ	**S1TP**	**S1TSP**	**S2TP**	**S2TSP**
5	6.78	8.40	7.03	8.25
10	7.80	9.03	7.46	8.35
15	8.27	9.18	7.68	8.43
20	8.42	9.26	7.77	8.47
25	8.43	9.29	7.77	8.47
30	8.46	9.29	7.78	8.49

Table 5.6: SNR Gain (ΔSNR) for different values of threshold in Eqn. (5.7). The abbreviations S1TP and S1TSP refer to speaker-1 separated by temporal and combined temporal and spectral processing methods, respectively. Similarly S2TP and S2TSP corresponding to speaker-2.

Condition / T_h	**S1TP**	**S1TSP**	**S2TP**	**S2TSP**
0.1	8.57	9.26	7.89	8.46
0.2	8.42	9.26	7.77	8.47
0.3	8.54	9.25	7.86	8.44
0.4	8.16	9.12	7.67	8.43
0.5	7.80	8.89	7.52	8.32

5.5.5 Subjective Quality Measures

Subjective tests are conducted to get subjective opinion about proposed method. Mean opinion score (MOS) is used [2]. The subjects are 20 graduate students who volunteered for the task. The evaluation was conducted by playing speech signals through head phone set in a laboratory environment. The degraded and enhanced speech signals for all the examples are pooled together, coded in filenames and presented in random order. The subjects are asked to rate each signal separately.

Two different subjective studies are performed. The objective of first study is to get opinion about separation of speech achieved in the combined method as well as individual processing methods, without taking into consideration about signal distortion present. The subjects are asked to rate the speech signal under test on a 5 point scale as given in Table 5.7. The objective of second study is to get opinion about perceptual distortion present in the signals processed by proposed combined as well as individual processing methods, without taking into account about separation achieved. The subjects are asked to rate the speech under test on a 5 point scale as given in Table 5.8.

To further validate the robustness of proposed method with reference to different acoustic environments, two other sets of two speaker data was collected in our laboratory. In case of real data the recording scenario involves collection of speech produced by two speakers speaking simultaneously using two microphones separated by about 1.5 m. The microphones are approximately 1 to 2 m from the speakers. The data is collected in a laboratory environment with associated background noise and reverberation. The locations of speakers and microphones are fixed throughout the recording, so that all the delays are constant. Further, the speakers are positioned in such a way that the delay is different for different speakers. The two cases are two male and one male and one female. The acoustic environment had both reverberation and background noise. The data was collected when the two speakers are speaking simultaneously. The two microphones are separated by 0.6 m, and they are about 1 m distance from the speakers.

The summary of MOS ratings for different speech signals obtained from different processing methods are given in Table 5.9. For speech separation, proposed combined method performs better compared to individual processing methods for both synthetic mixtures and real data. However, temporal processing is accompanied with slight distortion which is inevitable. Note that, spectral processing results slightly higher MOS rating than temporal processing. However this relative merit is mainly because of pitch estimate obtained from temporally processed speech. The rating for level of distortion

is improved from temporal processing to combined temporal and spectral processing. This may be due to the attenuation of spectral peaks of undesired speaker and further enhancement of spectral peaks of desired speaker. With reference to distortion rating, the high MOS rating for synthetic degraded speech signal is mainly because there is no background noise or reverberation present in the original speech mixture.

Table 5.7: The 5-point scale used for obtaining mean opinion scores (MOS) about speech separation achieved.

Rating	Speech Quality	Level of Separation
5	Excellent	Background speaker imperceptible and natural
4	Good	Background speaker just perceptible but not annoying
3	Fair	Background speaker perceptible and slightly annoying
2	Poor	Background speaker annoying but not objectionable
1	Unsatisfactory	Background speaker very annoying and objectionable

Table 5.8: The 5-point scale used for obtaining mean opinion scores (MOS) about distortion introduced.

Rating	Speech Quality	Level of Distortion
5	Excellent	Distortion imperceptible and natural
4	Good	Distortion just perceptible but not annoying
3	Fair	Distortion perceptible and slightly annoying
2	Poor	Distortion annoying but not objectionable
1	Unsatisfactory	Distortion very annoying and objectionable

Table 5.9: MOS for different speech signals of the examples of synthetic mixtures and the examples collected from the real acoustic laboratory environment. The abbreviations DEG, S1TP, S1SP and S1TSP refer to degraded speech, speaker-1 separated by temporal, spectral and combined temporal and spectral processing methods, respectively. Similarly, S2TP/S2SP/S2TSP corresponding to speaker-2.

	Synthetic Mixtures		Real Data	
Condition	MOS for speech separation	MOS for distortion	MOS for speech separation	MOS for distortion
DEG	1.02	4.03	1.15	3.15
S1TP	2.48	2.56	2.65	2.25
S1SP	2.73	3.27	2.95	3.05
S1TSP	3.68	3.38	3.40	3.20
S2TP	2.89	2.82	2.40	2.15
S2SP	2.97	3.14	3.00	3.40
S2TSP	3.55	3.33	3.35	3.15

5.6 Summary

In this chapter the significance of combined TSP for separating speech of desired speaker from two speaker speech is demonstrated. Temporal processing method uses the knowledge of excitation source information present in the degraded speech for separation. The information about instants of significant excitation is derived from the degraded speech. The extracted instants of significant excitation are used for estimating time delay for each speaker. The estimated time delays are used for separating significant excitation regions of each speaker. The separated excitation characteristics are used for deriving weight function for each speaker. The separation of speech for the desired speaker is achieved by weighting degraded speech signal using the weight function. The speech signal enhanced at temporal level is further subjected to spectral domain processing. In spectral domain, enhancement is obtained by sampling and enhancing pitch and harmonics of desired speaker spectra and attenuating pitch and harmonics of undesired speaker spectra. The performance of proposed algorithm is demonstrated for both synthetic mixtures and real data. The improved performance obtained by the proposed combined method demonstrates the complementary nature of information exploited by the temporal and spectral based methods for speech separation.

6

Evaluation of Combined TSP Methods for Speaker Recognition under Degraded Conditions

Contents

6.1 Objective of Evaluation of Combined TSP Methods

In the last three chapters we have presented the combined temporal and spectral processing (TSP) methods for the enhancement of noisy speech, reverberant speech and two speaker speech. The performance of the proposed TSP methods is evaluated using various objective quality measures depending on the nature of the degradation. The results showed that the combined TSP methods give relatively higher performance than the individual temporal and spectral processing method. This chapter presents an experimental evaluation of the combined TSP methods for speaker recognition task under uncontrolled environments. Automatic speaker recognition system gives good performance in controlled environments. Speech recorded in real environments by distant microphones is degraded by factors like background noise, reverberation and competing speaker. This degradation strongly affects the performance of the speaker recognition system. Combined TSP methods proposed in the earlier chapters are used for pre-processing to improve the speaker-specific features and hence the speaker recognition performance.

6.2 Introduction to Speaker Recognition under Degraded Conditions

Speaker recognition is the process of automatically recognizing the speakers on the basis of individuality information from the speech signals [315]. It can be divided into speaker identification and speaker verification. In speaker identification, the task is to identify the speaker from the speech signal. The task of a speaker verification system is to authenticate the claim of a speaker based on the test speech. [316]. Automatic speaker recognition is further divided into text-dependent and text-independent methods. In a text-dependent system the text of speech utterances to be used for training and testing should be the same. For a text-independent system there is no restriction on the text used for training and testing [316].

Generally, speaker recognition by a machine involves three stages. They are, (i) feature extraction, (ii) training or modeling and (iii) testing stage. The feature extraction module estimates a set of features from the speech signal that represent speaker-specific information. A speaker-specific model is constructed based on these features. In the testing stage, for an identification system, the test input speaker utterance undergoes feature extraction. Then, a pattern matching scheme finds the best matched trained model to the current unknown speaker utterance. On the testing stage of a

verification system, the pattern matching is made between the claimed identity trained model and the current unknown speaker utterance.

A majority of the speaker models are based on modeling the underlying distribution of the feature vectors from a speaker. When the speech is distorted by any of the below mentioned sources, the speaker-specific features are also distorted and so their distributions are modified. Thus, a speaker model trained using speech from one type of environment will generally perform poorly in recognizing the same speaker using speech collected under different conditions, since the feature distributions are different. Some of the sources that degrade the performance of speaker recognition are

(i) Variations in the speaker characteristics (speaking rate, speaking style and stress)

(ii) Different microphones used in the training and testing environments

(iii) Distortion introduced by the channel such as telephone network

(iv) The acoustic environments: Background noise, reverberation and speech from other speaker(s).

This work concentrates on the acoustical degradation. Speaker recognition systems are generally trained using data obtained under controlled conditions. This data is acquired from noise-free environment using high quality microphones. Practically in any speaker recognition application the input speech signal may not always be clean, in particular during testing, and may be corrupted in many ways that can degrade the quality of the speech signal and therefore reduce the performance of the speaker recognition system [317–320]. Many strategies have been adopted to deal with acoustical degradation and to provide the robustness to the recognizer. The process of providing robustness to the recognizer can be accomplished in different stages as follows

(i) **Robustness at the signal level (Speech enhancement):** In this class of approach speech signals are enhanced before the feature extraction stage. Accordingly before being transformed into feature vectors, the degraded speech undergoes an enhancement step which tries to filter out the degradation [318].

(ii) **Robustness at the feature level (Feature compensation):** The features representing the speech signal are designed in order to be less sensitive to the degraded conditions. This is achieved by analyzing the influence of the degradation on the speech signal and deriving feature extraction methods that reduce the influence of the degradation [321–324].

(iii) **Robustness at the classifier level (Model compensation):** The aim of the model compensation approach is to determine the influence of the degradation on the distribution of the speech features and to modify the models used in the recognition to take into account about the influence of the degradation. The robustness can be achieved by integrating a model for the speech signal distortion into the overall classifier model [325, 326] or mapping the classifier model obtained during training to better fit the testing condition [327].

The present work aims to provide the robustness at the signal level using the combined TSP methods as a pre-processing stage. Text-independent speaker identification task is taken for demonstration. The approach followed in the presented speaker identification system is schematically illustrated in Fig. 6.1. Here *pre-processing* block refers to temporal processing, spectral processing and the combined TSP methods. A brief summary of various steps involved in the combined TSP of noisy speech, reverberant speech and two speaker speech is given in the Tables 6.1, 6.2 and 6.3, respectively.

The rest of the chapter is organized as follows: Section 6.3 explains the database and experimental setup used in the present study. Section 6.4 describe the experimental results for the case of noisy speech, reverberant speech and multi-speaker speech. Finally, the summary of the work presented in this chapter is mentioned in Section 6.5.

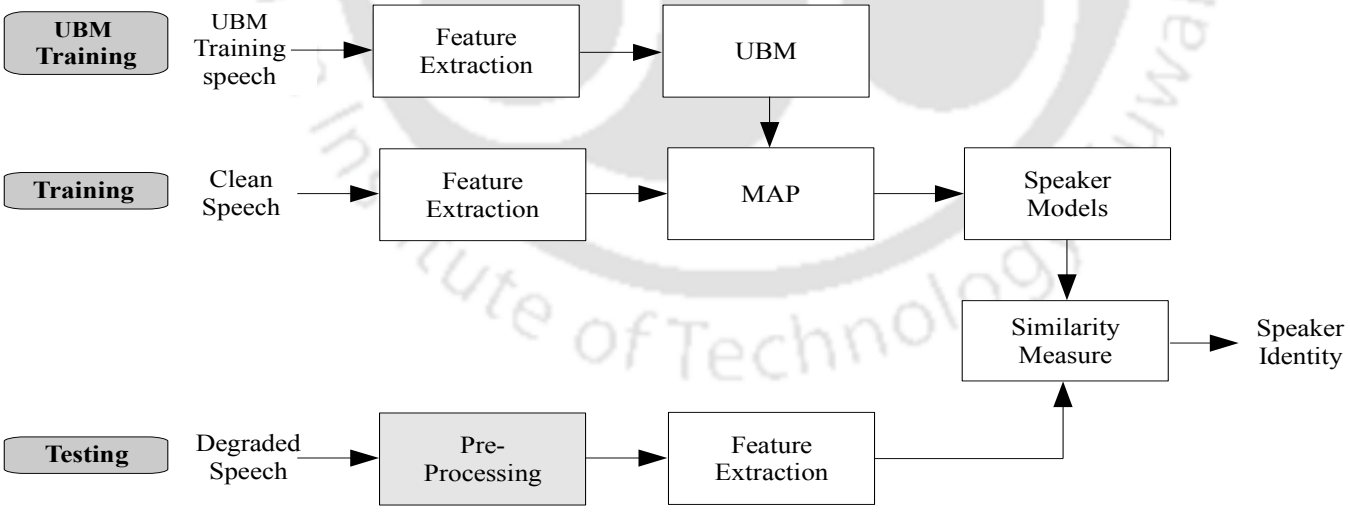


Figure 6.1: Block diagram of speaker recognition under degraded conditions.

Table 6.1: Combined TSP algorithm for enhancement of noisy speech

Temporal Processing:

Gross Level Processing

- Compute the sum of the ten largest peaks in the discrete Fourier transform (DFT) magnitude spectrum using frame size of 20 ms and shift of 10 ms . Mathematically, it is expressed as

$$s_d(l) = \sum_{j=1}^{10} |X(k_j, l)| \tag{6.1}$$

  where l is the frame index, k_j represents the frequency indexes of ten largest spectral peaks and $X(k,l)$ represents the DFT of a frame of noisy speech and is computed as

$$X(k,l) = \sum_{n=0}^{N-1} x(n)w(n-lR)e^{-\frac{j2\pi nk}{N}} \tag{6.2}$$

  where $w(n)$ is a Hamming window, N is the number of points used for computing the DFT and R is the frame shift in samples.

- Compute linear prediction (LP) residual of noisy speech using a frame size of 20 ms, shift of 10 ms and 10^{th} order LP analysis.

- Compute the Hilbert envelope (HE) of LP residual.

- Smooth the HE of LP residual using 50 ms Hamming window.

$$h_{sm}(n) = h_e(n) * h_{w1}(n) \tag{6.3}$$

  where $*$ denotes convolution operation and $h_{w1}(n)$ is Hamming window of 50 ms duration.

- Compute the modulation transfer function energies of the noisy speech signal. Mathematically the modulation transfer function energies are expressed as [284]

$$m(i) = \sum_{p=1}^{18} \left[\sum_{k=k_1}^{k=k_2} \left| \hat{X}_p(k,l) \right|^2 \right] \tag{6.4}$$

  where l is the frame index, p represents the critical band number, k_1 and k_2 represent frequency index of 4 Hz and 16 Hz, respectively. $\hat{X}_p(k)$ is the DFT of normalized envelope of p^{th} filter output and is computed as described in Eqn. (6.2).

- Enhance the evidences of SNR regions of each of the above parameters using the first order difference of the evidence plot .

- Sum all the enhanced parameters and normalize the sum with respect to maximum value.

- Nonlinearly map the normalized sum values using a sigmoid nonlinear function

$$w_g(n) = \frac{1}{1 + e^{-\lambda(s_i(n) - T)}} \tag{6.5}$$

  where slope parameter $\lambda = 20$ and T equal to average value of the normalized sum $s_i(n)$. This generates a gross weight function.

Fine Level Processing

- Compute the DFT magnitude and phase spectra for the noisy speech using 1024 point DFT.

- Pick the largest 8 peaks in the DFT magnitude spectrum and corresponding phase values and synthesize the speech.

- Compute the HE of LP residual of the signal obtained.

- Emphasize the regions around the instants of significant excitation using the neighborhood information of each sample in the HE.

- Obtain first order Gaussian differentiator (FOGD) given by

$$g_d(n) = \frac{1}{\sigma\sqrt{2\pi}} \left[e^{-\frac{(n+1)^2}{2\sigma^2}} - e^{-\frac{n^2}{2\sigma^2}} \right], \quad 1 \le n \le L_g \tag{6.6}$$

  where Gaussian window of length $L_g = 80$ samples and variance $\sigma = 8$.

- Convolve the negative of FOGD operator with the mean smoothed HE of the LP residual and determine negative to positive transitions.

- Derive the fine weight function $w_f(n)$ by convolving detected instants with the Hamming window

$$w_f(n) = \left(\sum_{i=1}^{N_i} \delta(n - a_i) \right) * h_w(n) \tag{6.7}$$

  where N_i represents total number of detected instants, a_i is the approximate location of instants and $h_w(n)$ is the Hamming window of 3 ms duration.

Final Weight Function

- Multiply the two weight functions (gross and fine weight functions) to generate the final weight function.

- Multiply the LP residual signal of noisy speech by the final weight function.

- Excite the time-varying all-pole filter (derived from noisy speech) using weighted residual to obtain the temporally processed speech.

Spectral Processing:

- Update the noise magnitude spectrum if 5 consecutive frames are detected as non-speech regions.

- Process the temporally processed speech by any of the conventional spectral processing (e.g., multi-band spectral subtraction [39] or MMSE estimator [14]) methods.

- Reconstruct the enhanced speech signal using IDFT and overlap-add (OLA) method.

Table 6.2: Combined TSP algorithm for enhancement of reverberant speech

Spectral Processing:

- Estimate the late reverberant spectral variance $\hat{S}_{yl}(l,k)$.

$$\hat{S}_{yl}(l,k) = \gamma\omega(l-N_1) * |Y(l,k)|^2 \tag{6.8}$$

where $Y(l,k)$ is the short time Fourier transform of reverberant speech $y(n)$. The symbol * denotes convolution in the time domain and $w(l)$ is a smoothing function. γ specifies the relative strength of the late-impulse components and is set to 0.32 and

$$w(l) = \begin{cases} \frac{l+a}{a^2} e^{\frac{-(l-a)^2}{2a^2}}, & l > -a \\ 0, & \text{otherwise} \end{cases} \tag{6.9}$$

where a controls the span of the smoothing function and is set to 5.

- Compute the *a posteriori* signal to reverberant ratio (SRR) and the *a priori* SRR values. The *a priori SRR* $\xi(l,k)$ is calculated as (decision directed approach) [14]

$$\xi(l,k) = \eta \frac{\left|\hat{S}(l-1,k)\right|^2}{|Y(l,k)|^2} + (1-\eta)\max\{\gamma(l,k)-1, 0\}. \tag{6.10}$$

The value of η is chosen as 0.98 [14] and

$$\gamma(l,k) = \frac{|Y(l,k)|^2}{\hat{S}_{yl}(l,k)} \tag{6.11}$$

where the term $\gamma(l,k)$ is interpreted as the *a posteriori SRR*. $\hat{S}(l,k)$ is computed as given in Eqn. (6.14). The following initial condition is used for the first frame

$$\xi_k(l,k) = \eta + (1-\eta)\max\{\gamma(l,k)-1, 0\} \tag{6.12}$$

- Determine the gain function $G(l,k)$ for the the spectral subtraction from the estimated SRR values.

$$G(l,k) = 1 - \frac{1}{\sqrt{1+\xi(l,k)}}. \tag{6.13}$$

- Multiply the gain function with the reverberant speech spectrum.

$$\hat{S}(l,k) = Y(l,k)G(l,k) \tag{6.14}$$

- Obtain the spectrally processed speech using IDFT and OLA method.

Temporal Processing:

Gross Level Processing

- Compute LP residual of spectrally processed speech using a frame size of 20 ms, shift of 10 ms and 10^{th} order LP analysis.
- Compute the sum of the ten largest peaks in the DFT magnitude spectrum.
- Compute the HE of LP residual and mean smooth using 50 ms Hamming window.
- Compute the modulation spectrum energies of the spectrally processed speech signal.
- Enhance the high SRR regions of each of the above parameters.
- Sum all the enhanced parameters and normalize the sum with respect to maximum value.
- Nonlinearly map the normalized sum values by using a sigmoid nonlinear function with slope parameter $\lambda = 20$ and T equal to average value of the normalized sum. This generates a gross weight function.

Fine Level Processing

- Band pass filter the LP residual of spectrally processed speech into four subbands whose cut-off frequencies are equally spaced in linear scale.
- Compute the HE of LP residual for each subband.
- Sum all the subband HEs.
- Compute the emphasized HE of the LP residual.
- Obtain FOGD operator from Gaussian window of length $L_g = 80$ samples and $\sigma = 8$.
- Convolve the negative of FOGD operator with the emphasized HE of the LP residual and determine negative to positive transitions.
- Derive the fine weight function by convolving detected instants with the Hamming window of 3 ms duration.

