EVENT-BASED ANALYSIS OF SPEECH

A THESIS

submitted by

S.R. MAHADEVA PRASANNA

for the award of the degree

of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

CHENNAI

MARCH 2004

To my guide **PROF.B. YEGNANARAYANA**

for his guidance and inspiration

To my wife **S.R. NIRMALA**

for her patience and cooperation

To my friend **S. VIJAYAKUMAR**

for his encouragement and moral support

THESIS CERTIFICATE

This is to certify that the thesis entitled **EVENT-BASED ANALYSIS OF SPEECH** submitted by **S.R. Mahadeva Prasanna** to the Indian Institute of Technology Madras, Chennai for the award of the degree of Doctor of Philosophy is a bonafide record of research work carried out by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Chennai 600 036	Prof. B. Yegnanarayana
Date	Dept. of Computer Science and Engg.

ACKNOWLEDGEMENTS

I would like to express my deep sense of gratitude and respect to Prof.B. Yegnanarayana for taking me as his student and guiding me to successfully complete this research work. His qualities as a teacher, researcher, guide and philosopher have attracted me immensely. His discipline and hardwork are the great sources of inspiration for me. Association with him for about five years is a golden part of my life. I sincerely thank him for the excellent research environment he has created for students to learn and pursue research work.

I thank Dr. Hema A. Murthy, Dr. C. Chandra Sekhar and Dr. Sukhendu Das for their valuable guidance and suggestions. I would like to thank Prof.C. Pandurangan and Prof.S. Raman, chairmen of the department during my stay for providing me an opportunity to work in this department and making available various facilities. Heartfelt thanks to Prof.S. Raman for supporting me financially during the final stage of my work. Sincere thanks to my doctoral committee members Prof.C. Siva Ram Murthy, Prof.D. Janaki Ram, Prof. Megha Singh and Prof.V.V. Rao for sparing their valuable time to evaluate the progress of my research work.

I owe my special thanks to Dr.S. Rajendran (sir) for maintaining an excellent computing facility and various resources useful for the research work in our speech and vision group. In my personal view the facilities we (speech and vision group members) are enjoying was not possible without his dedicated effort. I also thank him for the valuable suggestions he gave at different stages of my work.

Sincere thanks to my friend S. Vijayakumar for sparing his time with me to give moral support and help, throughout my stay at IIT Madras. His philosophical approach towards life and research boosted my self-confidence and also discussions with him encouraged me all the time.

I thank my friend K. Sreenivasa Rao for all his help and brotherly care without which my stay would not have been so comfortable in the lab. I also thank him for the time he had spent in patiently correcting all my papers and thesis. My special thanks to Mrs.K. Ranga Kumari

(Mrs. KSR) for all the help she rendered to me and my wife.

Student part of my life at IIT Madras would have been incomplete if I had not spent time with my friends anil (child), chaitu (SR), dhanu (Mr.?), guru (sink), murty (no ra ..), satish (LK) and suresh (KS) in discussing nontechnical matters. I also thank all my lab members anil, anitha, anvita, chaitu, CKM (sir), dhanu, guru, KSR, leena, murty, nayeem, pal, RKS, satish, shahina, suresh and surya for their help and cooperation.

I will be failing in my duty if I do not render my thanks to my senior Dr.P. Satyanarayana, whose programs I extensively used for my work.

I thank my friend Dr. Mallikarjun S. Holi for introducing me to this premier institute for the first time. I also thank my friend T. Nagarajan for all the discussions we had.

It is the time for me to thank my wife S.R. Nirmala, without her support and encouragement this research work would have been next to impossible. I thank her for all the help, care and moral support provided.

I would like to thank my teachers, parents and brothers for making me what I am today.

I also thank everyone who helped me directly or indirectly during my stay at IIT Madras.

Finally, this research work would not have been possible without the unlimited support from my guide Prof.B. Yegnanarayana, my wife S.R. Nirmala and my friend S. Vijayakumar and hence I sincerely dedicate this work to them.