Final Weight Function

- Multiply the two weight functions (gross and fine weight functions) to generate the final weight function.
- Multiply the LP residual signal of noisy speech by the final weight function.
- Excite the time-varying all-pole filter (derived from spectrally processed noisy speech) using weighted residual to obtain the enhanced speech.

Table 6.3: Combined TSP algorithm for two speaker separation

Temporal Processing

- Compute the LP residual of mic-1 and mic-2 signals using a frame size of 20 ms, shift of 10 ms and 10^{th} order LP analysis.
- Compute the HE of LP residual.
- Estimate the time delays for each speaker by computing cross-correlation of the HEs using a frame size of 50 ms and shift of 5 ms. In the normalized cross-correlation sequence, the displacement of peak with respect to the center sample is considered as the time delay value [23].
- Adjust the HEs using the estimated time delays to produce the coherently adjusted HE for each speaker. That is,

$$h_{s1}(n) = \min\left(h_1(n), h_2(n-d_1)\right) \tag{6.15}$$

$$h_{s2}(n) = \min\left(h_1(n), h_2(n-d_2)\right) \tag{6.16}$$

  where $h_1(n)$ and $h_2(n)$ be the normalized HE sequences of speech signals collected at mic-1 and mic-2, respectively and d_1 and d_2 are the time-delays between the two microphone signals.
- Compute the error function. That is,

$$h_{12}(n) = h_{s1}(n) - h_{s2}(n). \tag{6.17}$$

Gross Level Processing

- Compute smoothed error function using 50 ms Hamming window. The smoothed sequence is obtained as

$$h_{sm}(n) = |h_{12}(n)| * h_{w1}(n) \tag{6.18}$$

  where $*$ denotes convolution operation and $h_{w1}(n)$ is Hamming window of 50 ms duration. The major peaks in the smoothed difference values $h_{sm}(n)$ indicate approximate locations of instants of significant excitation. In particular, positive peaks correspond to the instants of desired speaker and negative peaks correspond to the instants of undesired speaker
- Nonlinearly map smoothed error function values by using a sigmoid nonlinear function with slope parameter $\lambda = 20$ and T equal to 0.2 times average value of the normalized sum.
- The nonlinearly mapped values is termed as gross weight function.

Fine Level Processing

- Smooth the error function using the Hamming window.

$$h_s(n) = h_{12}(n) * h_{w2}(n) \tag{6.19}$$

  where $h_{w2}(n)$ is the Hamming window of 3 ms duration.
- Convolve the negative of FOGD operator with the positive values of mean smoothed HE of the LP residual to determine the desired speaker instants location.
- Convolve the negative of FOGD operator with the negative values of mean smoothed HE of the LP residual to determine the interfering speaker instants location.
- Compute the fine weight function.

$$W_f(n) = [w_{\min} + (1 + w_{\min})W_a(n)] - w_{\min}W_b(n) \tag{6.20}$$

  where

$$W_a(n) = \sum_{i=1}^{N_a} \delta\left(n - a_i\right) * h_{w2}(n) \tag{6.21}$$

$$W_b(n) = \sum_{i=1}^{N_b} \delta\left(n - b_i\right) * h_{w2}(n) \tag{6.22}$$

  where a_i and b_i represent the approximate locations of instants of significant excitation of desired and undesired speaker, respectively. N_a and N_b represent total number of detected instants of desired and undesired speaker, respectively. $w_{\min}$ is set as 0.3 and $h_{w2}(n)$ is the Hamming window of 3 ms duration.

Final Weight Function

- Multiply the two weight functions (gross and fine weight functions) to generate the final weight function.
- Multiply the LP residual signal of noisy speech by the final weight function.
- Excite the time-varying all-pole filter (derived from degraded speech) using weighted residual to obtain temporally processed speech.

Spectral Processing

- Compute the HE of enhanced LP residual (i.e., LP residual weighted by the weight function).
- Perform the autocorrelation on the HE of LP residual using a frame size of 40 ms and a frame shift of 10 ms. The normalized autocorrelation is obtained as [288]
- Find the pitch estimate from the first major after the center peak in the range of 2.5 ms to 12.5 ms.
- Compute the similarity measure (C_m) and the magnitude of first major peak (R_p) in the normalized autocorrelation sequence.
- A frame of speech subjected to autocorrelation is considered as voiced frame only when the values of $R_p \geq 0.4$ [203] and $C_m \geq 0.7$ [211].
- For voiced regions sample and enhance the pitch and harmonics of the desired speaker.
- Reconstruct the enhanced speech signal by IDFT and OLA method.

6.3 Database and Experimental Description

(i) **Database :** The speaker recognition studies are carried out on the TIMIT database [289, 290]. The TIMIT database is collected from American English speakers divided into 8 accent regions including speakers that do not have strong regional accents. The database contains 630 speakers, of which 438 are males (70 %) and 192 females (30 %). Each speaker has contributed 10 sentences of approximately 3 sec each. The speech was recorded using a high quality microphone in a sound proof booth with no session interval between recordings. Speech files are stored in NIST "wav"-file format with a sampling frequency of 16 kHz and a quantization resolution of 16 bits per sample. The recordings are single-channel, the mean duration is 3.28 sec and the standard deviation is 1.52 sec [328].

Out of 630 speakers, 100 speakers are randomly selected for forming subset for the study. The selection of the subset was arbitrary. As mentioned, each speaker said 10 different utterances. These 10 utterances that are read can be divided into three groups. The first two utterances are common across all speakers in the database and are known as the SA1 and SA2 utterances. These utterances can be used for speaker normalization and accent identification. The next eight utterances are different across all the speakers and are known as the SA and SI utterances. The common way of using this database is to use the first eight utterances (including the SA1 and SA2 utterances) of each speaker for the model training and the last two utterances for testing [329–333]. Since most of the speech information is present up to 4 kHz, the speech samples are first down sampled to 8 kHz and then used in this work.

(ii) **Feature Extraction :** Feature extraction aims at giving a useful representation of the speech signal by capturing the important information from it. It transforms the speech signal into a compact but effective representation that is more stable and discriminative than the original signal [334]. The performance of speaker modelling strongly depends on the feature extraction step. Several feature extraction methods have been studied for the speaker recognition task [277, 321, 333, 335–361]. Some of the widely used speaker-specific features include linear predictive coefficients (LPC) [337], linear predictive cepstral coefficients (LPCC) [339], real cepstral coefficients (RCC), Mel-frequency cepstral coefficients (MFCC) [344] and perceptual linear prediction (PLP) coefficients [350].

MFCCs have been demonstrated to show good performance in speaker recognition. Experimental evaluation of recognition accuracy of the MFCC, LPCC and PLP coefficients is made in [362] and result of this study is that all features perform poorly without some form of channel compensation, however, with channel compensation MFCC slightly outperforms other types. MFCCs are estimated based on human perception of critical bandwidths. The mel-frequency scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. In the present work we have also used MFCC features to increase robustness and performance of the system. For calculation of MFCCs, the short term Fourier transform (STFT) analysis is performed on the speech signal using frame size of 20 ms with shift of 10 ms. For each frame the STFT magnitude spectrum is computed and is further processed by the 24 triangular shaped mel-filter banks to find out the filter bank energies. Then discrete cosine transform (DCT) is taken on the spectral energies to obtain MFCCs. We have used a 13-dimensional MFCC vector (excluding c_0) appended with delta (Δ) and delta-delta ($\Delta\Delta$) coefficients as feature vector of each frame. A detailed description of the MFCC feature extraction technique is given in Appendix-E.

(iii) **Speaker Modelling :** There are several methods for speaker modelling. In text-dependent speaker recognition the most popular methods are dynamic time warping (DTW) [345] and hidden Markov models (HMM) [363]. In text-independent speaker recognition the most popular methods are vector quantization (VQ) [364], artificial neural networks (ANN) [365,366], support vector machines (SVM) [367], and Gaussian mixture models (GMM) [368]. These models can be classified into parametric and non-parametric models. The parametric models have a particular model structure characterized by certain parameters for the distribution of feature vectors [369]. For instance, GMM. In non-parametric models no such assumption is made. For instance, VQ and ANN.

Literature shows that many researchers have implemented GMM models in the text-independent speaker recognition system [329,365,368–376]. Because of the ability to model multivariate densities the GMM approach is well suited for text-independent speaker recognition. This approach has become the most dominant in this field and furthermore has achieved state of the art performance. We therefore implemented a speaker model based on the universal background model (UBM)-GMM for modelling the speakers [373]. The UBM-GMM can be mainly divided into three parts: UBM training, Bayesian adaptation of speaker models and speaker identification.

A speaker-independent model, or UBM, is trained from the non-target speakers using the EM algorithm. Then for each target speaker, speaker models are created through MAP adaptation of the UBM using speaker-specific training speech. Based on experimental results, the best performance can be achieved using only the mean adaptation of Gaussian mixtures [373, 376, 377]. A detailed description about the GMM and UBM-GMM is given in Appendix-F.

(iv) **Testing :** The recognition accuracy of speaker identification is measured by identification rate and is defined by the number of correctly identified utterances (N_{crt}) to the total number of testing utterances (N_{tot}). That is,

$$P_i = \frac{N_{crt}}{N_{tot}} \times 100. \tag{6.23}$$

From previous experiments conducted for speaker recognition, Reynolds *et al.* [373] has found that only a few of the mixtures of a GMM contributes significantly to the likelihood value for a speech feature vector. In addition, the mixture components of the adapted model of each speaker share a certain correspondence with the UBM, therefore log-likelihood score of the speaker model can be computed by scoring only the more significant mixtures. More commonly only the top five mixtures are used [373, 376]. The computation requirement for recognition is reduced significantly by employing this mixture scoring strategy.

6.4 Experimental Results and Discussions

In this work first the speaker models are created using clean speech data by UBM-GMM concept. The UBM is trained on approximately one hour of data (excluding silence regions) of the TIMIT database using the EM algorithm. The UBM training data are taken from *train* set of the TIMIT database and for evaluation speech samples are taken from the *test* set of the TIMIT database. A first set of experiment is conducted to evaluate the performance of the speaker recognition system under clean condition and is found to be 97 %. In the next step to evaluate the performance of the combined TSP method on degraded speech the approach followed is schematically illustrated in Fig. 6.1. In figure the pre-processing block refers to temporal processing, spectral processing and the combined TSP methods. The recognition studies are carried out on test speech (degraded speech) data with and without pre-processing.

6.4.1 Speaker Recognition in Noisy Environment

The noisy speech is created by adding white noise from NOISEX-92 database to the test utterances. The noise waveform being added to the speech was scaled to give the desired global SNR. The value of SNR is varied over the range of 0 - 30 dB. While testing, the degraded speech signal is enhanced using individual and the proposed TSP methods as given in Table 6.1 prior to the feature extraction step. In the present work, the conventional multi-band spectral subtraction [39] and the MMSE-STSA estimator [14] methods are used for combination. For MMSE-STSA estimator the *a priori* SNR for each frequency component is estimated using a decision directed variance estimator approach [14]. In this approach, the variance estimator at a given frame uses the signal spectral magnitude estimate from the previous frame along with the current noisy spectral component.

To illustrate the merit of combined temporal and spectral processing, Figs. 6.2(a)-(e) show the excitation source signal (LP residual) spectrum of clean, degraded, temporal, spectral and combined temporal and spectral processed speech, respectively. Similarly, Figs. 6.2(f)-(j) show the vocal spectrum of clean, degraded, temporal, spectral and combined temporal and spectral processed speech, respectively. The combined processing shows improvement in both the excitation source and vocal tract spectrum, whereas individual processing methods show major improvement either at the excitation source signal or at the vocal tract spectrum only. As a result the extracted speaker-specific features of combined TSP method are more robust compared to individual processing methods and thus may result in improved speaker recognition performance.

Table 6.4 and 6.5 show the identification results for the various SNR levels. From Table 6.5, it can be seen that the recognition performance of the combined temporal and spectral processing is higher than individual processing methods. However for higher SNR values, in particular for 30 dB, the combined method results slightly lower performance than spectral processing alone. This is mainly because the underlying temporal processing method involves weighting of the LP residual for enhancement. For higher SNR values, since noise level is very low, the weighting may disturb the actual signal and thus results in slight reduction of performance.

6.4.2 Speaker Recognition in Reverberant Environment

The reverberant speech signal is generated through a linear convolution between the original speech data and a room impulse response. We have generated five impulse responses using image method

having the reverberation time in the range of 0.2 - 1.0 sec with source microphone distance of 1.5 m. The same procedure which is followed in the noisy speech experiment is repeated and the obtained results are given in Table 6.6 and 6.7. The TSP algorithm used for enhancing the reverberant speech is given in Table 6.2. From the results, it can be observed that the TSP method gives improved performance than individual processing method. This is mainly because

(i) Spectral processing removes the late reverberant portion of reverberant speech

(ii) Temporal processing enhances the early reverberant portion of reverberant speech.

As a result speech processed by the combined TSP method gives lesser spectral distance with reference to clean speech and therefore extracted MFCC features are more closer to that of the clean speech MFCC features. This leads to the improvement in the speaker identification rate.

6.4.3 Speaker Recognition in Multi-Speaker Environment

For this study, first a set of 100 speakers different from the UBM training and testing set is chosen as interfering speakers. The synthetic two speaker data for a pair of microphones are created with delays of $d_1 = 8$ and $d_2 = -16$ samples and used as a test signal for speaker recognition study. Two different types of two speaker data are created, they are (i) the gender of the interfering speaker is same as that of original test speaker, and (ii) interfering speaker gender is different from that of test speaker. The degraded signal is preprocessed according to the method given in Table 6.3 and the preprocessed speech signal is used for capturing MFCC features. Table 6.8 shows the result of this study for different pre-processing techniques. The relative improvement in the performance of speaker recognition with the different pre-processing techniques is clearly seen from the results. It can be observed that identification rate of same gender case is less than the different gender case. This may be interpreted in following way. In all pitch based separation methods, speech separation not only depends on the processing method used but also on the nature of the degraded signal. The more separated in the pitch and harmonics of each talker, the better the result to be expected. For same gender case the separation between the pitch of desired and interfering speaker may be minimum. Therefore for this case, the temporal and spectral processing may leave some of the interfering speaker information in the enhanced speech and thus results slightly poor performance than different gender case.

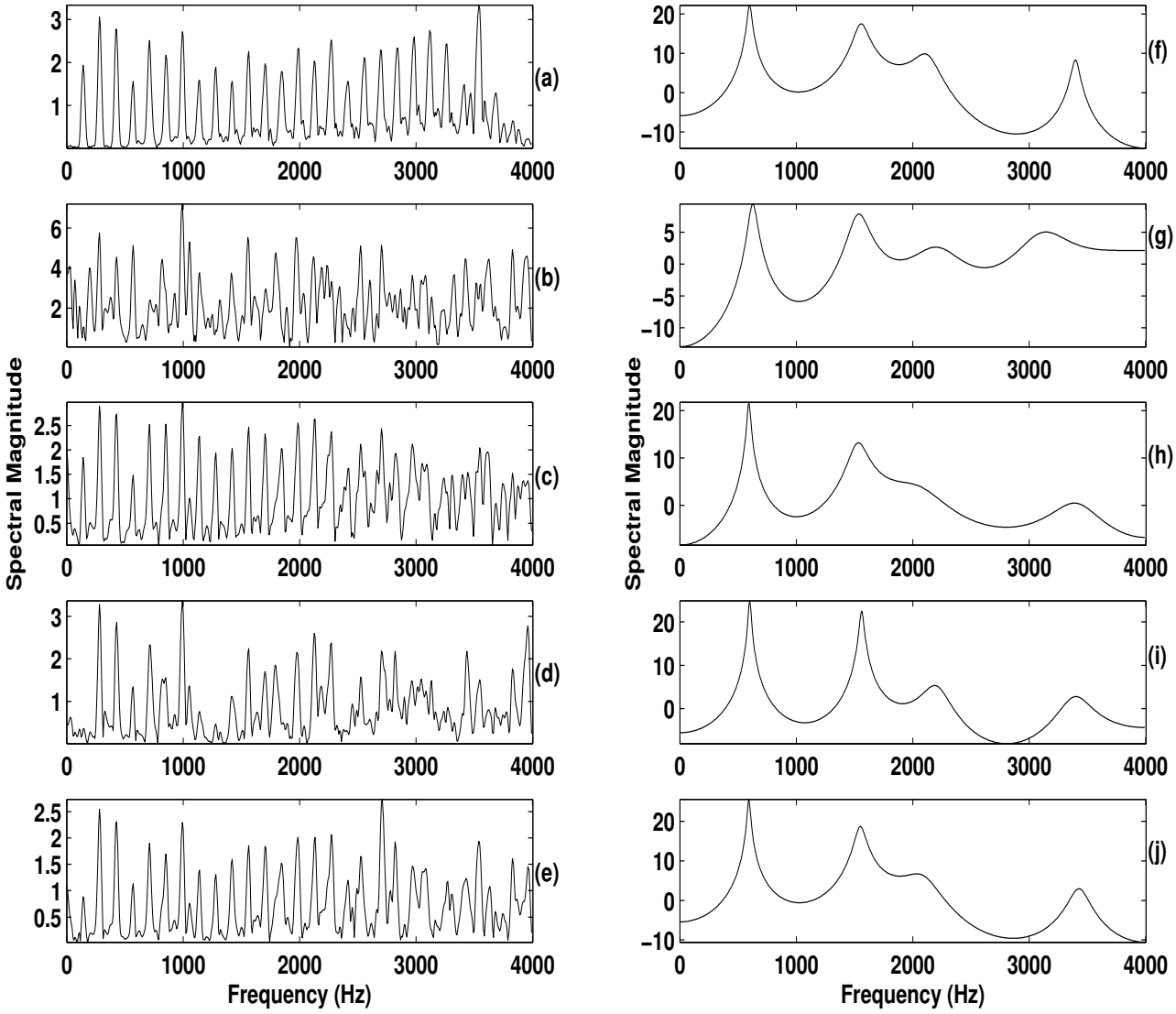


Figure 6.2: Noisy speech enhancement: Excitation source spectrum of a frame of (a) clean speech, (b) degraded speech, (c) speech processed by temporal processing, (d) speech processed by spectral processing, (e) speech processed by temporal and spectral processing, and Vocal tract (LP) spectrum of a frame of (f) clean speech, (g) degraded speech, (h) speech processed by temporal processing, (i) speech processed by spectral processing, and (j) speech processed by temporal and spectral processing.

Table 6.4: Speaker recognition performance under noisy environment. In table abbreviations DEG, TP, SP1, SP2, TSP1 and TSP2 refer to degraded speech, temporal processing, multi band spectral subtraction, MMSE-STSA estimator, combined temporal and multi-band spectral subtraction and combined temporal and MMSE-STSA estimator, respectively. P_i represents percentage of identification.

	No. of Gaussians									No. of Gaussians							
	8	**16**	**32**	**64**	**128**	**256**	**512**	**Max (P_i)**		**8**	**16**	**32**	**64**	**128**	**256**	**512**	**Max (P_i)**
Clean	85.0	92.5	94.0	95.5	95.5	97.0	96.5	**97.0**	**Clean**	85.0	92.5	94.0	95.5	95.5	97.0	96.5	**97.0**
	SNR = 0 dB									**SNR = 12 dB**							
DEG	0.5	1.0	0.5	0.5	1.5	1.5	1.5	**1.5**	**DEG**	5.5	5.0	6.0	8.0	8.5	10.5	9.5	**10.5**
TP	1.0	1.0	1.0	1.5	2.5	3.0	2.5	**3.0**	**TP**	18.5	24.0	29.0	32.5	33.5	33.0	30.0	**33.5**
SP1	4.5	5.0	6.0	3.5	5.0	6.0	6.0	**6.0**	**SP1**	38.5	39.5	49.5	44.5	45.0	43.0	40.5	**49.5**
SP2	4.5	4.5	3.5	2.5	5.5	4.0	5.5	**5.5**	**SP2**	43.0	48.0	55.5	53.0	54.5	50.0	57.0	**57.0**
TSP1	8.5	8.0	10.5	12.5	10.5	9.5	13.0	**13.0**	**TSP1**	54.0	56.5	67.0	66.5	71.0	72.0	72.5	**72.5**
TSP2	4.0	6.0	7.5	8.0	7.5	6.5	7.5	**8.0**	**TSP2**	55.0	56.5	66.0	67.0	66.5	63.5	67.5	**67.5**
	SNR = 3 dB									**SNR = 15 dB**							
DEG	1.0	1.0	1.0	1.5	2.0	2.0	2.0	**2.0**	**DEG**	13.0	12.0	17.0	20.0	19.5	22.0	23.5	**23.5**
TP	1.0	1.0	2.0	1.5	3.5	5.0	2.5	**5.0**	**TP**	27.0	37.0	41.5	42.5	42.5	40.5	40.5	**42.5**
SP1	7.5	12.0	11.0	13.5	14.5	15.0	19.0	**19.0**	**SP1**	56.5	60.5	65.0	67.0	69.5	70.5	68.0	**70.5**
SP2	5.5	8.0	6.5	11.0	10.0	8.0	12.0	**12.0**	**SP2**	60.0	67.5	70.5	77.0	77.0	75.5	74.0	**77.0**
TSP1	13.0	18.0	22.5	17.5	22.0	20.5	24.0	**24.0**	**TSP1**	60.0	65.5	76.5	77.5	83.0	82.5	79.0	**83.0**
TSP2	8.5	13.5	13.5	12.0	11.0	9.0	11.0	**13.5**	**TSP2**	61.0	67.5	73.0	77.0	77.0	80.5	78.5	**80.5**
	SNR = 6 dB									**SNR = 20 dB**							
DEG	1.0	1.0	2.0	2.0	2.0	2.0	2.0	**2.0**	**DEG**	26.5	35.0	38.5	49.0	50.0	51.5	51.0	**51.5**
TP	2.5	7.0	6.0	13.5	10.5	16.0	12.5	**16.0**	**TP**	52.0	60.5	71.0	72.5	78.5	76.0	76.0	**78.5**
SP1	16.5	21.5	27.0	27.0	32.0	28.0	31.5	**32.0**	**SP1**	73.5	79.5	82.5	85.5	87.0	87.0	86.0	**87.0**
SP2	9.0	15.0	16.5	20.0	19.0	15.0	21.0	**21.0**	**SP2**	71.0	76.0	80.0	86.5	84.0	86.0	85.0	**86.5**
TSP1	23.0	29.0	32.0	31.0	36.5	38.0	41.5	**41.5**	**TSP1**	67.0	75.5	78.5	83.5	88.0	86.5	84.5	**88.0**
TSP2	19.5	24.5	27.5	30.0	28.5	24.0	27.5	**30.0**	**TSP2**	70.5	76.0	80.5	85.5	85.5	87.0	83.5	**87.0**
	SNR = 9 dB									**SNR = 30 dB**							
DEG	3.0	2.5	3.0	3.5	3.0	2.5	2.0	**3.5**	**DEG**	59.0	72.5	79.0	86.5	87.0	86.0	89.0	**89.0**
TP	11.0	13.0	15.5	17.0	18.5	21.0	19.5	**21.0**	**TP**	66.0	76.0	82.0	86.0	86.5	87.5	85.0	**87.5**
SP1	31.0	39.0	44.5	42.0	49.5	48.5	49.0	**49.5**	**SP1**	80.0	84.0	86.0	92.0	91.0	90.0	90.5	**92.0**
SP2	22.5	25.0	32.5	32.0	33.5	28.0	36.0	**36.0**	**SP2**	78.0	82.5	85.0	91.5	90.5	91.0	91.5	**91.5**
TSP1	36.0	43.0	46.0	54.5	59.0	57.0	59.0	**59.0**	**TSP1**	71.5	78.5	83.0	86.0	87.5	88.5	89.0	**89.0**
TSP2	37.0	40.0	48.5	50.5	48.0	52.5	53.5	**53.5**	**TSP2**	71.5	79.0	82.5	89.0	87.0	88.0	87.0	**89.0**

Table 6.5: Speaker recognition performance (percentage of identification) under noisy environment. In table abbreviations DEG, TP, SP1, SP2, TSP1 and TSP2 refer to degraded speech, temporal processing, multi band spectral subtraction, MMSE-STSA estimator, combined temporal and multi-band spectral subtraction and combined temporal and MMSE-STSA estimator, respectively.

	SNR Level							
Condition	**0 db**	**3 db**	**6 db**	**9 db**	**12 db**	**15 db**	**20 db**	**30 db**
DEG	1.50	2.00	2.00	3.50	10.50	23.50	51.50	89.00
TP	3.00	5.00	16.00	21.00	33.50	42.50	78.50	87.50
SP1	6.00	19.00	32.00	49.50	59.50	70.50	87.00	92.00
SP2	5.50	12.00	21.00	36.00	57.00	77.00	86.50	91.50
TSP1	13.00	24.00	41.50	59.00	72.50	83.00	88.00	89.00
TSP2	8.00	13.50	30.00	53.50	67.50	80.50	87.00	89.00

Table 6.6: Speaker recognition performance under reverberant environment. In the table abbreviations DEG, TP, SP and TSP refer to degraded speech, temporal processing, spectral processing and combined temporal and spectral processing, respectively. P_i represents percentage of identification and D represents source microphone distance.

	No. of Gaussians							
	8	**16**	**32**	**64**	**128**	**256**	**512**	**Max (P_i)**
D = 1.5 m & T60 = 0.2 sec								
DEG	29.0	32.0	36.0	35.0	43.0	38.0	37.0	**43.0**
SP	42.0	45.5	51.5	56.5	53.0	54.0	53.5	**56.5**
TP	35.0	34.0	37.0	35.5	43.0	36.0	37.5	**43.0**
TSP	38.5	44.5	49.5	51.5	52.0	56.5	52.5	**56.5**
D = 1.5 m & T60 = 0.4 sec								
DEG	24.5	23.0	28.5	31.0	26.5	30.5	30.5	**31.0**
SP	36.0	35.0	40.0	45.5	45.0	49.5	47.0	**49.5**
TP	26.0	24.5	30.0	32.0	32.5	34.5	33.5	**34.5**
TSP	37.0	37.0	39.0	44.0	46.0	53.5	46.5	**53.5**
D = 1.5 m & T60 = 0.6 sec								
DEG	18.5	20.5	21.5	20.0	16.0	23.0	22.5	**23.0**
SP	17.0	18.0	23.0	21.5	17.0	26.0	21.5	**26.0**
TP	23.0	26.5	27.0	32.5	30.0	35.0	29.0	**35.0**
TSP	23.0	26.0	30.5	34.0	32.5	39.5	31.5	**39.5**
D = 1.5 m & T60 = 0.8 sec								
DEG	13.5	15.5	16.0	15.5	18.0	18.5	19.0	**19.0**
SP	20.0	18.0	18.5	23.0	23.0	24.5	24.5	**24.5**
TP	16.5	15.0	21.0	17.5	18.5	19.5	21.0	**21.0**
TSP	21.0	19.5	19.0	24.0	31.5	29.5	25.0	**31.5**
D = 1.5 m & T60 = 1.0 sec								
DEG	8.0	5.5	7.5	8.5	8.0	9.0	7.5	**9.0**
SP	12.5	14.0	17.5	19.0	14.5	19.0	19.0	**19.0**
TP	8.0	7.5	7.5	8.0	9.5	9.0	8.5	**9.5**
TSP	17.5	15.0	21.0	20.0	15.0	21.5	18.0	**21.5**

Table 6.7: Speaker recognition performance (percentage of identification) under reverberant environment. In the table abbreviations DEG, TP, SP and TSP refer to degraded speech, temporal processing, spectral processing and combined temporal and spectral processing, respectively.

	Reverberation Time T_{60} (sec)				
Condition	**0.20**	**0.40**	**0.60**	**0.80**	**1.00**
DEG	43.00	31.00	23.00	19.00	9.00
SP	56.50	49.50	35.00	24.50	19.00
TP	43.00	34.50	26.00	21.00	9.50
TSP	56.50	53.50	39.50	31.50	21.50

Table 6.8: Speaker recognition performance in two speaker environment. In the table abbreviations DEG, TP, SP and TSP refer to degraded speech, temporal processing, spectral processing and combined temporal and spectral processing, respectively. P_i represents percentage of identification.

	No. of Gaussians							
	8	**16**	**32**	**64**	**128**	**256**	**512**	**Max (P_i)**
Same Gender								
DEG	18.5	22.5	29.0	32.0	37.0	36.0	39.5	**39.5**
TP	26.0	35.0	38.0	44.0	45.0	50.5	50.0	**50.5**
SP	27.5	36.5	41.5	50.0	56.0	51.5	56.0	**56.0**
TSP	31.0	45.0	45.5	49.5	59.0	52.0	55.0	**59.0**
Different Gender								
DEG	28.0	31.0	38.0	38.0	41.5	40.0	40.0	**41.5**
TP	32.0	40.0	43.0	45.0	52.0	50.0	49.0	**52.0**
SP	41.0	47.5	48.5	57.0	55.5	58.0	55.0	**58.0**
TSP	42.0	49.5	50.0	58.5	59.5	62.5	60.0	**62.5**

6.5 Summary

The main objective of this chapter is to evaluate the performance of the combined TSP based speech enhancement methods in the speaker recognition task. For this study the TIMIT database is taken and MFCC features are extracted from the clean speech data and the speaker models are trained using UBM-GMM system. In testing stage, the synthetic degraded speech is generated for each type of degradation. The degraded speech is subjected to temporal, spectral and TSP methods before the feature extraction step to attenuate the degradation characteristics. The enhanced speech is subjected to testing. The recognition results show that combined processing method gives relatively higher performance than individual processing methods, except with some limitations. Like in very high SNR values and for lower reverberation times the combined method results slightly lower or equal performance than expected due to underlying processing steps involved in temporal processing.

7

Summary and Conclusions

Contents

7.1 Summary of the Present Work

The objective of the work presented in this thesis is to explore the combined temporal and spectral processing (TSP) method for speech enhancement. As a result we have proposed the combined TSP methods for the enhancement of speech degraded by background noise, reverberation and competing speaker speech.

For speech degraded by background noise, a combined TSP method is proposed by emphasizing high signal to noise ratio (SNR) regions in the temporal domain, and eliminating the degradation and enhancing the speech-specific components in the spectral domain. The temporal domain processing is performed at two levels: gross and fine levels. In gross level processing, a method is proposed for detecting high SNR regions using the sum of the 10 largest peaks in the discrete Fourier transform (DFT) spectrum, the smoothed Hilbert envelope (HE) of the linear prediction (LP) residual and the modulation spectrum values. In fine level processing, a method is proposed to identify the instants of significant excitation from noisy speech. The proposed method involves following: (i) Sinusoidal analysis of noisy speech, (ii) Convolving the HE of the LP residual of the speech obtained from sinusoidal analysis by the first order Gaussian differentiator (FOGD). The gross and fine level features are combined to derive the weight function for the excitation source signal which emphasizes the excitation around the instants of significant excitation and deemphasizes the random peaks of background noise. Enhancement is done in the LP residual domain, because processing of LP residual produces less distortion than directly manipulating the original speech waveform. The temporally processed speech signal is further subjected to spectral processing which consists of two stages: attenuation of spectral characteristics of background noise and enhancement of speech-specific spectral features. In the first stage, the spectral characteristics of the background noise is estimated and attenuated using conventional spectral processing methods based on spectral subtraction or minimum mean square error (MMSE) estimators. In the second stage, the region around pitch and harmonics are enhanced by estimating pitch from the temporally processed speech. Finally, the performance of the combined TSP method is evaluated using three different composite objective quality measures: *signal distortion, noise distortion and overall quality*.

For speech degraded by reverberation, a combined TSP method is proposed that suppresses the effect of early and late reverberations. First, the late reverberant signal component is suppressed using spectral processing. It first estimates the power spectrum of late reverberation, and then subtracts it

from the power spectrum of the reverberant speech. Temporal processing is used to primarily attenuate the early reverberation. Temporal processing first identifies the high signal to reverberation ratio (SRR) region. Then, it enhances the instants of significant excitation within this high SRR region. The sum of 10 largest peaks in the DFT spectrum, the smoothed HE of the LP residual and the modulation spectrum values are used as indicators for identifying high SRR regions. Then, the instants of significant excitation are determined from the sum of bandpass filtered HE of the LP residual. A weight function is derived for the LP residual of spectrally processed speech to attenuate the reverberant peaks in the residual signal. Finally, the enhanced residual signal and the vocal-tract system characteristics derived from spectrally processed speech are used for synthesizing enhanced speech. The performance of the combined TSP method is evaluated using objective quality measures *Segmental SRR (SegSRR) and log spectral distance (LSD)*. In this study the proposed temporal processing method is also compared with the conventional LP residual based temporal processing method. The objective quality measures showed that the proposed temporal processing method provides improved performance than the conventional one.

A combined TSP method is proposed for separating speech of individual speakers from the mixture of two speaker speech signals, collected over a pair of microphones. The time delay in the arrival of speech of each speaker at a pair of microphones is exploited for speech separation. The mixed speech signals are first subjected to temporal processing. In temporal processing, speech of each speaker is enhanced with respect to the other by relatively emphasizing the speech around the instants of significant excitation of desired speaker by deriving speaker-specific weight function. To further improve the separation, the temporally processed speech is subjected to spectral processing. This involves enhancing the regions around the pitch and harmonic peaks of short time spectra computed from the temporally processed speech. To do so the pitch estimate is obtained from the temporally processed speech. Lastly, the performance of the proposed method is evaluated using (i) objective quality measures: *percentage of energy loss*, *percentage of noise residue and SNR gain*, and (ii) subjective quality measure: *mean opinion score (MOS)*.

An experimental evaluation is made to evaluate the performance of the combined TSP methods in speaker recognition (speaker identification) under degraded condition. For this purpose, the speaker recognition system is developed using the *universal background model (UBM) - gaussian mixture model (GMM)* concept with Mel-frequency cepstral coefficients (MFCC) as speaker-specific features.

For evaluation, testing is carried out with (i) clean speech, (ii) degraded speech (i.e., noisy speech or reverberant speech or two speaker speech), (iii) degraded speech preprocessed by temporal processing, (iv) degraded speech preprocessed by spectral processing and (v) degraded speech preprocessed by combined TSP.

In all the studies it is found that the combined TSP provides improved performance compared to temporal or spectral processing alone.

7.2 Contributions of the Present Work

The important contribution of the research work reported in this thesis is the development of combined TSP methods for speech enhancement. These include,

(i) Combined TSP for enhancement of noisy speech.

(ii) Combined TSP for enhancement of reverberant speech.

(iii) Combined TSP for enhancement of two speaker speech.

(iv) Evaluation of combined TSP methods in speaker recognition under degraded conditions.

While developing these methods the other contributions of the thesis are as follows:

(i) Set of features are proposed for gross level detection of speech regions in noisy, reverberant and two speaker speech.

(ii) Method to determine the instants of significant excitation in noisy speech is proposed.

(iii) Method is proposed to determine the instants of significant excitation in reverberant speech.

(iv) Pitch estimation method for two speaker speech is proposed using time delay estimation.

(v) A new spectral enhancement method is proposed in spectral processing.

7.3 Suggestions for Future Research

In this section we will provide some suggestions for further research.

(i) The performance of the proposed gross level detection may be increased by adaptively updating the thresholds or by combining the proposed features with existing voice activity detection features. For this thorough assessment of the strengths and weaknesses of the individual parameters needs to be done to compare the performance with the existing voice activity detection methods.

(ii) The LP coefficients derived from the noisy speech signal not accurate. Therefore, an iterative Wiener filter approach [10] can also be exploited to derive cleaner LP coefficients and subsequently a better residual. In this approach it may be viewed as spectral processing followed by the temporal processing. Since in the first stage itself vocal-tract characteristics will be modified. Therefore there is no need for further spectral processing.

(iii) To demonstrate the significance of temporal and spectral processing for enhancement of reverberant speech, the present work uses a simple spectral subtraction method. The performance of spectral processing can be further improved by exploiting statistical model based methods to estimate the late reverberant speech spectrum. The perceptual quality of the enhanced speech may also be improved by incorporating the proposed spectral enhancement technique.

(iv) In the case of two speaker separation, even though proposed method was illustrated using speech data for two speakers only, the method can be extended for enhancing speech of desired speaker from multi-speaker (more than two speakers) data also. In such a case first the number of speakers and corresponding delay values can be obtained by estimating time delays. It is also possible to obtain additional improvement in signal enhancement from multi-speaker data, if speech data is collected from a number (more than two) of spatially distributed microphones.

(v) In two speaker separation study the speakers are assumed to be stationary during recording sessions. However in practice there may be some movement of speakers. In such a case variation in the time delays must be computed as a function of time. For this case, methods can also be developed for estimating time-delays by combining excitation source and spectral based methods to improve the performance.

(vi) In this work no explicit study is made for processing unvoiced regions of degraded speech. Therefore a rigorous analysis of unvoiced regions can be done both in temporal and spectral processing. In particular, methods need to be developed for (i) identification of unvoiced sounds, (ii) defining and identification of instants of significant excitation for unvoiced sound, and (iii) identification of speech-specific spectral features of unvoiced sounds.

(vii) A robust method for determining instants of significant excitation for noisy speech can be developed by integrating (i) sinusoidal analysis, (ii) temporal larynx cycle averaging, and (iii) bandpass filtering the HE of the LP residual.

(viii) A robust method for determining instants of significant excitation for reverberant speech can be developed by incorporating temporal larynx cycle averaging in the proposed bandpass filtering approach.

(ix) The combined TSP method of noisy and reverberant speech can be extended to multi-microphone case.

(x) An efficient and simple pitch estimation and tracking algorithm for multi-speaker speech, in particular more than two speakers, can be developed using time delay estimation.

(xi) The proposed spectral enhancement method can be further refined and also be extended to unvoiced regions.

(xii) The combined TSP method needs to be developed when the speech contains all three types of degradations.

(xiii) In practical conditions, methods can be developed to identify the type of degradation and also the level of degradation. Based on this, the temporal weight function can be adaptively updated to improve the performance especially in high SNR/SRR conditions.

(xiv) In the present work existing temporal and spectral processing methods are used sequentially to obtain combined TSP method. The next step is to develop joint TSP methods where the degraded speech is processed in a parallel or simultaneous manner in both temporal and spectral domains.

A

MMSE-STSA Estimator

Contents

A.1 Derivation of the MMSE-STSA Estimator

Let $y(n) = s(n) + d(n)$ be the sampled noisy speech consisting of the clean speech $s(n)$ and the background noise $d(n)$. Taking the short-time Fourier transform of $y(n)$, we get [16]

$$Y(\omega_k) = S(\omega_k) + D(\omega_k). \tag{A.1}$$

In the case of discrete-time domain, $\omega_k = 2\pi k/N$ where $k = 0, 1, 2, ..., N-1$, and N is the number of points used for computing DFT. The above equation is expressed in polar form as [16]:

$$Y_k e^{j\theta_y(k)} = S_k e^{j\theta_s(k)} + D_k e^{j\theta_d(k)}. \tag{A.2}$$

where Y_k, S_k and D_k denote, respectively, the magnitudes of the noisy speech, clean speech and background noise spectra, and $\theta_y(k)$, $\theta_s(k)$ and $\theta_d(k)$ are their corresponding phase components. Since the spectral components are assumed to be statistically independent, the MMSE amplitude estimator $\hat{S}_k$ is derived from $Y(\omega_k)$ [144]. That is,

$$\hat{S}_k = E\left\{S_k|Y(\omega_k)\right\} = \int_0^\infty S_k p\left(S_k|Y(\omega_k)\right) dS_k \tag{A.3}$$

$$= \frac{\int_0^\infty \int_0^{2\pi} S_k p\left(Y(\omega_k)|S_k,\theta_k\right) p\left(S_k|\theta_k\right) dS_k d\theta_k}{\int_0^\infty \int_0^{2\pi} p\left(Y(\omega_k)|S_k,\theta_k\right) p\left(S_k|\theta_k\right) dS_k d\theta_k} \tag{A.4}$$

where $\theta_k = \theta_s(k)$, $E\{.\}$ is the expectation operator, and $p(.)$ is the Probability Density Function (PDF). If both noise and speech spectra are modelled as a Gaussian distribution, the conditional PDF, $p\left(Y(\omega_k)|S_k,\theta_k\right)$ is written as [16]

$$p\left(Y(\omega_k)|S_k,\theta_k\right) = \frac{1}{\pi\sigma_d^2} \exp\left\{-\frac{1}{\sigma_d^2}\left|Y_k - S_k e^{j\theta_s(k)}\right|^2\right\} \tag{A.5}$$

where $\sigma_d^2 \triangleq E\left\{|D_k|^2\right\}$ is the variance of the k^{th} spectral component of background noise.