S.R. Mahadeva Prasanna

ABSTRACT

KEYWORDS: *events in speech; event-based analysis; glottal closure event; vowel onset point event.*

This thesis proposes an *event-based* approach for the analysis of speech, and is inspired by the nature of speech production. A sequence of changes takes place during the production of speech. These changes are manifested in the speech signal and are treated as *events*. From the speech production point of view, events indicate the instants of significant activity, and hence important and discriminatory information for the analysis of speech is present around the events. Events are used as the anchor points, and analysis of the characteristics of the signal around the events is carried out to develop new methods for processing speech. The event-based approach involves defining the chosen event in terms of the changes that occur during the production of speech, deriving methods for the detection of the event and developing methods based on the chosen event for different applications of speech. The proposed event-based approach is illustrated using two important events, namely, the Glottal Closure (GC) and the Vowel Onset Point (VOP). The GC event is defined as the instant at which the closure of vocal folds takes place within a pitch period. The VOP event is defined as the instant at which the onset of vowel takes place.

The major contributions of the thesis are:

- Methods are discussed for the detection of the GC events.
- A Method is proposed for the extraction of pitch in adverse conditions by exploiting the properties of the signal around the GC events.
- A speech enhancement method based on the GC event information is proposed for processing degraded speech collected over a single channel
- A method for the estimation of time-delay between the speech signals collected over a pair of spatially distributed microphones is developed using the GC event information.
- A method based on the GC event information is proposed to process degraded speech signals collected from multiple microphones to produce enhanced speech.

- A method for enhancing speech of the desired speaker from the speech collected in a multispeaker environment is developed using the knowledge of the GC events.
- Methods for the detection of the VOP events are proposed using the GC events as anchor points for the analysis of speech.
- A method for the detection of the end-points of a speech utterance is developed using the knowledge of the VOP events.

Contents

List of Tables

List of Figures

ABBREVIATIONS

GC	–	Glottal Closure
VOP	–	Vowel Onset Point
FT	–	Fourier Transform
DFT	–	Discrete Fourier Transform
LP	–	Linear Prediction
LPCs	–	Linear Prediction Coefficients
CV	–	Consonant Vowel
SCV	–	Stop Consonant Vowel
UVUA	–	Unvoiced Unaspirated
UVA	–	Unvoiced Aspirated
VUA	–	Voiced Unaspirated
VA	–	Voiced Aspirated
SNR	–	Signal to Noise Ratio
MLED	–	Maximum Likelihood Epoch Detection
SVD	–	Singular Value Decomposition
TDOA	–	Time Differences of Arrival
GCC	–	Generalized Crosscorrelation
MCESA	–	Minimum-Cross-Entropy Spectrum Analysis
HMS	–	Harmonic Magnitude Suppression
HES	–	Harmonic Enhancement and Suppression
BSS	–	Blind Source Separation
DTW	–	Dynamic Time Warping
SIFT	–	Simple Inverse Filtering Technique
WLPCC	–	Weighted Linear Prediction Cepstral Coefficients

NOTATION

a_k	– kth linear prediction coefficient
$\mathbf{a}$	– vector of linear prediction coefficients: $[a_1 \ a_2 \cdots a_p]^T$
n	– discrete time index
p	– linear prediction order
$s(n)$	– speech signal as a function of time index n
$x(n)$	– signal variable
η	– normalized prediction error
ω	– frequency variable in radians for a discrete time signal
$e(n)$	– residual signal
$e_h(n)$	– Hilbert transform
$h_e(n)$	– Hilbert envelope
$\phi(\omega)$	– Fourier transform phase
$-\phi'(\omega)$	– Negative derivative of Fourier transfrom phase
$(.)^*$	– complex conjugate