It is known that, for complex Gaussian random variables with zero mean, their amplitude and phase components are statistically independent [16]. Therefore,

$$p\left(S_k|\theta_k\right) = p(S_k)p(\theta_k) \tag{A.6}$$

where $p\left(S_k\right)$ is a Rayleigh density with

$$p\left(S_k\right) = \frac{2S_k}{\sigma_s^2} \exp\left\{-\frac{S_k^2}{\sigma_s^2}\right\} \tag{A.7}$$

where $\sigma_s^2 \triangleq E\left\{|S_k|^2\right\}$ is the variances of the k^{th} spectral component of speech and $p\left(\theta_k\right)$ is a uniform density with

$$p\left(\theta_k\right) = \frac{1}{2\pi}. \tag{A.8}$$

Substituting Eqns. (A.5) and (A.6) into Eqn. (A.4) gives

$$\hat{S}_k = \frac{\int_0^\infty S_k^2 \exp\left\{-\frac{1}{\sigma_d^2}\left(Y_k^2 + \frac{\sigma_s^2+\sigma_d^2}{\sigma_d^2}\right)\right\} I_0\left(\frac{2S_kY_k}{\sigma_d^2}\right) dS_k}{\int_0^\infty S_k \exp\left\{-\frac{1}{\sigma_d^2}\left(Y_k^2 + \frac{\sigma_s^2+\sigma_d^2}{\sigma_d^2}\right)\right\} I_0\left(\frac{2S_kY_k}{\sigma_d^2}\right) dS_k} \tag{A.9}$$

With some mathematical manipulation, the estimator is expressed in the following form [14]

$$\hat{S}_k = \Gamma(1.5)\frac{\sqrt{\nu_k}}{\gamma_k} \exp\left(-\frac{\nu_k}{2}\right)\left[(1+\nu_k)I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right] Y_k \tag{A.10}$$

where $\Gamma(.)$ is the Gamma function (with $\Gamma(1.5) = \sqrt{\pi}/2$) and $I_0(.)$ and $I_1(.)$ are the zeroth and first order modified Bessel functions, respectively, defined as

$$I_n(z) \triangleq \frac{1}{2\pi}\int_0^{2\pi} \cos(\beta n)\exp(z\cos\beta)d\beta. \tag{A.11}$$

In Eqn. (A.10), ν_k is defined as [14]

$$\nu_k = \frac{\xi_k}{1+\xi_k}\gamma_k \tag{A.12}$$

where

$$\xi_k = \frac{E\left\{|S_k|^2\right\}}{E\left\{|D_k|^2\right\}} \tag{A.13}$$

$$\gamma_k = \frac{E\left\{|Y_k|^2\right\}}{E\left\{|D_k|^2\right\}} \tag{A.14}$$

where the terms ξ_k and γ_k are referred as *a priori* SNR and *a posteriori* SNR, respectively [14].

B

Linear Prediction Analysis of Speech

Contents

B.1 Basic Principles of Linear Predictive Analysis of Speech

Linear prediction (LP) is one of the most important tools in speech analysis. The philosophy behind linear prediction is that a speech sample can be approximated as a linear combination of past samples. Then, by minimizing the sum of the squared differences between the actual speech samples and the linearly predicted ones over a finite interval, a unique set of predictor coefficients can be determined [378, 379]. LP analysis decomposes the speech into two highly independent components, the vocal tract parameters (LP coefficients) and the glottal excitation (LP residual). It is assumed that speech is produced by exciting a linear time-varying filter (the vocal tract) by random noise for unvoiced speech segments, or a train of pulses for voiced speech. Fig. B.1 shows a model of speech production for LP analysis [6]. It consists of a time varying filter $H(z)$ which is excited by either a quasi periodic or a random noise source.

The most general predictor form in linear prediction is the autoregressive moving average (ARMA) model where the speech sample $s(n)$ is modelled as a linear combination of the past outputs and the present and past inputs [267, 337, 380]. It can be written mathematically as follows

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + G\sum_{l=0}^{q} b_l u(n-l), \quad b_0 = 1 \tag{B.1}$$

where $a_k, 1 \leq k \leq p, b_l, 1 \leq l \leq q$ and gain G are the parameters of the filter. Equivalently, in frequency domain, the transfer function of the linear prediction speech model is [337]

$$H(z) = \frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{1 + \sum_{k=1}^{p} a_k z^{-k}}. \tag{B.2}$$

$H(z)$ is referred to as a pole-zero model. The zeros represent the nasals and the poles represent the resonances (formants) of the vocal-tract. When $a_k = 0$ for $1 \leq k \leq p$, $H(z)$ becomes an all-zero or moving average (MA) model. Conversely, when $b_l = 0$ for $1 \leq l \leq q$, $H(z)$ becomes an an all-pole or autoregressive (AR) model [267]. For non-nasal voiced speech sounds the transfer function of the vocal-tract has no zeros whereas the nasals and unvoiced sounds usually includes the poles (resonances) as well as zeros (anti resonances) [6].

Generally the all-pole model is preferred for most applications because it is computationally more efficient and its the acoustic tube model for speech production. It can model sounds such as vowels well enough. The zeros arise only in nasals and in unvoiced sounds like fricatives. These zeros are

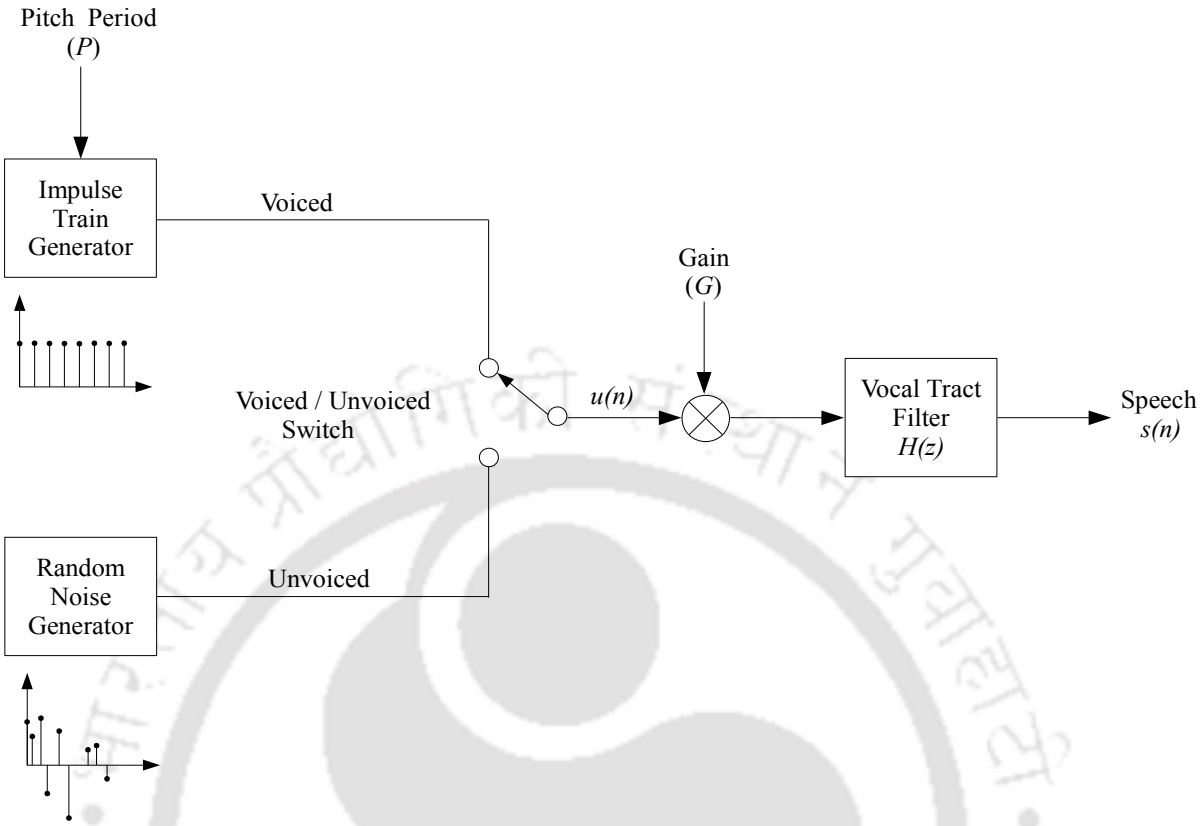


Figure B.1: Model of speech production for LP analysis.

approximately modelled by including more poles [337]. In addition, the location of a poles considerably more important perceptually than the location of a zero [381]. Moreover, it is easy to solve an all-pole model. To solve a pole-zero model, it is necessary to solve a set of nonlinear equations, but in the case of an all-pole model, only a set of linear equations need to be solved. The transfer function of the all-pole model is [267]

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}. \tag{B.3}$$

The number p implies that the past p output samples are being considered, which is also the order of the linear prediction. With this transfer function, we get a difference equation for synthesizing the speech samples $s(n)$ as

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{B.4}$$

where the coefficients a_k's are known as linear predictive coefficients (LPCs) and p is the order of the LP filter. It should be selected such that there is at least pair of poles per each formant. Generally,

the prediction order is chosen using the relation [382]

$$p = 2 \times (BW + 1) \tag{B.5}$$

where BW is the speech bandwidth in kHz.

B.2 The Prediction Error Signal

The error signal or the residual signal $e(n)$ is the difference between the input speech and the estimated speech [267].

$$e(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k). \tag{B.6}$$

Here the gain G is usually ignored to allow the parameterizations to be independent of the signal intensity. In z-domain $e(n)$ can be viewed as the output of the prediction filter $A(z)$ to the input speech signal $s(n)$ which is expressed as

$$E(z) = A(z)S(z) \tag{B.7}$$

where

$$A(z) = \frac{1}{H(z)} = 1 + \sum_{k=1}^{p} a_k z^{-k}; \quad G = 1. \tag{B.8}$$

The LP residual represents the excitations for production of speech [268]. The residual is typically a series of pulses, when derived from voiced speech or noise-like, when derived from unvoiced speech. The whole LP model can be decomposed into the two parts, the analysis part and the synthesis part as shown in Fig. B.2. The LP analysis filter removes the formant structure of the speech signal and leaves a lower energy output prediction error which is often called the LP residual or excitation signal. The synthesis part takes the error signal as an input [267]. The input is filtered by the synthesis filter $1/A(z)$, and the output is the speech signal.

B.3 Estimation of Linear Prediction Coefficients

There are two widely used methods for estimating the LP coefficients (LPCs): (i) Autocorrelation and (ii) Covariance. Both methods choose the short term filter coefficients (LPCs) a_k in such a way that the energy in the error signal (residual) is minimized. For speech processing tasks, the autocorrelation method is almost exclusively used because of its computational efficiency and inherent stability whereas the covariance method does not guarantee the stability of the all-pole LP synthesis

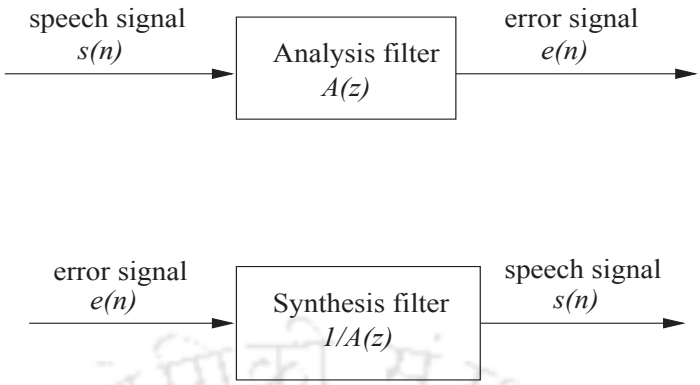


Figure B.2: LP analysis and synthesis model.

filter [6, 351, 383]. The autocorrelation method of computing LPCs is described below:

First, speech signal $s(n)$ is multiplied by a window $w(n)$ to get the windowed speech segment $s_w(n)$. Normally, a Hamming or Hanning window is used. The windowed speech signal is expressed as

$$s_w(n) = s(n)w(n). \tag{B.9}$$

The next step is to minimize the energy in the residual signal. The residual energy E_p is defined as [267]

$$E_p = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left(s_w(n) + \sum_{k=1}^{p} a_k s_w(n-k) \right)^2. \tag{B.10}$$

The values of a_k that minimize E_p are found by by setting the partial derivatives of the energy E_p with respect to the LPC parameters equal to zero.

$$\frac{\partial E_p}{\partial a_k} = 0, \quad 1 \leq k \leq p. \tag{B.11}$$

This results in the following p linear equations for the p unknown parameters $a_1, \ldots, a_p$

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n-k) = -\sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n), \quad 1 \leq i \leq p. \tag{B.12}$$

This linear equations can be expressed in terms of the autocorrelation function. This is because the autocorrelation function of the windowed segment $s_w(n)$ is defined as

$$R_s(i) = \sum_{n=-\infty}^{\infty} s_w(n)s_w(n+i), \quad 1 \leq i \leq p. \tag{B.13}$$

Exploiting the fact that the autocorrelation function is an even function i.e., $R_s(i) = R_s(-i)$. By substituting the values from Eqn. (B.13) in Eqn. (B.12), we get

$$\sum_{k=1}^{p} R_s\left(|i-k|\right) a_k = -R_s(i), \qquad 1 \leq i \leq p. \tag{B.14}$$

These set of p linear equations can be represented in the following matrix form as [6]

$$\begin{bmatrix} R_s(0) & R_s(1) & \cdots & R_s(p-1) \\ R_s(1) & R_s(0) & \cdots & R_s(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_s(p-1) & R_s(p-2) & \cdots & R_s(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_s(1) \\ R_s(2) \\ \vdots \\ R_s(p) \end{bmatrix} \tag{B.15}$$

This can be summarized using vector-matrix notation as

$$\mathbf{R_s a} = -\mathbf{r_s} \tag{B.16}$$

where the $p \times p$ matrix $\mathbf{R_s}$ is known as the autocorrelation matrix. The resulting matrix is a Toeplitz matrix where all elements along a given diagonal are equal. This allows the linear equations to be solved by the Levinson-Durbin algorithm. Because of the Toeplitz structure of $\mathbf{R_s}$, $A(z)$ is minimum phase [6]. At the synthesis filter $H(z) = 1/A(z)$, the zeros of $A(z)$ become the poles of $H(z)$. Thus, the minimum phase of $A(z)$ guarantees the stability of $H(z)$.

C

Sinusoidal Analysis and Synthesis of Speech

Contents

C.1 Sinusoidal Analysis System

A common model of speech production states that speech is the result of passing a glottal excitation waveform through a time-varying linear filter that models the resonant characteristics of the vocal-tract [384]. It is appropriate to assume that the excitation signal, in most of the cases, be in one of the two possible states, corresponding to voiced or unvoiced speech. In the sinusoidal speech model, the excitation signal is represented as the sum of a finite number of corresponding sinusoidal parameters at the pitch and harmonics during voiced speech regions, and is represented as numbers of corresponding sinusoidal parameters at peaks in the spectral domain during unvoiced speech regions [384]. Accordingly the input speech signal $s(n)$ is as [384]

$$s(n) = \sum_{l=1}^{L} A_l \cos\left(\omega_l n + \theta_l\right) \tag{C.1}$$

where L is the number of sinusoidal parameters, and A_l, ω_l and θ_l represent the time-varying amplitude, frequency, and phase of each sine wave.

C.2 Estimation of Speech Parameters using Sinusoidal Analysis

The sine wave parameters are estimated by applying short-time Fourier transform (STFT) to a quasi stationary part of the speech signal. The STFT of speech will have peaks occurring at all pitch harmonics and formants. Therefore the frequencies of underlying sine waves correspond to the peaks of STFT. The amplitudes and phases are estimated at peaks from the high resolution STFT using a simple peak picking algorithm [285, 384]. Generally a frame length of at least four times the longest expected pitch period is chosen to obtain sufficient spectral resolution [285, 384].

C.3 Sinusoidal Synthesis System

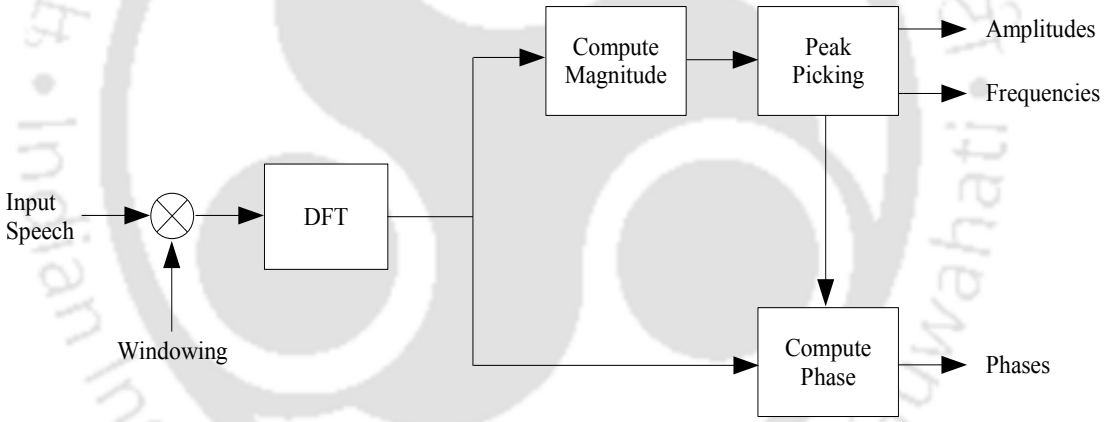
Synthesis in the sinusoidal model is achieved by overlap-add of the modeled segments. If the amplitudes, frequencies, and phases that are estimated for the k^{th} segment are denoted by A_l^k, ω_l^k, and θ_l^k respectively, the synthetic speech signal $\tilde{s}^k(n)$ can be represented as [384]

$$\tilde{s}^k(n) = \sum_{l=1}^{L^k} A_l^k \cos\left(\omega_l^k n + \theta_l^k\right) \tag{C.2}$$

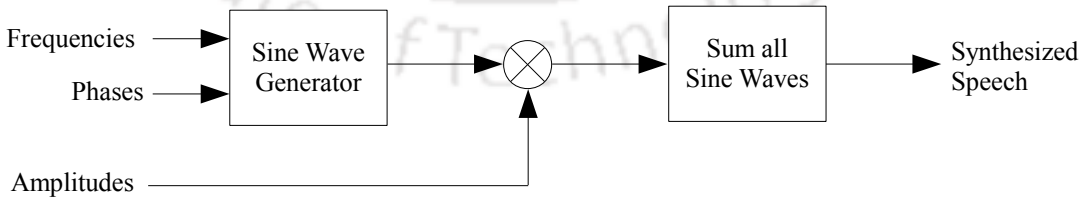
where L^k is the number of sinusoidal components in the frame.

Fig C.1 shows an analysis/synthesis system of the classical sinusoidal speech model [384]. In analysis system, first, the input speech signal is windowed and DFT is taken. Peak picking is then applied to the magnitude spectrum of the DFT to obtain a list of frequencies and corresponding amplitudes at those frequencies. Then, in the synthesis system, synthetic speech signal is reconstructed with these sinusoidal parameters.

Fig. C.2(a) shows a frame of voiced portion of the clean speech signal and the corresponding DFT log magnitude spectrum is given in Fig. C.2(b), from which the sinusoids are estimated. The peaks in the DFT magnitude spectrum are indicated by an * symbol in Fig. C.2(b). Fig. C.2(c)-(e) shows the synthesized speech signals from the sinusoidal analysis by considering the different number of sinusoidal components such as 4, 8 and 16, respectively. The similar steps are followed for a frame of unvoiced portion of the speech shown in Fig. C.2(f).