Chapter 1

INTRODUCTION

1.1 Objective of the Thesis

Speech is produced as a sequence of changes, and these changes are viewed as events in this work. Important information for processing speech is present around the events [1, 2]. For effective representation and analysis of speech, it is useful to know what these events are, and then extract and use the knowledge of such events for processing speech. Conventional block processing approach processes speech in uniform blocks of 10-30 ms, and it does not exploit the event nature of speech. This thesis proposes an event-based approach for the analysis of speech. The focus is on identifying and detecting some events occurring in speech and developing new methods for analysis of speech using these events. By the nature of production of speech, the events are generally high Signal-to-Noise Ratio (SNR) regions. By that we mean, level of signal is high around the events compared to other places. Hence the proposed event-based approach may provide robustness and may also result in improved performance, since feature extraction and processing is anchored around the events.

1.2 Events in Speech

The dictionary meaning of the term *event* is something that happens. Thus any happening that draws the attention may be viewed as an event. Hence the term event is used in different fields and even in case of speech itself in different contexts. To name a few, in case of speech there are phonetic events and acoustic events. Any feature which can be attributed to the activity of the speech organs is a phonetic event. For instance, voicing and closure are phonetic events [3–5]. Any feature which is present in the acoustic signal may be treated as an acoustic event. For instance, burst, frication and Voice Onset Time (VOT) are the acoustic events. Thus so far in case of speech, event indicates a property that exist over a region. In this work the term event is used in a slightly different sense. Any significant change during the production of speech, manifested in the acoustic signal is viewed as an event. One important deviation in our definition of event is that event is an instant property, where as in the earlier definitions event may represent a region property.

When information is to be conveyed, formulation of message takes place in the mind of the speaker. The formulated message is coded using the sound units and suprasegmental features of the language. The coded message generates a sequence of neuromuscular commands. These commands change the shape of vocal tract system and the nature of excitation, which results in the production of speech. The sequence of changes in the shape of the vocal tract and the nature of excitation are reflected as events in the speech signal. From the perception point of view, events and regions around them are known to contain important information [1,6–26]. A brief discussion about the significance of events for perception is given in Appendix–A.

Speech can be considered as a sequence of events, where an event can be interpreted as change in some characteristics of speech production reflected in the speech signal. The speech signal is also affected by the changes in the environmental characteristics (noise, reverberation, other speech signals), microphone and channel characteristics. In the present work, changes caused by deliberate attempt of producing speech alone are considered as events. In defining events, changes with respect to time only are considered. The changes occurring in the vocal tract system and the excitation source characteristics may be viewed at various levels such as signal level, production level, acoustic system level, phonetic level, sound unit level, suprasegmental level, speaker level and language level.

At the signal level, changes in the time domain characteristics and changes in the frequency domain characteristics as a function of time may be treated as events. Each instant of significant excitation is an event at the signal level. Similarly, changes in the formant (resonant) frequency values are also events at the signal level. At the production level, speech may be characterized in terms of production features such as voicing, aspiration, frication and burst. Onset of any of these features and change from one feature to the other may be treated as events at the production level. The characteristics of the vocal tract (acoustic) system depends on the positioning of various articulators, which in turn decides the type of speech sound produced. The changes in the positioning of articulators may be treated as events at the acoustic system level. For instance, during the production of bilabial sounds, opening of lips from initial closure is an event. At the phonetic level speech may be interpreted in terms of sequence of phonemes such as consonants and vowels, and transition from one phoneme to other may be treated as an event. Speech may also be viewed as a sequence of sound units such as syllables. Onset of syllable and change from one syllable to the other may be treated as events at the sound units level. By defining the onset and changing characteristics of the suprasegmental features as events, it may be possible to analyze and extract suprasegmental features in a better way. For example, onset of raising and lowering of pitch contour may be treated as events at the suprasegmental level. In a conversation of two or more speakers, change from one speaker to the other is an event at the speaker level. In a multi-lingual scenario change from one language to the other is an event at the language level.

As discussed above, speech may be viewed as a sequence of events at various levels, and a summary of this discussion is given in Table 1.1. The present work focuses on two events, namely, the Glottal Closure (GC) and the Vowel Onset Point (VOP) events. The GC event is the instant at which the closure of vocal folds takes place within a pitch period. The VOP event is the instant at which the onset of vowel takes place. GC is an event at the signal level and VOP is an event at the phonetic level.

1.3 Event-based Analysis

The proposed event-based approach for analysis of speech involves the following steps:

Table 1.1: Grouping of events in speech and examples for each group.

Sl.No.	Category of event	Category of linguistic unit	Examples of some events
1	Signal	Signal amplitude	Instants of significant excitation
2	Production	Manner of articulation	Onset of voicing, burst, fricative
3	Acoustic system	Gestural closure and release	Opening of lips, raising of velum
4	Phonetic	Melodic properties	Change from consonant to vowel
5	Sound unit	Syllable affiliation	Change from one syllable to other
6	Suprasegmental	Tonal-metrical-prosodic tier	Onset of raising or lowering of pitch
7	Speaker	Idiolect identity	Change from one speaker to the other
8	Language	Grammatical tier	Change from one language to the other

- Defining the event for study in terms of the changes occurring during the production of speech.
- Observation of the speech signal to identify the changes around the event
- Proposing a set of acoustic cues for the detection of the event.
- Developing a method for the automatic detection of the event.
- Proposing methods for processing speech for various applications using the knowledge of the event.

1.4 Significance of the GC and VOP Events

Knowledge of the GC events is useful for accurate estimation of pitch period. The closed glottis interval starts at the GC event, and analysis of the speech signal in the closed glottis interval provides an accurate estimate of the frequency response of the vocal tract system [2, 27]. The GC events may be used as pitch markers for prosodic manipulation which is useful in several applications like text-to-speech synthesis, voice conversion and speech rate conversion [28]. The GC event is useful in detecting events at higher levels like VOP events [29]. Knowledge of GC events may be used for estimating time-delay between speech signals collected over a pair of spatially distributed microphones [30]. The SNR of speech signal is high around the GC events, and hence it is possible to enhance speech by exploiting the characteristics of speech signals around the GC events [31]. Enhancement of speech in a multispeaker environment may be achieved by extracting the unique sequence of the GC events corresponding to each speaker, and synthesizing speech using a modified excitation sequence [32].

As vowel is the nucleus of a syllable, segmentation of speech into syllable-like units may be done better at the signal level with the knowledge of the VOP events [29]. Knowledge of the VOP event helps in extracting a fixed duration pattern that contains most of the necessary information for recognition of CV units [33]. The regions immediately after the onset of vowel are less noisy than other regions. Hence these regions can be used for time-delay estimation and enhancement of speech [30]. The VOP event may also be used for detection of end-points, which is important in applications like text-dependent speaker verification [34].

1.5 Scope of the Present Work

The scope of this work is to illustrate the effectiveness of using the knowledge of events for analysis of speech. The GC and VOP events are chosen to discuss various issues involved in the proposed event-based approach. Since it is difficult to detect the events directly from the speech signal under all conditions, a set of acoustic cues are proposed. Methods for automatic detection of the GC and VOP events are discussed. The knowledge of the GC and VOP events are used in the following studies:

- Extraction of pitch in adverse conditions.
- Enhancement of degraded speech collected over a single microphone.
- Estimation of time-delay between the speech signals collected over a pair of spatially distributed microphones.

- Enhancement of degraded speech collected over a set of spatially distributed microphones.
- Enhancement of speech of desired speaker degraded by speech of other speakers.
- Detection of end-points of a speech utterance.

1.6 Organization of the Thesis

The evolution of ideas presented in this thesis are listed in Table 1.2. The contents of the thesis are organized as follows:

In **Chapter 2**, a review of the existing methods for detection of GC events, extraction of pitch, estimation of time-delay, enhancement of speech, detection of the VOP events and detection of end-points is presented.

Chapter 3 discusses issues related to the detection of GC events. Identification of acoustic cues for detection of GC events is explained. An algorithm for automatic detection of GC events is discussed.