(a) Sinusoidal analysis system



(b) Sinusoidal synthesis system

Figure C.1: Analysis/synthesis system of classical sinusoidal speech model.

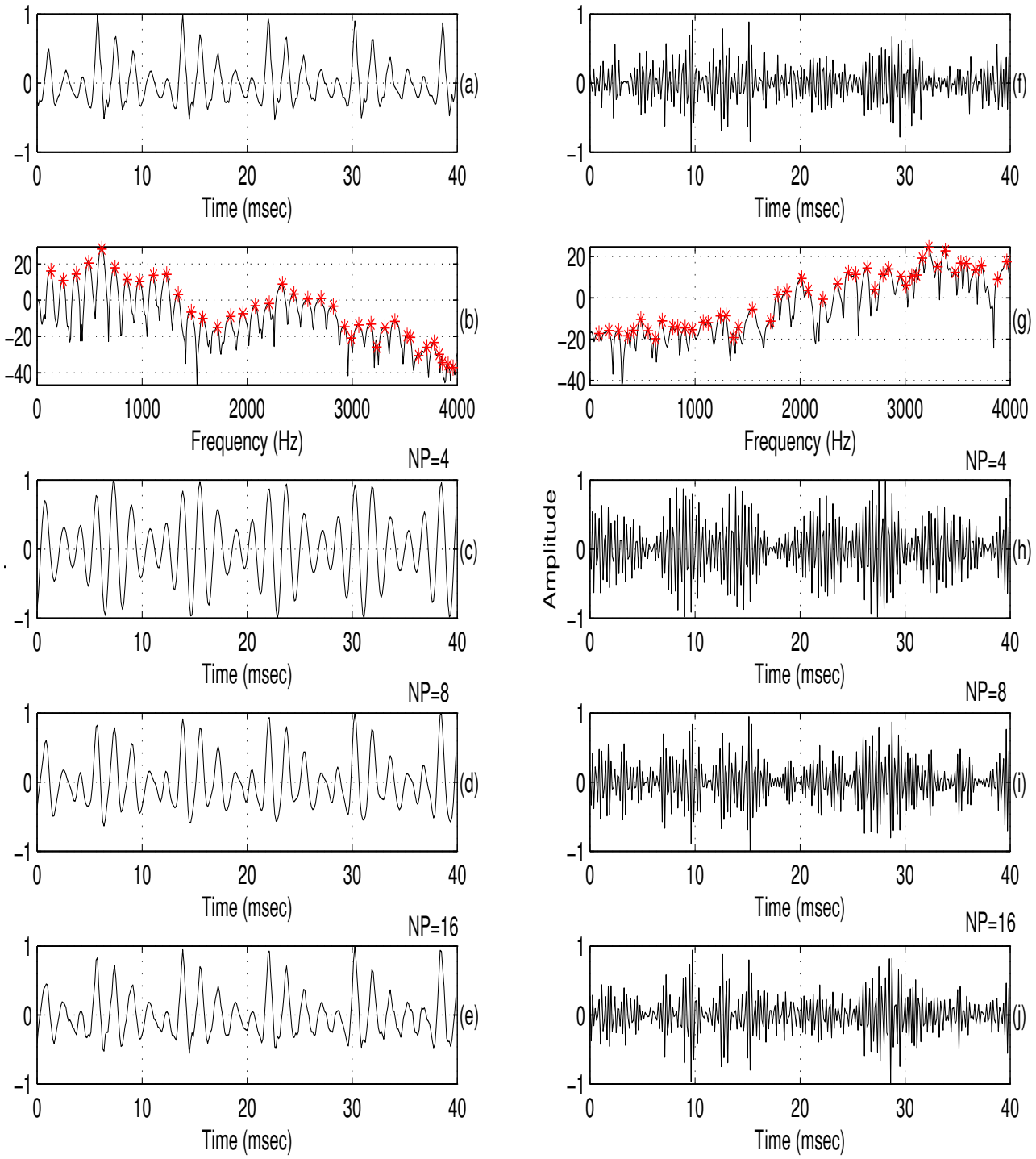


Figure C.2: Sinusoidal Analysis/synthesis (a) a frame of voiced speech, (b) its log magnitude spectrum, (c)-(e) synthesized speech signals by considering 4, 8 and 16 largest peaks, (f) a frame of unvoiced speech, (g) its log magnitude spectrum, (h)-(j) synthesized speech signals by considering 4, 8 and 16 largest peaks. In figure NP represents the number of largest peaks (sinusoidal components) considered for synthesizing the speech signal.

D

Composite Objective Quality Measures

Contents

D.1 Speech Quality Measures

In speech signal processing the quality of speech enhancement algorithms is validated using the subjective and objective distortion measures. The two main aspects of human perception of speech are speech intelligibility and quality. The quality of speech signals is a subjective measure which reflects the way the signal is perceived by listeners. It can be expressed in terms of how pleasant the signal sounds or how much effort is required on behalf of the listeners in order to understand the message. Intelligibility, on the other hand, is an objective measure of the amount of information which can be extracted from the listeners from the given signal, whether the signal is clean or noisy. A given signal may be of high quality but low intelligibility, and vice versa [3–5, 385]. Hence the two measures are independent of each other.

Subjective distortion measures are based on the opinion of a listener or a group of listeners. These measures are time-consuming and costly to obtain, requiring a set of discriminating listeners. In addition, a consistent listening environment is required since the perceived distortion can vary with such factors as the playback volume and type of listening instrument used [386]. However, subjective distortion measures provide the most accurate assessment of the performance since the degree of perceptual quality and intelligibility is determined by the human auditory system.

Objective quality measures are based on a mathematical comparison of the original and processed (enhanced) speech signals. It can be evaluated automatically from the speech signal, its spectrum or some parameters obtained thereof. Since they do not require listening tests, these measures can give an immediate estimate of the perceptual quality of a speech enhancement algorithm. The two main factors in selecting an objective distortion measure are its performance and complexity. The performance of an objective distortion measure can be established by its correlation with a subjective distortion measure of the same features (quality or intelligibility) [166].

D.2 Composite Objective Quality Measures

As mentioned the most accurate method for evaluating speech quality is through subjective listening tests. Although subjective evaluation of speech enhancement algorithms is always accurate and preferable, it is time consuming and cost expensive. Another promising approach to estimating the subjective quality is the use of composite objective measures, which is the result of the evaluation of the relationship between subjective analysis and the single objective measures [294, 296, 387]. The

reason behind the use of composite measures is that different objective measures capture different characteristics of the distorted or enhanced signal, and therefore combining them in a linear or non-linear fashion can potentially yield a significant gains in correlations (i.e., correlation with subjective measures such as mean opinion scores (MOS)). Hu and Loizou [294] proposed one such composite objective measures to evaluate speech enhancement algorithms by combining the various objective quality measures. The composite measure evaluates the quality of the speech by three different measures called signal distortion (C_{sig}), noise distortion (C_{bak}) and overall quality (C_{ovl}) [294]. These values are in between 1 and 5. The higher value of C_{sig}, C_{bak} and C_{ovl} represents lower signal distortion, lower background intrusiveness and higher quality of speech, respectively. These values are obtained by linearly combining the existing objective measures by the following relations [294]

$$C_{sig} = 3.093 - 1.029LLR + 0.603PESQ - 0.009WSS \tag{D.1}$$

$$C_{bak} = 1.634 + 0.478PESQ - 0.007WSS + 0.063segSNR \tag{D.2}$$

$$C_{ovl} = 1.594 + 0.805PESQ - 0.512LLR - 0.007WSS \tag{D.3}$$

where LLR, $PESQ$, WSS and $segSNR$ represents the log likelihood ratio, perceptual evaluation of speech quality, weighted slope spectral distance and segmental SNR, respectively. The coefficients of the linear equations are determined by computing the correlation coefficient between the subjective quality measure (MOS) and the objective quality measure [294]. A brief description of the individual objective quality measures used in Eqns. (D.1) - (D.3) is given below.

D.2.1 Segmental Signal-to-Noise Ratio (SegSNR)

The SNR is the ratio of signal energy to noise energy expressed in decibels dB and is given by [388]

$$SNR_{dB} = 10\log_{10}\frac{\sum\limits_{n} s^2(n)}{\sum\limits_{n}\left[s(n) - \hat{s}(n)\right]^2} \tag{D.4}$$

where $s(n)$ is the clean speech and $\hat{s}(n)$ is the degraded speech or enhanced speech signal. However, mathematically simple, the SNR measure carries with it the drawback of being a poor estimator of subjective quality. This is because SNR is not particularly well related to any subjective attribute of speech quality and weights all time domain errors in the speech waveform equally. A high SNR value, is thus, not necessarily indicative of good perceptual quality of the speech [2].

The speech energy in general is time varying. If we assume that noise distortion is broadband with

little energy fluctuation, then SNR measures should vary on a frame by fame basis. A much improved quality measure can be obtained if SNR is measured over short frames and the results averaged. The Segmental SNR is such frame based SNR and is estimated as follows [389]

$$SegSNR_{dB} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\left[\frac{\sum_{n=0}^{L-1}|s(n+mL)|^2}{\sum_{n=0}^{L-1}|s(n+mL)-\hat{s}(n+mL)|^2}\right] \tag{D.5}$$

where $s(n)$ is the clean speech signal and $\hat{s}(n)$ is the degraded speech or enhanced speech signal. M represents the number of frames, and L is the number of samples per frame. Generally only frames with segmental SNR in the range of -10dB to 35dB are considered in the average [389].

D.2.2 Log-Likelihood Ratio (LLR)

The Log-Likelihood Ratio measure is also referred to as the Itakura distance measure [389]. LLR measure is based on the dissimilarity between the all pole models of the reference and enhanced speech [390]. This distance measure is computed between sets of linear predictive (LP) parameters over synchronous frames in the original and enhanced speech. This measure is heavily influenced by spectral dissimilarity due to mismatch in formant locations whereas the locations of spectral valleys do not heavily contribute to the distance. This is desirable, since the auditory system is more sensitive to errors in formant location and bandwidth than to the spectral valleys between peaks [2]. The LLR measure is found as [389]

$$LLR = \log_{10}\left[\frac{a_x R_x a_x^T}{a_y R_y a_y^T}\right] \tag{D.6}$$

where a_x and a_y are the LP coefficient vectors for the degraded or enhanced and clean speech segments, respectively. R_x and R_y are the autocorrelation matrices of the degraded or enhanced and clean speech segments, respectively.

D.2.3 Weighted Spectral Slope (WSS) Measure

The weighted spectral slope measure proposed by Klatt is based on critical filter band analysis (auditory model) in which 36 overlapping filters of progressively larger bandwidth are used to estimate the smoothed short time speech spectrum and is therefore more closely related to the aspects of listener intelligibility [391]. This measure finds a weighted difference between the spectral slopes in each band. The magnitude of each weight reflects whether the band is near a spectral peak or valley and whether

the peak is the largest in the spectrum. A per frame WSS measure is in decibels found as [389]

$$WSS = K_{spl}(K - \hat{K}) + \sum_{k=1}^{36} W_a(k) \left[s(k) - \hat{s}(k)\right]^2 \tag{D.7}$$

where K and $\hat{K}$ are related to overall sound pressure level of the original and enhanced utterances, K_{spl} is a parameter which can be varied to increase overall performance, $W_a(k)$ is the weight of each band and $s(k)$ and $\hat{s}(k)$ are are the slopes in each critical band k for the original and degraded or enhanced speech respectively.

D.2.4 Perceptual Evaluation of Speech Quality (PESQ) Measure

The PESQ measure is proposed to predict the subjective opinion score of a degraded or enhanced speech. It is recommended by International Telecommunications Union for speech quality assessment [392]. The aim of this measure is to estimate the perceptual quality of narrow band voice codecs. PESQ is intended to address factors such as packet loss, variable delay, coding distortions and channel errors which are very poorly handled by conventional methods of comparison [393, 394].

In PESQ measure a reference signal and the processed signal are first aligned in both time and level. This is followed by a range of perceptually significant transforms which include Bark spectral analysis, frequency equalization, gain variation equalization and loudness mapping. After the two signals have undergo these transformations, two parameters (average disturbance value and average asymmetrical disturbance value) are computed [296]. These parameters are then combined in a mapping function to give an estimate of mean opinion score. For normal subjective test material the PESQ score ranges from 1.0 to 4.5, with higher score indicating better quality [392].

E

MFCC Feature Extraction

Contents

E.1 MFCC Feature Extraction

The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude and then warping the frequencies on a Mel scale, followed by applying the inverse DCT. The detailed description of various steps involved in the MFCC feature extraction is explained below.

(i) **Pre-emphasis:** Pre-emphasis refers to filtering that emphasizes the higher frequencies. Its purpose is to balance the spectrum of voiced sounds that have a steep roll-off in the high frequency region. For voiced sounds, the glottal source has an approximately -12 dB/octave slope [351]. However, when the acoustic energy radiates from the lips, this causes a roughly +6 dB/octave boost to the spectrum. As a result, a speech signal when recorded with a microphone from a distance has approximately a -6 dB/octave slope downward compared to the true spectrum of the vocal tract. Therefore, pre-emphasis removes some of the glottal effects from the vocal tract parameters. The most commonly used pre-emphasis filter is given by the following transfer function

$$H(z) = 1 - bz^{-1} \tag{E.1}$$

where the value of b controls the slope of the filter and is usually between 0.4 to 1.0 [351].

(ii) **Frame Blocking and Windowing:** The speech signal is a slowly time-varying or quasi-stationary signal. It means that when speech is examined over a sufficiently short period of time it has quite stable acoustic characteristics. Therefore, speech analysis must always be carried out on short segments across which the speech signal is assumed to be stationary. Short-term spectral measurements are typically carried out over 20 ms windows and advanced every 10 ms [2, 144]. Advancing the time window every 10 ms enables the temporal characteristics of individual speech sounds to be tracked and the 20 ms analysis window is usually sufficient to provide good spectral resolution of these sounds and at the same time short enough to resolve significant temporal characteristics. The purpose of the overlapping analysis is that each speech sound of the input sequence would be approximately centered at some frame. On each frame a window is applied to taper the signal towards the frame boundaries. Generally, Hanning or Hamming windows are used [351]. This is done to enhance the harmonics, smooth the edges and to reduce the edge effect while taking the DFT on the signal.

(iii) **DFT Spectrum:** Each windowed frame is converted into magnitude spectrum by applying DFT.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}; \qquad 0 \leq k \leq N-1 \tag{E.2}$$

where N is the number of points used to compute the DFT.

(iv) **Mel-Spectrum:** Mel-Spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as mel-filter bank. A mel is a unit of measure based on the human ears perceived frequency. It does not correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive pitch linearly. The mel scale is approximately a linear frequency spacing below 1 kHz, and a logarithmic spacing above 1 kHz [395]. The approximation of mel from physical frequency can be expressed as [2]

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{E.3}$$

where f denotes the physical frequency in Hz, and f_{mel} denotes the perceived frequency.

Filter banks can be implemented in both time domain and frequency domain. For MFCC computation, filter banks are generally implemented in frequency domain. The center frequencies of the filters are normally evenly spaced on the frequency axis. However, in order to mimic the human ears perception, the warped axis according to the non-linear function given in Eqn. (E.3), is implemented. The most commonly used filter shaper is triangular, and in some cases the Hanning filter can be found [351]. The triangular filter banks with mel-frequency warping is given in Fig. E.1.

The mel spectrum of the magnitude spectrum $X(k)$ is computed by multiplying the magnitude spectrum by each of the of the triangular mel weighting filters.

$$s(m) = \sum_{k=0}^{N-1} \Big[|X(k)|^2 H_m(k) \Big]; \qquad 0 \leq m \leq M-1 \tag{E.4}$$

where M is total number of triangular mel weighting filters [396,397]. $H_m(k)$ is the weight given

to the k^{th} energy spectrum bin contributing to the m^{th} output band and is expressed as :

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{E.5}$$

with m ranging from 0 to $M-1$.

(v) **Discrete Cosine Transform (DCT):** Since the vocal tract is smooth, the energy levels in adjacent bands tend to be correlated. The DCT is applied to the transformed mel frequency coefficients produces a set of cepstral coefficients. Prior to computing DCT the mel spectrum is usually represented on a log scale. This results in a signal in the cepstral domain with a que-frequency peak corresponding to the pitch of the signal and a number of formants representing low quefrequency peaks. Since most of the signal information is represented by the first few MFCC coefficients, the system can be made robust by extracting only those coefficients ignoring or truncating higher order DCT components [351]. Finally, MFCC is calculated as [351]

$$c(n) = \sum_{m=0}^{M-1} \log_{10}\left(s(m)\right) \cos\left(\frac{\pi n(m-0.5)}{M}\right); \qquad n = 0, 1, 2, ..., C-1 \tag{E.6}$$

where $c(n)$ are the cepstral coefficients and C is the number of MFCCs. Traditional MFCC systems use only 8 to 13 cepstral coefficients. The zeroth coefficient is often excluded since it represents the average log-energy of the input signal, which only carries little speaker-specific information.

(vi) **Dynamic MFCC Features:** The cepstral coefficients are usually referred to as static features, since they only contain information from a given frame. The extra information about the temporal dynamics of the signal is obtained by computing first and second derivatives of cepstral coefficients [1, 345, 398]. The first order derivative is called delta coefficients, and the second order derivative is called delta-delta coefficients. Delta coefficients tells about the speech rate, and the delta-delta coefficients gives an information similar to acceleration of speech. The

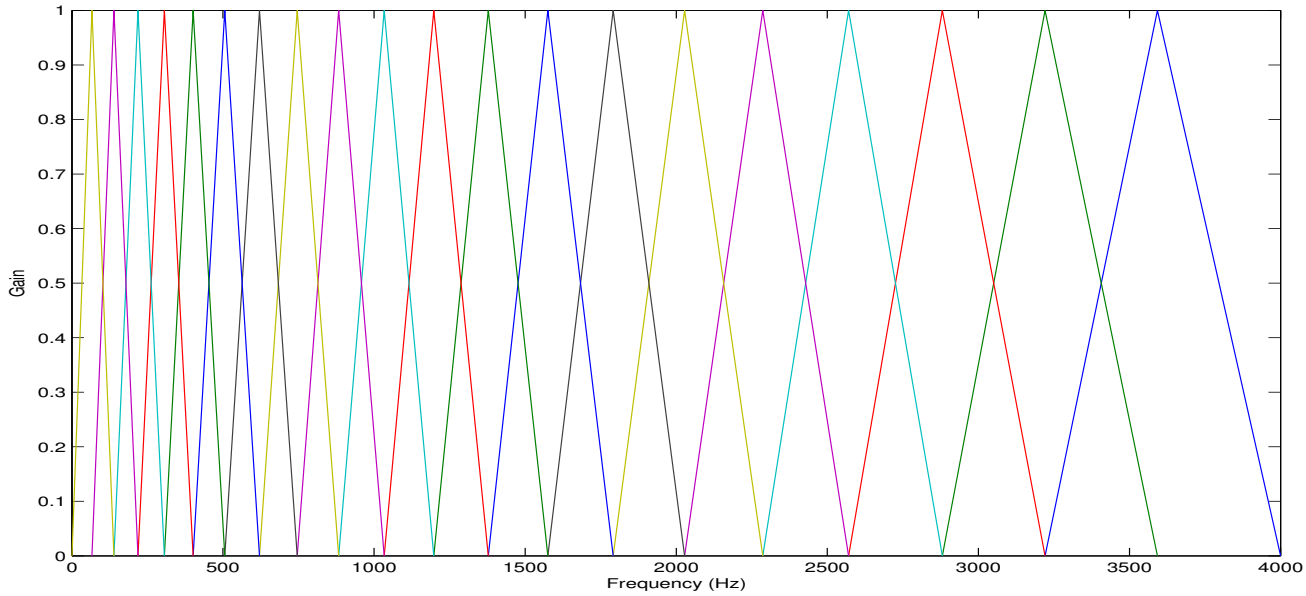


Figure E.1: Mel-filter bank

commonly used definition for computing dynamic parameter is [1]

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \tag{E.7}$$

where $c_m(n)$ denotes the m^{th} feature for the n^{th} time frame, k_i is the i^{th} weight and T is the number of successive frames used for computation. Generally T is taken as 2. The delta-delta coefficients are computed by taking the first order derivative of the delta coefficients.