Chapter 4 illustrates the usefulness of GC events in two applications namely, extraction of pitch and speech enhancement in single channel case. Using information about GC events, a method is proposed for extraction of pitch in adverse conditions. A method for enhancement of degraded speech collected over a single channel using GC events is also proposed.

In **Chapter 5** three more applications of GC events namely, estimation of time-delay, speech enhancement in multichannel case and speech enhancement in multi-speaker environment are discussed. A method for time-delay estimation between a pair of spatially distributed microphones using GC events information is proposed. A method based on the knowledge of GC events derived from the multiple microphone signals for enhancement of speech is proposed. A method based on estimated time-delays and GC events is proposed for enhancement of speech in multispeaker environment.

Chapter 6 focuses on the detection of VOP events. Acoustic cues for the detection of VOP events are proposed. An algorithm for automatic detection of VOP events is proposed.

Chapter 7 proposes a method based on the knowledge of VOP events for detection of the end-points. The proposed end-points detection method is evaluated by conducting speaker verification studies.

A summary of the present work is given in **Chapter 8** by listing major contributions of the present work and some directions for further research in the area of event-based analysis of speech.

Table 1.2: Evolution of ideas presented in the thesis.

Event-based Analysis of Speech

- Human beings produce speech as a sequence of events
- Important information for processing speech is present around the events
- Event-based approach is an attractive alternative to block processing
- Steps in event-based analysis
  - Defining the event for study
  - Visual observation of changes at the event
  - Proposing acoustic cues for detection of the event
  - Proposing a method for automatic detection of the event
  - Illustrating usefulness of the event for different applications
- GC and VOP events are chosen for study
- Acoustic cues for detection of the GC and VOP events
- Automatic detection of the GC and VOP events
- Applications of the GC and VOP events
  - Extraction of pitch in adverse conditions
  - Enhancement of speech in single channel case
  - Time-delay estimation
  - Enhancement of speech in a multichannel case
  - Enhancement of speech in a multispeaker environment
  - Detection of end-points

Chapter 2

ISSUES IN SPEECH ANALYSIS - A REVIEW

This chapter reviews some of the issues in the analysis of speech which are addressed in the present work on event-based approach. Methods proposed in the literature for detecting GC events are discussed in Section 2.1. The GC events are also termed as epochs or instants of significant excitation, and hence these terms will be used interchangeably in this work. Using information about GC events, methods are proposed for extraction of pitch, enhancement of degraded speech, multispeaker processing and detection of VOP events. Section 2.2 discusses the existing methods for extraction of pitch. Time-delay estimation methods proposed in the literature are discussed in Section 2.3. Section 2.4 reviews methods for speech enhancement against background noise and reverberation. Approaches for enhancement of speech in a multispeaker environment are discussed in Section 2.5. Some methods have been proposed for detecting VOP events, and these are reviewed in Section 2.6. Methods used for the detection of end-points are discussed in Section 2.7. A summary of the issues discussed in this chapter is given in Section 2.8.

2.1 Detection of GC Events

The first contribution to the detection of the GC event is due to Sobakin [35]. A slightly modified version is proposed by Strube [36]. In Strube's work, some predictor methods based on Linear Prediction (LP) analysis for the determination of the GC events are reviewed, which do not always yield reliable and unequivocal results. Then Sobakin's method using the determinant of the autocovariance matrix is examined critically, and reinterpreted such that the determinant is maximum if the beginning of the interval on which the autocovariance matrix is calculated coincides with the glottal closure.

Method based on the decomposition of composite signals is proposed for epoch extraction of voiced speech [37]. The general epoch filter theory is applied to the outputs of models of voiced speech and to actual speech data. It is shown that the points of excitation of the vocal tract can be precisely identified for continuous speech. However this method is suitable for analyzing only clean speech. A large value in LP residual is supposed to indicate the epoch location [38]. However, there are often ambiguities in the direct use of the LP residual since samples of either polarity occur around the epochs. A detailed study is made on the determination of the epochs from the LP residual [27]. Finally a method for unambiguous identification of epochs from the LP residual is proposed [27].

A least squares approach for glottal inverse filtering from the acoustic speech waveform is proposed [39]. In this work covariance analysis as a least squares approach for accurately performing glottal inverse filtering from the acoustic speech waveform is discussed. A method based on maximum-likelihood theory for epoch determination is proposed for detecting the GC event [40]. The speech signal is processed to get Maximum-Likelihood Epoch Detection (MLED) signal. The strongest positive pulse indicates the GC event within a pitch period. However the MLED signal creates not only a strong and sharp epoch pulse, but also a set of weaker pulses which represent the suboptimal epoch candidates within a pitch period. Hence a selection function is derived using the input signal and its Hilbert transform, which emphasizes the contrast between the epoch pulse and the subpulses. Using MLED signal and selection signal with appropriate threshold, the epochs are detected. The limitation of this method is the choice of window for deriving the selection function and also the use of threshold for deciding the epochs. A Frobenius norm approach to the detection of GC events is also proposed [41]. In this work a new approach based on Singular Value Decomposition(SVD) is proposed. The SVD method amounts to calculating the Frobenius norms of signal matrices, and is therefore, computationally efficient. The limitation of this approach is that it is shown to be working only for vowels. No attempt has been made in detecting the GC events in difficult cases like nasals, voiced consonants and semivowels.

A method for detecting the GC events in speech using the properties of minimum phase signals and group delay functions is proposed [42,43]. The method is based on the global phase characteristics of minimum phase signals. The average slope of the unwrapped phase (phase slope) of the short-time Fourier transform of linear prediction residual is calculated as a function of time. Instants where the phase slope function makes a positive zero-crossing are identified as GC events.

The GC event is an instant property. But, in most of the methods discussed above (except [27,43]), the GC events are detected by employing block processing approach, which results in ambiguity about the precise location of the GC events. In general, it is difficult to detect the GC events in case of low voiced consonants, nasals, semivowels, breathy voices and female speakers. A summary of the discussion related to the detection of GC events is given in Table 2.1.

Table 2.1: Summary of the review of detection of GC events.

- All the methods processes either the speech signal or the LP residual signal to generate an output signal in which there will be prominent peaks at the GC events. The output signal is interpreted properly to detect the location of the GC events.

- The limitation of existing methods (except [27,43]) is that even though GC event is an instant property, they employ block processing, which results in ambiguity about the precise location of the GC events.

- Methods ([27,43]) based on the instant property will detect the GC events with high accuracy.

2.2 Extraction of Pitch

There are several algorithms proposed in the literature for the extraction of pitch. These algorithms may be broadly classified into three categories [44], namely, algorithms using time domain properties, algorithms using frequency domain properties and algorithms using both time and frequency domain properties of speech signals. The algorithms based on time domain properties operate directly on the speech signal to estimate pitch. Most often, the measurements for these algorithms are peak and valley detection, zero-crossings and autocorrelation. The basic assumption is that if a quasiperiodic signal has been suitably processed to minimize the effects of formant structure, then simple time domain measurement will provide good estimate of pitch period. The algorithms based on frequency domain properties of speech signals assume that if the signal is periodic in the time domain, then the frequency spectrum of the signal contains a series of impulses at the fundamental frequency and its harmonics. Thus simple measurements can be made on the frequency spectrum of the signal or a nonlinearly transformed version of it, as in the cepstral method [45], to estimate the pitch period of the signal. In the third category, frequency domain approach may be used to spectrally flatten the time domain signal, and then an autocorrelation measurement is used for the extraction of pitch.

The cepstrum method for extraction of pitch utilizes the frequency domain properties of speech signals [45]. In the short-term spectrum of a given voiced frame, the information about the vocal tract appear as slowly varying component, and that of the excitation source as high frequency variations. These two components may be separated by considering the logarithm of the spectrum, and then applying inverse Fourier transform to obtain the cepstrum. This operation transforms the information from frequency domain to cepstral domain, which has a strong peak corresponding to the pitch period of the voiced speech frame being analyzed.