F

Gaussian Mixture Models

Contents

F.1 Gaussian Mixture Model (GMM) Description

In the speech and speaker recognition the acoustic events are usually modeled by Gaussian probability density functions (PDFs), described by the mean vector and the covariance matrix. However unimodel PDF with only one mean and covariance are unsuitable to model all variations of a single event in speech signals. Therefore, a mixture of single densities is used to model the complex structure of the density probability. For a D-dimensional feature vector denoted as x_t, the mixture density for speaker Ω is defined as weighted sum of M component Gaussian densities as given by the following [329]

$$P(x_t|\Omega) = \sum_{i=1}^{M} w_i P_i(x_t) \tag{F.1}$$

where w_i are the weights and $P_i(x_t)$ are the component densities. Each component density is a D-variate Gaussian function of the form

$$P_i(x_t) = \frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(x_t-\mu_i)'\Sigma_i^{-1}(x_t-\mu_i)\right]} \tag{F.2}$$

where μ_i is a mean vector and Σ_i covariance matrix for i^{th} component. The mixture weights have to satisfy the constraint [329]

$$\sum_{i=1}^{M} w_i = 1. \tag{F.3}$$

The complete Gaussian mixture density is parameterized by the mean vector, the covariance matrix and the mixture weight from all component densities. These parameters are collectively represented by

$$\Omega = \{w_i, \mu_i, \Sigma_i\}\,; \qquad i = 1, 2, ...M. \tag{F.4}$$

F.2 Training the GMMs

To determine the model parameters of GMM of the speaker, the GMM has to be trained. In the training process, the maximum likelihood (ML) procedure is adopted to estimate model parameters. For a sequence of training vectors $X = \{x_1, x_2, .., x_T\}$, the GMM likelihood can be written as (assuming observations independence) [329]

$$P(X|\Omega) = \prod_{t=1}^{T} P(x_t|\Omega). \tag{F.5}$$

Usually this is done by taking the logarithm and is commonly named as log-likelihood function. From Eqns. (F.1) and (F.5), the log-likelihood function can be written as

$$\log\left[P(X|\Omega)\right] = \sum_{t=1}^{T} \log\left[\sum_{i=1}^{M} w_i P_i(x_t)\right]. \tag{F.6}$$

Often, the average log-likelihood is used value is used by dividing $\log\left[P(X|\Omega)\right]$ by T. This is done to normalize out duration effects from the log-likelihood value. Also, since the incorrect assumption of independence is underestimating the actual likelihood value with dependencies, scaling by T can be considered a rough compensation factor [376]. The parameters of a GMM model can be estimated using maximum likelihood (ML) estimation. The main objective of the ML estimation is to derive the optimum model parameters that can maximize the likelihood of GMM. The likelihood value is, however, a highly nonlinear function in the model parameters and direct maximization is not possible. Instead, maximization is done through iterative procedures. Of the many techniques developed to maximize the likelihood value, the most popular is the iterative expectation maximization (EM) algorithm [399].

F.2.1 Expectation Maximization (EM) Algorithm

The EM algorithm begins with an initial model Ω and tends to estimate a new model such that the likelihood of the model increasing with each iteration. This new model is considered to be an initial model in the next iteration and the entire process is repeated until a certain convergence threshold is obtained or a certain predetermined number of iterations have been made. A summary of the various steps followed in the EM algorithm are described below.

(i) **Initialization:** In this step an initial estimate of the parameters is obtained. The performance of the EM algorithm depends on this initialization. Generally, LBG [400] or K-means algorithm [401, 402] is used to initialize the GMM parameters.

(ii) **Likelihood Computation:** In each iteration the posterior probabilities for the i^{th} mixture is computed as [329]:

$$\Pr(i|x_t) = \frac{w_i P_i(x_t)}{\sum_{j=1}^{M} w_j P_j(x_t)}. \tag{F.7}$$

(iii) **Parameter Update:** Having the posterior probabilities, the model parameters are updated according to the following expressions [329].

Mixture weight update:

$$\overline{w_i} = \frac{\sum_{i=1}^{T} \Pr(i|x_t)}{T}. \tag{F.8}$$

Mean vector update:

$$\overline{\mu_i} = \frac{\sum_{i=1}^{T} \Pr(i|x_t) x_t}{\sum_{i=1}^{T} \Pr(i|x_t)}. \tag{F.9}$$

Covariance matrix update:

$$\overline{\sigma_i^2} = \frac{\sum_{i=1}^{T} \Pr(i|x_t) \left|x_t - \overline{\mu_i}\right|^2}{\sum_{i=1}^{T} \Pr(i|x_t)}. \tag{F.10}$$

In the estimation of the model parameters, it is possible to choose, either full covariance matrices or diagonal covariance matrices. It is more common to use diagonal covariance matrices for GMM, since linear combination of diagonal covariance Gaussians has the same model capability with full matrices [403]. Another reason is that speech utterances are usually parameterized with cepstral features. Cepstral features are more compactable, discriminative, and most important, they are nearly uncorrelated, which allows diagonal covariance to be used by the GMMs [329, 368]. The iterative process is normally carried out 10 times, at which point the model is assumed to converge to a local maximum [329].

F.2.2 Maximum *a posteriori* (MAP) Adaptation

Gaussian mixture models for a speaker can be trained using the modeling described earlier. For this, it is necessary that sufficient training data is available in order to create a model of the speaker. Another way of estimating a statistical model, which is especially useful when the training data available is of short duration, is by using maximum *a posteriori* adaptation (MAP) of a background model trained on the speech data of several other speakers [404]. This background model is a large GMM that is trained with a large amount of data which encompasses the different kinds of speech that may be encountered by the system during training. These different kinds may include different channel conditions, composition of speakers, acoustic conditions, etc. A summary of MAP adaptation steps are given below.

For each mixture i from the background model, $Pr(i|x_t)$ is calculated as [373]

$$\Pr(i|x_t) = \frac{w_i P_i(x_t)}{\sum_{j=1}^{M} w_j P_j(x_t)}. \tag{F.11}$$

Using $Pr(i|x_t)$, the statistics of the weight, mean and variance are calculated as follows [373]

$$n_i = \sum_{i=1}^{T} \Pr(i|x_t) \tag{F.12}$$

$$E_i(x_t) = \frac{\sum_{i=1}^{T} \Pr(i|x_t) x_t}{n_i} \tag{F.13}$$

$$E_i(x_t^2) = \frac{\sum_{i=1}^{T} \Pr(i|x_t) x_t^2}{n_i}. \tag{F.14}$$

These new statistics calculated from the training data are then used adapt the background model, and the new weights ($\hat{w}_i$), means ($\hat{\mu}_i$) and variances ($\hat{\sigma}_i^2$) are given by [373]

$$\hat{w}_i = \left[\frac{\alpha_i n_i}{T} + (1-\alpha_i) w_i\right] \gamma \tag{F.15}$$

$$\hat{\mu}_i = \alpha_i E_i(x_t) + (1-\alpha_i)\mu_i \tag{F.16}$$

$$\hat{\sigma}_i^2 = \alpha_i E_i(x_t^2) + (1-\alpha_i)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2. \tag{F.17}$$

A scale factor γ is used, which ensures that all the new mixture weights sum to 1. α_i is the adaptation coefficient which controls the balance between the old and new model parameter estimates. α_i is defined as [373]

$$\alpha_i = \frac{n_i}{n_i + r} \tag{F.18}$$

where r is a fixed relevance factor, which determines the extent of mixing of the old and new estimates of the parameters. Low values for α_i ($\alpha_i \to 0$), will result in new parameter estimates from the data to be de-emphasized, while higher values ($\alpha_i \to 1$) will emphasize the use of the new training data-dependent parameters. Generally only mean values are adapted [376]. It is experimentally shown that mean adaptation gives slightly higher performance than adapting all three parameters [373].

F.3 Testing

In identification phase, mixture densities are calculated for every feature vector for all speakers and speaker with maximum likelihood is selected as identified speaker. For example, if S speaker models $\{\Omega_{1,}\Omega_{2,..,}\Omega_S\}$ are available after the training, speaker identification can be done based on a new speech data set. First, the sequence of feature vectors $X = \{x_1, x_2, .., x_T\}$ is calculated. Then the speaker model $\hat{s}$ is determined which maximizes the a posteriori probability $P(\Omega_S|X)$. That is, according to the Bayes rule [329]

$$\hat{s} = \max_{1\leq s\leq S} P(\Omega_S|X) = \max_{1\leq s\leq S} \frac{P(X|\Omega_S)}{P(X)} P(\Omega_S). \tag{F.19}$$

Assuming equal probability of all speakers and the statistical independence of the observations, the decision rule for the most probable speaker can be redefined as

$$\hat{s} = \max_{1\leq s\leq S} \sum_{t=1}^{T} \log P(x_t|\Omega_s) \tag{F.20}$$

with T the number of feature vectors of the speech data set under test and $P(x_t|\Omega_s)$ given by Eqn. (F.1).

Decision in verification is obtained by comparing the score computed using the model for the claimed speaker Ω_S given by $P(\Omega_S|X)$ to a predefined threshold θ. The claim is accepted if $P(\Omega_S|X) > \theta$, and rejected otherwise [376].

Bibliography

[1] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[2] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete Time Processing of Speech Signals*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[3] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 122, no. 3, pp. 1777–1786, 2007.

[4] Y. Ephraim, H. Lev Ari, and W. J .J. Roberts, "A brief survey of speech enhancement," in *The Electronic Handbook*, 2nd ed. CRC Press, 2005.

[5] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook*. CRC Press, 2006, ch. 15, pp. 12–26.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, 1st ed. Pearson Education, 1978.

[7] C. Cherry and R. Wiley, "Speech communication in very noisy environments," *nature*, vol. 214, p. 1184, Jun. 1967.

[8] P. Satyanarayana, "Short segment analysis of speech for enhancement," Ph.D. dissertation, Indian Insititute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India, Feb. 1999. [Online]. Available: http://speech.cs.iitm.ernet.in/Main/publications/PhDTheses/MurthyThesis.ps.gz

[9] S. R. M. Prasanna, "Event based analysis of speech," Ph.D. dissertation, Indian Insititute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India, Mar. 2004. [Online]. Available: http://speech.cs.iitm.ernet.in/Main/publications/PhDTheses/prasannaPhdThesis.pdf

[10] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[11] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. Hyderabad, India: Universities Press (India) Pvt., Ltd., 2007.

[12] R. Munkong and B.-H. Juang, "Auditory perception and cognition," *IEEE Signal Process. Magazine*, vol. 25, no. 3, pp. 98–117, May 2008.

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

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

[15] ——, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[16] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL.: CRC, 2007.

[17] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. Satyanarayana Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, pp. 25–42, May 1999.

[18] B. Yegnanarayana, S. R. Mahadeva Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Orlando, USA, 2002, pp. I–541–I–544.

[19] E. Habets, "Single-and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, The Netherlands, Jun. 2007. [Online]. Available: http://alexandria.tue.nl/extra2/200710970.pdf

[20] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, pp. 774–784, May 2006.

[21] K. Lebart and J. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[22] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 267–281, May 2000.

[23] B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 1110–1118, Nov. 2005.

[24] T. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60, pp. 911–918, Oct. 1976.

[25] D. Morgan, E. George, L. Lee, and S. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, pp. 407–424, Sep. 1997.

[26] M. Portnoff, "Short-time fourier analysis of sampled speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 3, pp. 364–373, Jun. 1981.

[27] B. Yegnanarayana, S. R. M. Prasanna, and M. Mathew, "Enhancement of speech in multispeaker environment," in *Proc. European Conf. Speech Process., Technology*, Geneva, Switzerland, 2003, pp. 581–584.

[28] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*. Berlin, Germany: Springer-Verlag, 2005.

[29] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[30] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, pp. 679–681, Aug. 1982.

[31] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons, 2006.

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

[33] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1979, pp. 208–211.

[34] P. Vary, "Noise suppression by spectral magnitude estimation - mechanism and theoretical limits," *Signal Process.*, vol. 8, pp. 387–400, 1985.

[35] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, pp. 453–466, Jun. 2008.

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

[37] G. Kang and L. Fransen, "Quality improvement of LPC-processed noisy speech by using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-37, pp. 939–942, Jun. 1989.

[38] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215–228, 1992.

[39] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, USA, May 2002.

[40] P. Crozier, B. Cheetham, C. Holt, and E. Munday, "Speech enhancement employing spectral subtraction and linear predictive analysis," *Electronics Letters*, vol. 29, pp. 1094–1095, Jun. 1993.

[41] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. European Signal Process. Conference (EUSIPCO)*, Edinburgh, UK, 1994, pp. 1182–1185.

[42] Z. Goh, Kah-Chye Tan, and T. G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 287–292, May 1998.

[43] B. L. Sim, Y. C. Tong, J. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 328–337, Jul. 1998.

[44] J. W. Seok and K. S. Bae, "Reduction of musical noise in spectral subtraction method using subframe phase randomisation," *Electronics Letters*, vol. 35, pp. 123–125, Jan. 1999.

[45] W. Kim, S. Kang, and H. Ko, "Spectral subtraction based on phonetic dependency and masking effects," *IEE Proc. Vision, Image and Signal Process.*, vol. 147, no. 5, pp. 423–427, Oct. 2000.

[46] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.,*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[47] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 799–807, Nov. 2001.

[48] M. Li, H. McAllister, N. Black, and T. A. D. Perez, "Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids," *IEEE Trans. Biomedical Engineering*, vol. 48, no. 9, pp. 979–988, Sep. 2001.

[49] H. T. Hu, F. J. Kuo, and H. J. Wang, "Supplementary schemes to spectral subtraction for speech enhancement," *Speech Communication*, vol. 36, pp. 205–218, Mar. 2002.

[50] M. Hasan, S. Salahuddin, and M. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Process. Letters*, vol. 11, pp. 450–453, Apr. 2004.

[51] Y. Hu and P. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Process. Letters*, vol. 11, pp. 270–273, Feb. 2004.

[52] K. Yamashita and T. Shimamura, "Nonstationary noise estimation using low-frequency regions for spectral subtraction," *IEEE Signal Process. Letters*, vol. 12, pp. 465–468, Jun. 2005.

[53] L. P. Yang and Q. J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *J. Acoust. Soc. Am.*, vol. 117, pp. 1001–1004, Mar. 2005.

[54] N. W. D. Evans, J. S. Mason, W. M. Liu, and B. Fauve, "An assessment on the fundamental limitations of spectral subtraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Toulouse, France, May 2006, pp. I–145–I–148.

[55] H. T. Hu and C. Yu, "Adaptive noise spectral estimation for spectral subtraction speech enhancement," *IET Signal Process.*, vol. 1, no. 3, pp. 156–163, Sep. 2007.

[56] C.-T. Lu, "Reduction of musical residual noise for speech enhancement using masking properties and optimal smoothing," *Advances on Pattern Recognition Letters*, vol. 28, no. 11, pp. 1300–1306., Aug. 2007.

[57] P. Händel, "Power spectral density error analysis of spectral subtraction type of speech enhancement methods," *EURASIP J. Applied Signal Process.*, vol. 2007, Article ID 96384, 9 pages, no. 1.

[58] R. M. Udrea, N. D. Vizireanu, and S. Ciochina, "An improved spectral subtraction method for speech enhancement using a perceptual weighting filter," *Digital Signal Process.*, vol. 18, no. 4, pp. 581–587, Jul. 2008.

[59] Q. Zeng and W. H. Abdulla, "Speech enhancement by multichannel crosstalk resistant ANC and improved spectrum subtraction," *EURASIP J. Applied Signal Process.*, vol. 2006, Article ID 61214, 10 pages.

[60] R. M. Udrea, N. Vizireanu, S. Ciochina, and S. Halunga, "Nonlinear spectral subtraction method for colored noise reduction using multi-band bark scale," *Signal Process.*, vol. 88, no. 5, pp. 1299–1303, 2008.

[61] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Adaptive β-order generalized spectral subtraction for speech enhancement," *Signal Process.*, vol. 88, no. 11, pp. 2764–2776, 2008.

[62] K. Wojcicki, M. Milacic, A. Stark, J. Lyons, and K. Paliwal, "Exploiting conjugate symmetry of the short-time fourier spectrum for speech enhancement," *IEEE Signal Process. Letters*, vol. 15, pp. 461–464, 2008.

[63] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, pp. 497–514, Nov. 1997.

[64] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 126–137, Mar. 1999.

[65] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 59–67, Jan. 2004.

[66] A. Natarajan, J. H. L. Hansen, K. H. Arehart, and J. Rossi-Katz, "An auditory-masking-threshold-based noise suppression algorithm GMMSE-AMT [ERB] for listeners with sensorineural hearing loss," *EURASIP J. Applied Signal Process.*, vol. 2005, no. 18, pp. 2938-2953, 2005.

[67] H. Liu, Q. Zhao, M. Wan, and S. Wang, "Application of spectral subtraction method on enhancement of electrolarynx speech," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 398–406, 2006.

[68] M.-C. You, C.-Y. Mao, J.-S. Wang, and F.-C. Chuang, "A recursive parametric spectral subtraction algorithm for speech enhancement." in *Advanced Intelligent Computing Theories and Applications*, D.-S. Huang, L. Heutte, and M. Loog, Eds., vol. 2. Springer, 2007, pp. 826–835.

[69] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[70] B. J. Shannon, "Speech recognition and enhancement using autocorrelation domain processing," Ph.D. dissertation, School of engineering, Griffith University, Brisbane, Australia, Aug. 2006.

[71] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, "Nonlinear speech enhancement: An overview," in *Progress in Nonlinear Speech Processing*. Springer Berlin / Heidelberg, 2007, pp. 217–248.

[72] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Communication*, vol. 24, pp. 249–257, Jun. 1998.

[73] B. Chen and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Communication*, vol. 49, pp. 134–143, Feb. 2007.

[74] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. 11th IEEE Workshop on Statistical Signal Process.*, Orchid Country Club, Singapore, Aug. 2001, pp. 496–499.

[75] M.Marzinzik and B.Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 109–118, Feb. 2002.

[76] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 845–856, Sep. 2005.

[77] B. Chen and P. Loizou, "Speech enhancement using a MMSE short time spectral amplitude estimator with Laplacian speech modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Philadelphia, PA, USA, 2005, pp. 1097–1100.

[78] S. Gazor and W. Zhang, "Speech enhancement employing Laplacian-Gaussian mixture," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 896–904, Sep. 2005.

[79] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 345–349, Apr. 1994.

[80] I. Cohen, "On the decision-directed estimation approach of Ephraim and Malah," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Montreal, Quebec, Canada, May 2004, pp. I–293–I–296.

[81] ——, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Process. Letters*, vol. 11, no. 9, pp. 725–728, Sep. 2004.

[82] ——, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sep. 2005.

[83] ——, "Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation," *Speech Communication*, vol. 47, pp. 336–350, Nov. 2005.

[84] H. Tasmaz and E. Ercelebi, "Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE-STSA estimation in various noise environments," *Digital Signal Process.*, vol. 18, no. 5, pp. 797–812, Sep. 2008.

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

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

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

[88] C. T. Lu and H. C. Wang, "Enhancement of single channel speech based on masking property and wavelet transform," *Speech Communication*, vol. 41, pp. 409–427, Oct. 2003.

[89] S.-H. Chen and J.-F. Wang, "Speech enhancement using perceptual wavelet packet decomposition and teager energy operator," *J. VLSI Signal Process. System*, vol. 36, no. 2-3, pp. 125–139, 2004.

[90] C.-T. Lu and H.-C. Wang, "Speech enhancement using perceptually-constrained gain factors in critical-band-wavelet-packet transform," *Electronics Letters*, vol. 40, no. 6, pp. 394–396, Mar. 2004.

[91] ——, "Speech enhancement using robust weighting factors for critical-band-wavelet-packet transform," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Montreal, Quebec, Canada, May 2004, pp. I–721–I–724.

[92] S. Ayat, M. M. Shalmani, and R. Dianat, "An improved wavelet-based speech enhancement by using speech signal features," *Computers and Electrical Engineering*, vol. 32, no. 6, pp. 411–425, Nov. 2006.

[93] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on timescale adaptation," *Speech Communication*, vol. 48, pp. 1620–1637, Dec. 2006.

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

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

[96] C. T. Lu and H. C. Wang, "Speech enhancement using hybrid gain factor in critical-band-wavelet-packet transform," *Digital Signal Process.*, vol. 17, no. 1, pp. 172–188, Jan. 2007.

[97] J.-H. Chang, S. Gazor, N. S. Kim, and S. K. Mitra, "Multiple statistical models for soft decision in noisy speech enhancement," *Pattern Recognition*, vol. 40, pp. 1123–1134, Mar. 2007.

[98] S. Senapati, S. Chakroborty, and G. Saha, "Speech enhancement by joint statistical characterization in the Log Gabor Wavelet domain," *Speech Communication*, vol. 50, pp. 504–518, Jun. 2008.

[99] M. K. Hasan, S. Salahuddin, and M. R. Khan, "Reducing signal-bias from mad estimated noise level for dct speech enhancement," *Signal Process.*, vol. 84, no. 1, pp. 151–162, 2004.

[100] L. Rabiner and M. Samber, "An algorithm for determining the endpoints of isolated utterances,," *The Bell System Technical Journal*, vol. 54, pp. 297–315, 1975.

[101] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Letters*, vol. 6, pp. 1–3, Jan. 1999.

[102] S. G. Tanyer and H. Ozer, "Voice activity detection in non stationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 478–482, Jul. 2000.

[103] K.-H. Woo, T.-Y. Yang, K.-J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, pp. 180–181, Jan. 2000.

[104] L.-S. Huang and C.-H. Yang, "A novel approach to robust speech endpoint detection in car environments," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Process.*, vol. 3, 2000, pp. 1751–1754.

[105] Y. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Process.*, vol. 2, 2001, p. 711.

[106] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 217–231, Mar. 2001.

[107] Q. Li, J. Zheng, Q. Zhou, and C.-H. Lee, "A robust, real-time endpoint detector with energy normalization for ASR in adverse environments," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Process.* IEEE, 2001.

[108] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 109–118, Feb 2002.

[109] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 498–505, Sep. 2003.

[110] J.-H. Chang and N. S. Kim, "Voice activity detection based on complex laplacian model," *Electronics Letters*, vol. 39, no. 7, pp. 632–634, Apr. 2003.

[111] J.-H. Chang, J. Shin, and N. Kim, "Voice activity detector employing generalised Gaussian distribution," *Electronics Letters*, vol. 40, no. 24, pp. 1561–1563, Nov. 2004.

[112] K. Li, M. Swamy, and M. O. Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 965–974, Sep. 2005.

[113] J. Ramirez, J. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Letters*, vol. 12, no. 10, pp. 689–692, Oct. 2005.

[114] J.-H. Chang, N. S. Kim, and S. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[115] A. Martin and L. Mauuary, "Robust speech/non-speech detection based on LDA-derived parameter and voicing parameter for speech recognition in noisy environments," *Speech Communication*, vol. 48, pp. 191–206, Feb. 2006.

[116] J. W. Shin, J.-H. Chang, and N. S. Kim, "Voice activity detection based on a family of parametric distributions," *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1295–1299, 2007.

[117] J. Ramrez, J. M. Grriz, and J. C. Segura, *Voice Activity Detection: Fundamentals and Speech Recognition System Robustness.* Vienna, Austria: I-Tech Education and Publishing, 2007, pp. 1–22.

[118] ——, *New Advances in Voice Activity Detection using HOS and Optimization Strategies.* Vienna, Austria: I-Tech Education and Publishing, 2007, pp. 1–22.

[119] M. Pwint and F. Sattar, "Speech/nonspeech detection using minimal Walsh basis functions," *EURASIP J. Audio, Speech, and Music Process.*, vol. 2007, Article ID 39546, 9 pages.

[120] J. M. Gorriz, J. Ramirez, E. W. Lang, and C. G. Puntonet, "Jointly gaussian pdf-based likelihood ratio test for voice activity detection," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 8, pp. 1565–1578, Nov. 2008.

[121] A. Gersho, "Advances in speech and audio compression," *Proc. IEEE*, vol. 82, no. 6, pp. 900–918, Jun 1994.

[122] R. L. Bouquin-Jeanns and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Communication*, vol. 16, pp. 245–254, Apr. 1995.

[123] R. Martin, "An efficient algorithm to estimate the instantaneous SNR of speech signals," in *Proc. EUROSPEECH-93*, Berlin, Germany, Sep. 1993, pp. 1093–1096.

[124] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Letters*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[125] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[126] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Process.*, vol. 86, no. 6, pp. 1215–1229, 2006.

[127] Z. Lin, R. A. Goubran, and R. M. Dansereau, "Noise estimation using speech/non-speech frame decision and subband spectral tracking," *Speech Communication*, vol. 49, pp. 542–557, Aug. 2007.

[128] K. V. Srensen and S. V. Andersen, "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions," *EURASIP J. Applied Signal Process.*, vol. 2005, no. 18, pp. 2954-2964, 2005, no. 18.

[129] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 6, pp. 1112–1123, Aug. 2008.

[130] Y.-S. Park and J.-H. Chang, "A probabilistic combination method of minimum statistics and soft decision for robust noise power estimation in speech enhancement," *IEEE Signal Process. Letters*, vol. 15, pp. 95–98, 2008.

[131] W. Jin and M. S. Scordilis, "Speech enhancement by residual domain constrained optimization," *Speech Communication*, vol. 48, pp. 1349–1364, Oct. 2006.

[132] K. Hermus, P. Wambacq, and H. V. hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. Advances in Signal Process.*, vol. 2007, Article ID 45821, 15 pages.

[133] Y. Ephraim and H. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 251–266, Jul. 1995.

[134] J. Sun, J. Zhang, and M. Small, "Extension of the local subspace method to enhancement of speech with colored noise," *Signal Process.*, vol. 88, no. 7, pp. 1881–1888, 2008.

[135] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: a regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–67, 1991.

[136] S. Jensen, P. Hansen, S. Hansen, and J. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 439 – 448, Nov. 1995.

[137] F. Asano, S. Hayamizu, T. Yamada, and S.Nakamura, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 159–167, Mar. 2000.

[138] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 87–95, Feb. 2001.

[139] C. H. You, S. N. Koh, and S.Rahardja, "An invertible frequency eigendomain transformation for masking-based subspace speech enhancement," *IEEE Signal Process. Letters*, vol. 12, pp. 461–464, Jun. 2005.

[140] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 700–708, Nov. 2003.

[141] M. Kahrs and K. Brandenburg, Eds., *Applications of digital signal processing to audio and acoustics*. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[142] S. Neely and J. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, pp. 165–169, 1979.

[143] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of acoustics*, 2nd ed. New York: Wiley Eastern, 1982.

[144] J. Benesty, M. M. Sondhi, and Y. A. Huang, *Springer Handbook of Speech Processing*. Springer-Verlag New York, Inc., 2008.

[145] S. Hameed, J. Pakarinen, K. Valde, and V. Pulkki, "Psychoacoustic cues in room size perception," *Audio Engineering Society Convention paper 6084. Presented at the 116th Convention 2004*, May 8-11,Berlin, Germany. 2004.

[146] J. Sandvad, "Auditory perception of reverberant surroundings," *J. Acoust. Soc. Am.*, vol. 105, no. 2, p. 1193, Feb. 1999.

[147] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, 1999.

[148] A. J. Watkins, "Perceptual compensation for effects of echo and of reverberation on speech identification," *Acta Acustica united with Acustica*, vol. 91, pp. 892–901, 2005.

[149] A. Watkins and S. Makin, "Perceptual compensation for reverberation in speech identification: Effects of single-band, multiple-band and wideband contexts," *Acta Acustica united with Acustica*, vol. 93, pp. 403–410, 2007.

[150] B. Gillespie and L. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Orlando, USA, 2002, pp. I–557–I–560.

[151] J. Flanagan and R. Lummis, "Signal processing to reduce multipath distortion in small rooms," *J. Acoust. Soc. Am.*, vol. 47, pp. 1475–1481, 1970.

[152] J. Allen, D. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, pp. 912–915, 1977.

[153] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Hong Kong, China PR, Apr. 2003, pp. 92–95.

[154] E. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proc. 15th Annual Workshop on Circuits, Systems and Signal Process.*, Nov. 2004, pp. 250–254.

[155] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Mar. 2005, pp. 173–176.

[156] I. Tashev and D. Allred, "Reverberation reduction for improved speech recognition," in *Proc. Workshop on Hands-free Speech Communication*, Rutgers, USA., March 2005.

[157] F. S. Pacheco and R. Seara, "Spectral subtraction for reverberation reduction applied to automatic speech recognition," in *Proc. Int. Telecommunications Symposium*, Sep. 2006, pp. 795–800.

[158] E. A. P. Habets, S. Gannot, I. Cohen, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 8, pp. 1433–1451, Nov. 2008.

[159] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 534–545, May 2009.

[160] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound and Vibration*, vol. 102, pp. 217–228, 1985.

[161] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Toronto, Canada, Apr. 1991, pp. 977–980.

[162] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptivefiltering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 6, Salt Lake City, USA, 2001, pp. 3701–3704.

[163] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-36, pp. 145–152, Feb. 1988.

[164] S. Subramaniam, A. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 392–396, Sep. 1996.

[165] T. Yoshioka, T. Hikichi, and M. Miyoshi, "Dereverberation by using time-variant nature of speech production system," *EURASIP J. Advances in Signal Process.*, vol. 2007, Article ID 65698, 15 pages, 2007.

[166] H. Wang and F. Itakura, "Realization of acoustic inverse filtering through multi-microphone sub-band processing," *IEICE Trans. Fundamentals of Electronics, Communications and Computer*, vol. E75-A, no. 11, pp. 1474–1483, 1992.

[167] D. Cole, M. Moody, and S. Sridharan, "Intelligibility of reverberant speech enhanced by inversion of room response," in *IEEE Proc. Int. Sym. Speech, Image Processing and Neural Networks*, vol. 1, Apr. 1994, pp. 241–244.

[168] M. I. Gurelli and C. L. Nikias, "EVAM: An eigenvector-based algorithm for multi-channel blind deconvolution of input colored signals," *IEEE Trans. Signal Process.*, vol. 43, pp. 134–149, Jan. 1995.

[169] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Applied Signal Process.*, vol. 2003, no. 11, pp. 1074-1090, 2003.

[170] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

[171] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Fast estimation of a precise dereverberatioin filter based on speech harmonicity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, 2005.

[172] B. Radlovic, R. Williamson, and R. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 311–319, May 2000.

[173] A. Oppenheim, R. Schafer, and J. T.G. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264–1291, Aug. 1968.

[174] M. Tohyama, R. Lyon, and T. Koike, "Source waveform recovery in a reverberant space by cepstrum dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Minneapolis, USA, Apr. 1993, pp. 157–160.

[175] Qing-Guang Liu, B. Champagne, and P. Kabal, "A microphone array processing technique for speech enhancement in a reverberantspace," *Speech Communication*, vol. 18, pp. 317–334, Jun. 1996.

[176] K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: experimental validation," *EURASIP J. Audio, Speech, and Music Process.*, vol. 2007, Article ID 51831, 19 pages, 2007.

[177] C. Avendano, "Temporal processing of speech in a time-feature space," Ph.D. dissertation, Oregon Graduate Institute, Apr. 1997, http://www.bme.ogi.edu/~hynek/cgi-bin/publications/showbib_asp.pl?all.

[178] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, pp. 1069–1077, 1985.

[179] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 1982, pp. 156–159.

[180] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. Fourth Int. Conf. Spoken Language Process.*, vol. 2, Oct. 1996, pp. 889–892.

[181] J. Mourjopoulos and J. Hammond, "Modelling and enhancement of reverberant speech using an envelope convolution method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 8, Apr. 1983, pp. 1144–1147.

[182] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "An improved method based on the MTF concept for restoring the power envelope from a reverberant signal," *Acoustical Science and Technology*, vol. 25, no. 4, pp. 232–242, 2004.

[183] M. Unoki, K. Sakata, M. Furukawa, and M. Akagi, "A speech dereverberation method based on the MTF concept in power envelope restoration," *Acoustical science and technology*, vol. 25, no. 4, pp. 243–254, 2004.

[184] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Communication*, vol. 45, pp. 101–113, Feb. 2005.

[185] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "A method based on the MTF concept for dereverberating the power envelope from the reverberant signal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2003, pp. I–888–I–891.

[186] N. Gaubitch and P. Naylor, "Spatiotemporal averagingmethod for enhancement of reverberant speech," in *Proc. 15th Inter. Conf. Digital Signal Process.*, Cardiff, Wales, UK, Jul. 2007, pp. 607–610.

[187] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Mulitmicrophone speech dereverberation using spatiotemporal averaging," in *In Proc. European Signal Process. Conf.*, Sep. 2004, pp. 809–812.

[188] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[189] N. Gaubitch, P. Naylor, and D. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Int. Workshop Acoust., Echo Noise Control*, Sep. 2003.

[190] N. D. Gaubitch, D. B. Ward, and P. A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. Am.*, vol. 120, pp. 4031–4039, Dec. 2006.

[191] M. Wu, "Pitch tracking and speech enhancement in noisy and reverberant environments," Ph.D. dissertation, The Ohio State University, 2003.

[192] K. Furuya, S. Sakauchi, and A. Kataoka, "Speech dereverberation by combining MINT-based blind deconvolution and modified spectral subtraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 2006, pp. I–813–I–816.

[193] K. Furuya and A. Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, pp. 1579–1591, Jul. 2007.

[194] N. D. Gaubitch, E. A. P. Habets, and P. A. Naylor, "Multimicrophone speech dereverberation using spatiotemporal and spectral processing," in *Proc. IEEE Int. Symp. Circuits and Systems*, Seattle, Washington, USA, May 2008, pp. 3222–3225.

[195] E. Habets, N. Gaubitch, and P. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Apr. 2008, pp. 4577–4580.

[196] A. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computing*, vol. 7, pp. 1129–1159, 1995.

[197] K. Abed-Meraim, E. Moulines, and P. Loubaton, "Prediction error method for second-order blind identification," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 694–705, Mar. 1997.

[198] F. Ehlers and H. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment," *IEEE Trans. Signal Process.*, vol. 45, no. 10, pp. 2608–2612, Oct. 1997.

[199] K. Furuya and Y. Kaneda, "Two-channel blind deconvolution for nonminimum phase impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, Apr. 1997, pp. 1315–1318.

[200] M. D. T. Hikichi and M. Miyoshi, "Blind dereverberation based on estimates of signal transmission channels without precise informationon channel," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, USA, Mar. 2005, pp. 1069–1072.

[201] Y. Huang, J. Benesty, and J. Chen, "Blind channel identification based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 882–896, Sep. 2005.

[202] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP J. Audio Speech Music Process.*, vol. 2007, Article ID 84186, 15 pages, 2007.

[203] J. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio and Electroacoustics*, vol. 20, pp. 367–377, Dec. 1972.

[204] J. Wise, J. Caprio, and T. Parks, "Maximum likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, pp. 418–423, Oct. 1976.

[205] D. Krubsack and R. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Trans. Signal Process.*, vol. 39, pp. 319–329, Feb. 1991.

[206] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. Signal Process.*, vol. 39, no. 1, pp. 40–48, Jan. 1991.

[207] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound." in *Proceedings of the Institute of Phonetic Sciences*, vol. 17. University of Amsterdam, 1993, pp. 97–110.

[208] I. Atkinson, A. Kondoz, and B. Evans, "Pitch detection of speech signals using segmented autocorrelation," *Electronics Letters*, vol. 31, pp. 533–535, Mar. 1995.

[209] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 727–730, Oct. 2001.

[210] J. Hu, S. Xu, and J. Chen, "A modified pitch detection algorithm," *IEEE Communications Letters*, vol. 5, pp. 64–66, Feb. 2001.

[211] S. R. M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Montreal, Quebec, Canada, May 2004, pp. I–109–I–112.

[212] J.-X. Xu and J. C. Principe, "A pitch detector based on a generalized correlation function," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 8, pp. 1420–1432, Nov. 2008.

[213] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, pp. 293–309, 1967.

[214] S. Ahmadi and A. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 333–338, May 1999.

[215] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Am.*, vol. 43, pp. 829–834, 1968.

[216] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.*, vol. 83, pp. 257–264, Jan. 1988.

[217] C. K. Lee and D. G. Childers, "Cochannel speech separation," *J. Acoust. Soc. Am.*, vol. 83, pp. 274–280, Jan. 1988.

[218] B. Hanson and D. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 9, Mar. 1984, pp. 65–68.

[219] D. Childers and C. Lee, "Co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 12, Dallas, TX, Apr. 1987, pp. 181–184.

[220] T. Quatieri and R. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-38, pp. 56–69, Jan. 1990.

[221] Y. A. Mahgoub and R. M. Dansereau, "Time domain method for precise estimation of sinusoidal model parameters of co-channel speech," *Research Letters in Signal Process.*, vol. 2008, Article ID 364674, 5 pages.

[222] J. Naylor and J. Porter, "An effective speech separation system which requires no a priori information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Toronto, Canada, Apr. 1991, pp. 937–940.

[223] A. S. Bregman, *Auditory scene analysis*. Cambridge, MA: MIT Press, 1990.

[224] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communication*, vol. 27, no. 3-4, pp. 209–222, 1999.

[225] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, vol. 21, pp. 191–207, Apr. 1997.

[226] G. Hu and D. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo and Noise Control*, E. Hansler and G. Schmidt, Eds. Springer, Heidelberg, 2006, pp. 485–515.

[227] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[228] D. L. Wang and G. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 684–697, May 1999.

[229] G. J. Brown and D. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer Berlin Heidelberg, 2005, pp. 371–402.

[230] M. Slaney, "The history and future of CASA ," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, Mass, USA: Kluwer Academic, 2005, pp. 199–211.

[231] D. F. Rosenthal and H. G. Okuno, Eds., *Com.* Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., 1998.

[232] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[233] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[234] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech and Language Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[235] A. van der Kouwe, D. Wang, and G. Brown, "A comparison of auditory and blind separation techniques for speech segregation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 189–195, Mar. 2001.

[236] R. Stubbs and Q. Summerfield, "Evaluation of two voice separation algorithms using normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 84, pp. 1236–1249, Oct. 1988.

[237] R. J. Stubbs and Q. Summerfield, "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 87, no. 1, pp. 359–372, 1990.

[238] ——, "Effects of signal-to-noise ratio, signal periodicity, and degree of hearing impairment on the performance of voice-separation algorithms," *J. Acoust. Soc. Am.*, vol. 89, no. 3, pp. 1383–1393, 1991.

[239] D. Childers, D. Skinner, and R. Kemerait, "The cepstrum: A guide to processing," *Proc. IEEE*, vol. 65, pp. 1428–1443, Oct. 1977.

[240] A. Oppenheim and R. Schafer, "From frequency to quefrency: a history of the cepstrum," *IEEE Signal Process. Magazine*, vol. 21, pp. 95–106, Sep. 2004.

[241] J. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, pp. 2009–2025, 1998.

[242] C. Jutten and J. Herault, "Blind separation of sources, part 1: an adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, 1991.

[243] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Applied Signal Process.*, vol. 2003, no. 11, pp. 1135-1146, 2003.

[244] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, Y. Ikeda, H. Hashimoto, and T. Morita, "Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking," *EURASIP J. Applied Signal Process.*, vol. 2006, Article ID 34970, 17 pages, 2006.

[245] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[246] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Information Process. - Letters and Reviews*, vol. 6, no. 1, pp. 1–57, Jan. 2005.

[247] N. Das, A. Routray, and P. K. Dash, "ICA methods for blind source separation of instantaneous mixtures: A case study," *Neural Information Process. Letters and Reviews*, vol. 11, no. 11, pp. 225–246, Nov. 2007.

[248] A. Hyvrinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley & Sons, Inc., 2001.

[249] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[250] J. V. Stone, *Encyclopedia of Statistics in Behavioral Science*. Chichester: John Wiley & Sons, Ltd, 2005, vol. 2, ch. A Brief Introduction to Independent Component Analysis, p. 907912.

[251] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures," *IEEE Trans. Neural Networks*, vol. 19, no. 3, pp. 475–492, Mar. 2008.

[252] Q. Pan and T. Aboulnasr, "Time-domain convolutive blind source separation employing selective-tap adaptive algorithms," *EURASIP J. Audio Speech Music Process.*, vol. 2007, Article ID 92528, 11 pages, 2007.

[253] J.-F. Cardoso, "Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, Apr. 1990, pp. 2655–2658.

[254] R. Lambert and A. Bell, "Blind separation of multiple speakers in a multipath environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 1997, pp. 423–426.

[255] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 204–215, May 2003.

[256] G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single-channel source separation," *J. Machine Learning Research, Special issue on independent components analysis*, vol. 4, pp. 1365–1392, 2003.

[257] Z. Koldovsky and P. Tichavsky, "Time-domain blind audio source separation using advanced ICA methods," in *Proc.Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 27–31.

[258] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, Nov. 1998.

[259] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[260] N. Mitianoudis and M. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 489–497, Sep. 2003.

[261] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[262] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A, no. 7, pp. 1640–1655, Jul. 2005.

[263] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[264] G.-J. Jang, T.-W. Lee, and Y.-H. Oh, "Single-channel signal separation using time-domain basis functions," *IEEE Signal Process. Letters*, vol. 10, no. 6, pp. 168–171, Jun. 2003.

[265] D. Smith, J. Lukasiak, and I. Burnett, "Blind speech separation using a joint model of speech production," *IEEE Signal Process. Letters*, vol. 12, no. 11, pp. 784–787, Nov. 2005.

[266] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation based on multi-stage ICA combining frequency-domain ICA and time-domain ICA," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Orlando, USA, 2002, pp. I–917–I–920.

[267] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[268] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, pp. 309–319, Aug. 1979.

[269] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 325–333, Sep. 1995.

[270] P. Satyanarayana Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 609–619, Nov. 1999.

[271] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group delay function," *IEEE Signal Process. Letters*, vol. 14, no. 10, pp. 762–765, Oct. 2007.

[272] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[273] J. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, Sep. 1999.

[274] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Prentice Hall, 1975.

[275] K. Sri Rama Murty, B. Yegnanarayana, and S. Guruprasad, "Voice activity detection in degraded speech using excitation source information," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 2941–2944.

[276] B. Yegnanarayana and K. Sri Rama Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 614–624, May 2009.

[277] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, Oct. 2006.

[278] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, Apr. 1997, pp. 1647–1650.

[279] H. Maganti, P. Motlicek, and D. Gatica-Perez, "Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Honolulu, HI, USA, Apr. 2007, pp. IV–1037–IV–1040.

[280] N. Kanederaa, T. Araib, H. Hermanskyc, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43–55, May 1999.

[281] N. Kanedera, H. Hermansky, and T. Arai, "Desired characteristics of modulation spectrum for robust automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Seattle WA, USA, 1998, pp. 613–616.

[282] T. Houtgast and H. J. M. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, Mar. 1985.

[283] B. E. D. Kingsbury, "Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments," Ph.D. dissertation, University of California, Berkeley, 1998. [Online]. Available: http://www.icsi.berkeley.edu/~bedk/bedk_thesis.pdf

[284] S. R. M. Prasanna, B. V. Sandeep Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks and modulation spectrum energies," *IEEE Trans. Speech, Audio and Language-Process.*, vol. 17, no. 4, pp. 556–565, May. 2009.

[285] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, pp. 744–754, Aug. 1986.

[286] S. R. M. Prasanna and A. Subramanian, "Finding pitch markers using first order Gaussian differentiator," in *IEEE Proc. Third Int. Conf. Intelligent Sensing Information Process.*, Bangalore, India, Dec. 2005, pp. 140–145.

[287] M. Schroeder, "Parameter estimation in speech: A lesson in unorthodoxy," *Proc. IEEE*, vol. 58, no. 5, pp. 707–712, May 1970.

[288] J. G. Proakis and D. G. Manolakis, *Digital signal processing-principles, algorithms, and applications*, 3rd ed. Prentice Hall, 1996.

[289] "TIMIT acoustic-phonetic continuous speech corpus," NTIS Order PB91-505065, National Institute of Standards and Technology, Gaithersburg, Md, USA, 1990, Speech Disc 1-1.1.

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

[291] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, May 2006, pp. I–153–I–156.

[292] Y. Hu and Philipos C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.

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

[294] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc.Interspeech*, Philadelphia, PA, USA, Sep. 2006.

[295] Y. Hu and P. Loizou, "PESQ and other objective measures for evaluating quality of speech processed by noise suppression algorithms." [Online]. Available: http://www.utdallas.edu/~loizou/speech/software.htm

[296] ——, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[297] P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T, ITU-T Recommendation P.835*, Nov. 2003. [Online]. Available: http://www.itu.int/rec/T-REC-P.835/en

[298] A. K. Nábelek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, vol. 86, no. 4, pp. 1259–1265, 1989.

[299] M. Wu and D. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Hong Kong, China PR, Apr. 2003, pp. 892–895.

[300] E. A. P. Habets, S. Gannot, and I. Cohen, "Dereverberation and residual echo suppression in noisy environments," in *Speech and Audio Processing in Adverse Environments*. Springer, 2008, pp. 185–227.

[301] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 751–761, Sep. 2005.

[302] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.

[303] P. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Am.*, vol. 80, pp. 1527–1529, Nov. 1986.

[304] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Am.*, vol. 124, pp. 269–277, 2008.

[305] E. A. P. Habets, "Room impulse response generator for matlab." [Online]. Available: http://home.tiscali.nl/ehabets/rir_generator.html

[306] D. Wang and G. Hu, "Method for accurate pitch estimation and voice separation," US Patent Pending, File Number: 07032, Technology Licensing & Commercialization, The Ohio State University, Columbus.

[307] D. Morgan, E. George, L. Lee, and S. Kay, "Co-channel speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Detroid, USA., May 1995, pp. 828–831.

[308] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[309] J. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 6, pp. 998–1003, Dec. 1982.

[310] A. Kumar and Y. Bar-Shalom, "Time-domain analysis of cross correlation for time delay estimation with an autocorrelated signal," *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1664–1668, Apr. 1993.

[311] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Applied Signal Process.*, vol. 2006, Article ID 26503, 19 pages.

[312] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. Applied Signal Process.*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[313] R. Kumara Swamy, K. Sri Rama Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Process. Letters*, vol. 14, no. 7, pp. 481–484, Jul. 2007.

[314] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, Mass, USA: Kluwer Academic, 2005, pp. 181–197.

[315] B. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, Apr. 1976.

[316] J. P. Campbell, "Speaker recognition: a tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[317] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[318] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proc. Fourth Int. Conf. Spoken Language*, vol. 2, Oct. 1996, pp. 929–932.

[319] D. Reynolds, M. Zissman, T. Quatieri, G. O'Leary, and B. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Detroid, USA., May 1995, pp. 329–332.

[320] D. Reynolds, "The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Atlanta, USA, May 1996, pp. 113–116.

[321] K. Assaleh and R. Mammone, "New LP-derived features for speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 630–638, Oct. 1994.

[322] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," *IEEE Signal Process. Magazine*, vol. 13, no. 5, pp. 58–71, Sep. 1996.

[323] M. Zilovic, R. Ramachandran, and R. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 260–267, May 1998.

[324] L. P. Heck, Y. Konig, M. K. Sönmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Communication*, vol. 31, no. 2-3, pp. 181–192, 2000.

[325] R. Rose, E. Hofstetter, and D. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 245–257, Apr. 1994.

[326] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.

[327] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 3, pp. 190–202, May 1996.

[328] M. Kotti, L. Martins, E. Benetos, J. Cardoso, and C. Kotropoulos, "Automatic speaker segmentation using multiple features and distance measures: A comparison of three approaches," in *IEEE Int. Conf. Multimedia and Expo*, Jul. 2006, pp. 1101–1104.

[329] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, Aug. 1995.

[330] D. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Process. Letters*, vol. 2, no. 3, pp. 46–48, Mar. 1995.

[331] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, Nov. 2003.

[332] D. J. Mashao and M. Skosan, "Combining classifier decisions for robust speaker identification," *Pattern Recognition*, vol. 39, no. 1, pp. 147–155, 2006.

[333] D. Hosseinzadeh and S. Krishnan, "On the use of complementary spectral features for speaker recognition," *EURASIP J. Advances in Signal Process.*, vol. 2008, Article ID 258184, 10 pages, 2008.

[334] T. Kinnunen, "Spectral features for automatic text-independent speaker recognition," Ph.D. dissertation, Univ. Joensuu, Joensuu, Finland, 2003.

[335] J. E. Luck, "Automatic speaker verification using cepstral measurements," *J. Acoust. Soc. Am.*, vol. 46, no. 4B, pp. 1026–1032, 1969.

[336] W. S. Mohn, "Two statistical feature evaluation techniques applied to speaker identification," *IEEE Trans. Computers*, vol. C-20, no. 9, pp. 979–987, Sep. 1971.

[337] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, pp. 637–655, Aug. 1971.

[338] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Am.*, vol. 51, no. 6B, pp. 2044–2056, 1972.

[339] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[340] U. G. Goldstein, "Speaker-identifying features based on formant tracks," *J. Acoust. Soc. Am.*, vol. 59, no. 1, pp. 176–182, 1975.

[341] C. LaRiviere, "Contributions of fundamental frequency and formant frequencies to speaker identification," *Phonetica*, vol. 31, pp. 185–197, 1975.

[342] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, no. 2, pp. 176–182, Apr. 1975.

[343] J. D. Markel, B. T. Oshika, and H. Gray, Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 4, pp. 330–337, Aug 1977.

[344] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[345] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 3, pp. 342–350, Jun. 1981.

[346] ——, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[347] C. C. Johnson, H. Hollien, and J. W. Hicks., "Speaker identification utilizing selected temporal speech features," *J. Phonetics*, vol. 12, pp. 319–326, 1984.

[348] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 52–59, Feb. 1986.

[349] C. Berasconi, "On instantaneous and transitional spectral information for text-dependent speaker verification," *Speech Communication*, vol. 9, no. 4, pp. 129–139, 1990.

[350] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, Apr. 1990.

[351] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, pp. 1215–1247, Sep. 1993.

[352] N. Fakotakis, A. Tsopanoglou, and G. Kokkinakis, "A text-independent speaker recognition system based on vowel spotting," *Speech Commun.*, vol. 12, no. 1, pp. 57–68, 1993.

[353] B. Imperl, Z. Kacic, and B. Horvat, "A study of harmonic features for the speaker recognition," *Speech Communication*, vol. 22, no. 4, pp. 385–402, Sep. 1997.

[354] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing Mel-frequency cepstral coefficients on the power spectrums," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Salt Lake City, USA, 2001, pp. 73–76.

[355] C.-T. Hsieh, E. Lai, and Y.-C. Wang, "Robust speech features based on wavelet transform with application to speaker identification," *IEE Proc. Vision, Image and Signal Process.*, vol. 149, no. 2, pp. 108–114, Apr. 2002.

[356] A. Petry and D. A. C. Barone, "Speaker identification using nonlinear dynamical features," *Chaos, Solitons, & Fractals-Elsevier*, vol. 13, no. 2, pp. 221–231, 2002.

[357] A. G. Adami, "Modeling prosodic differences for speaker recognition," *Speech Communication*, vol. 49, no. 4, pp. 277–291, 2007.

[358] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Apr. 2003, pp. IV–788–IV–791.

[359] E. Shriberg, L. Ferrer, S. S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition." *Speech Communication*, vol. 46, no. 3-4, pp. 455–472, 2005.

[360] N. Zheng, T. Lee, and P. C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Process. Letters*, vol. 14, no. 3, pp. 181–184, Mar. 2007.

[361] K. S. R Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Letters*, vol. 13, no. 1, pp. 52–55, Jan. 2006.

[362] D. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans., Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, Oct. 1994.

[363] N. Tisby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 3, pp. 563–570, Mar. 1991.

[364] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 10, Apr. 1985, pp. 387–390.

[365] K. Farrell, R. Mammone, and K. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 194–205, Jan. 1994.

[366] B. Yegnanarayana and S. P. Kishore, "AANN: an alternative to GMM for pattern recognition," *Neural Netw.*, vol. 15, no. 3, pp. 459–469, 2002.

[367] M. Schmidt and H. Gish, "Speaker identification via support vector classifiers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Atlanta, USA, May 1996, pp. 105–108.

[368] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[369] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Process. Magazine*, vol. 11, no. 4, pp. 18–32, Oct. 1994.

[370] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 3, pp. 456–459, Jul. 1994.

[371] K. Yu, J. Mason, and J. Oglesby, "Speaker recognition using Hidden Markov Models, dynamic time warping and vector quantisation," *IEE Proc. Vision, Image and Signal Process.*, vol. 142, no. 5, pp. 313–318, Oct. 1995.

[372] B. Pellom and J. Hansen, "An efficient scoring algorithm for Gaussian mixture model based speaker identification," *IEEE Signal Process. Letters*, vol. 5, no. 11, pp. 281–284, Nov. 1998.

[373] D. A. Reynolds, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[374] X. Yue, D. Ye, C. Zheng, and X. Wu, "Neural networks for improved text-independent speaker identification," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, no. 2, pp. 53–58, Mar/Apr. 2002.

[375] B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural Gaussian mixture models and neural network," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 447–456, Sep. 2003.

[376] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, M. I. Chagnolleau, S. Meignier, T. Merlin, O. J. Garcia, P. Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Applied Signal Process.*, vol. 2004, no. 4, pp. 430-451, 2004.

[377] V. Prakash and J. Hansen, "In-set/out-of-set speaker recognition under sparse enrollment," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 7, pp. 2044–2052, Sep. 2007.

[378] R. Schafer and L. Rabiner, "Digital representations of speech signals," *Proc. IEEE*, vol. 63, no. 4, pp. 662–677, April 1975.

[379] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and Trends in Signal Processing*, vol. 1, no. 1, pp. 1–194, Jan. 2007.

[380] B. S. Atal and M. R. Schroeder, "Linear prediction analysis of speech based on a pole-zero representation," *J. Acoust. Soc. Am.*, vol. 64, no. 5, pp. 1310–1318, 1978.

[381] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, Feb. 1988.

[382] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music.* John Wiley & Sons, Inc., 1999.

[383] J. E. Markel and A. H. Gray, *Linear Prediction of Speech.* Berlin: Springer-Verlag New York, Inc., 1976.

[384] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1st ed. Prentice Hall, 2001.

[385] Y. Hu and P. C. Loizou, "A comparative intelligibility study of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Honolulu, HI, USA, Apr. 2007, pp. IV–561 – IV–564.

[386] S. Dimolitsas, F. Corcoran, and C. Ravishankar, "Dependence of opinion scores on listening sets used in degradation category rating assessments," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 421–424, Sep. 1995.

[387] U. Halka, "A new objective quality measure for speech-coding systems based on the estimation of their nonlinear properties," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toronto, Canada, 1991, pp. 497–500.

[388] A. Spanias, "Speech coding: a tutorial review," *Proc. IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct 1994.

[389] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Language Process.*, Dec. 1998, pp. 2819–2822.

[390] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 1, pp. 67–72, Feb. 1975.

[391] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1982, pp. 1278–1281.

[392] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, USA, 2001, pp. 749–752.

[393] A. P. H. Antony W. Rix, Michael P. Hollier and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part I - time-delay compensation," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 755–764, Oct. 2002.

[394] A. W. R. John G. Beerends, Andries P. Hekstra and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part II - psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 765–778, Oct. 2002.

[395] J. Volkmann, S. Stevens, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Am.*, vol. 8, pp. 185–190, Jan. 1937.

[396] Z. Fang, Z. Guoliang, and S. Zhanjiang, "Comparison of different implementations of MFCC," *J. Computer Science and Technology*, vol. 16, no. 6, pp. 582 – 589, 2001.

[397] G. K. T. Ganchev, N. Fakotakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proc. SPECOM*, vol. 1, 2005, pp. 191–194.

[398] J. Mason and X. Zhang, "Velocity and acceleration features in speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, Toronto, Canada, Apr. 1991, pp. 3673–3676.

[399] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[400] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[401] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. the fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[402] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[403] Q. Y. Hong and S. Kwong, "A discriminative training approach for text-independent speaker recognition," *Signal Process.*, vol. 85, no. 7, pp. 1449–1463, 2005.

[404] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

List of Publications

Journal Publications

1. P. Krishnamoorthy and S. R. M. Prasanna, "Processing noisy speech for enhancement," *J. IETE Technical Review, Special issue on spoken language processing*, vol. 24, pp. 349–355, Sep.-Oct. 2007.

2. P. Krishnamoorthy and S. R. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *IEEE Trans. Speech, Audio and Language Processing*, vol.17, no.2, pp. 253-266, Feb. 2009.

3. P. Krishnamoorthy and S. R. M. Prasanna, "Temporal and Spectral Processing Methods for Processing of Degraded Speech: A Review," *J. IETE Technical Review*, vol. 26, Issue 2, pp. 137-148, Mar.-Apr. 2009.

4. P. Krishnamoorthy and S. R. M. Prasanna, "Application of Combined Temporal and Spectral Processing Methods for Speaker Recognition under Noisy, Reverberant or Multi-Speaker Environments," to appear in *Sadhana-Academy Proceedings in Engineering Sciences (springer).*

5. S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks and modulation spectrum energies," *IEEE Trans. Speech, Audio and Language Processing*, vol.17, no.4, pp. 556-565, May 2009.

Manuscripts Submitted

1. P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," Feb. 2008, submitted to *Speech Communication (Elsevier).*

2. P. Krishnamoorthy and S. R. M. Prasanna, "Two speaker speech separation by temporal and spectral processing," Nov. 2008, submitted to *Digital Signal Processing (Elsevier).*

3. P. Krishnamoorthy and S. R. M. Prasanna, "Temporal and Spectral Processing Methods for Processing of Degraded Speech," Mar. 2009, submitted to *International Journal of Parallel, Emergent and Distributed Systems (Taylor & Francis).*

Conference and Workshop Publications

1. P. Krishnamoorthy and S. R. M. Prasanna, "Modified spectral subtraction method for enhancement of noisy speech," in *IEEE Proc. Third Int. Conf. Intelligent Sensing and Information Processing (ICISIP 2005)*, Bangalore, India, Dec. 2005, pp. 146 – 150.

2. S. R. M. Prasanna, P. Krishnamoorthy, and B. Yegnanarayana, "Speech enhancement using source features and group delay analysis," in *IEEE Proc. Annual IEEE INDICON (INDICON 2005)*, Chennai, India, Dec. 2005, pp. 19 – 23.

3. P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by spectral subtraction and residual modification," in *IEEE Proc. Annual IEEE INDICON (INDICON 2006)*, New Delhi, India, Sep. 2006.

4. N. Gandhi, P. Krishnamoorthy, and S. R. M. Prasanna, "Reduction of musical noise in spectral subtracted speech using excitation source information," in *IEEE Proc. Annual IEEE INDICON (INDICON 2006)*, New Delhi, India, Sep. 2006.

5. P. Krishnamoorthy and S. R. M. Prasanna, "Processing noisy speech by noise components subtraction and speech components enhancement," in *Proc. Int. Conf. Systemics, Cybernetics and Informatics (ICSCI 2007)*, Hyberabad, India, Jan. 2007.

6. P. Krishnamoorthy and S. R. M. Prasanna, "Temporal and spectral processing for enhancement of noisy speech," in *Proc. Workshop on signal and image processing (WISP 2007)*, Guwahati, India, Dec. 2007, pp. 51–56.

7. B. V. Sandeep Reddy, P. Krishnamoorthy and S. R. M. Prasanna, "Keyword Spotting System using Vowel Onset Point Events," accepted in *International Symposium Frontiers of Research on Speech and Music (FRSM 2008)*, Kolkata, India, Feb. 2008.

8. P. Krishnamoorthy and S. R. M. Prasanna, "Temporal and spectral processing of degraded speech," *IEEE Proc. Int. Conf. Advanced Computing and Communications 2008 (ADCOM 2008)*, Chennai, India, Dec. 2008, pp. 112-118.

9. S. R. M. Prasanna, Kiran Bakki, and P. Krishnamoorthy, "Time-delay estimation using source and spectral information from speech," in *Proc. Fifteenth National Conference on Communications 2009 (NCC 2009)*, Guwahati, India, 16-18 Jan. 2009. pp. 272-275.