Simple Inverse Filtering Technique (SIFT) algorithm uses both time and frequency domain properties of the speech signal [46]. In the SIFT algorithm, speech signal is spectrally flattened and autocorrelation analysis is performed for the extraction of pitch. Due to spectral flattening, a prominent peak will be present at the pitch period of the voiced speech frame being analyzed.

Most of the existing methods for extraction of pitch work well for clean speech, and their performance will degrade severely for degraded conditions. This is because, peaks in autocorrelation function or cepstrum may not be prominent or unambiguous due to degradations. A summary of the discussion related to the extraction of pitch is given in Table 2.2.

Table 2.2: Summary of the review of extraction of pitch.

- Existing methods processes speech either in time domain or frequency domain or both for the extraction of pitch.
- Performance of existing methods will be poor for degraded conditions.
- Knowledge of GC events may be used for the extraction of pitch.

2.3 Estimation of Time-Delay

The problem of time-delay estimation has been handled traditionally by exploiting spectral characteristics of speech signals [47, 48]. Three broad strategies used in these studies are [49]: (1) Steered response power of a beamformer, (2) high resolution spectrum estimation, and (3) time difference of arrival estimation. In the steered beamformer the microphone array is steered to various locations to search for a peak in the output power. The delay and sum beamformer shifts the array signals in time to compensate for propagation delays in the arrival of the source signal at each microphone. In this case the signals are time aligned and summed together to form a single output signal. Sophisticated beamformers apply filtering to the array signals before time alignment and summing. These beamformers depend on the spectral content of the source signal. A *priori* knowledge of the independent background noise is used to improve the performance [50].

The second category of time-delay estimators based on high resolution spectrum estimation use spatio-spectral correlation matrix derived from the signals received at the microphones. This matrix is derived using an ensemble average of the signals over the intervals in which noise and speakers are assumed to be stationary, and their estimation parameters are assumed to be fixed [51]. But in the case of speech these assumptions are not valid. These high resolution methods are designed for narrowband stationary signals, and hence it is difficult to apply them for wideband nonstationary signals like speech.

Methods based on Time Differences of Arrival (TDOA) estimation are more suitable for time-delay estimation than the previous two approaches [49]. For accurate estimation of time-delays, weighted Generalized Cross-Correlation (GCC) method is often used [52]. The method relies on the spectral characteristics of the signal. Since the spectral characteristics of the received signal are modified by the multipath propagation in a room, the GCC function is made more robust by deemphasizing the frequency-dependent weightings [53]. Phase transform is one approach where the magnitude spectrum is flattened. However low SNR portions of the spectrum are given equal emphasis as those of high SNR portions. Cepstral prefiltering used to reduce the effects of reverberation, is also difficult to apply for speech signals due to the nonstationary nature of the signal [54]. Moreover, this approach is not suitable for estimating time-delays from short (50-100 ms) segments, which is essential for tracking a moving speaker.

Most of the methods for time-delay estimation rely on spectral characteristics of the speech signal, and the knowledge of degrading noise and environment. The spectrum of the received signal depends on how the waveform gets modified due to distance, noise and reverberation. Therefore, the performance of a time-delay estimation method depends on how the effect of the degrading components is minimized. A summary of the discussion related to the estimation of time-delays is given in Table 2.3.

2.4 Enhancement of Degraded Speech

When speech is transmitted in an acoustical environment like an office room, it will be degraded by background noise and reverberation [55, 56]. Several approaches have been proposed in the literature for enhancement of degraded speech [31, 57–66]. These

Table 2.3: Summary of the review of estimation of time-delays.

- Existing methods rely on the spectral characteristics for the estimation of time-delays.
- Performance depends on how the effect of degradation is minimized in the collected signals.
- Knowledge of excitation source information (GC events) may be used for the estimation of time-delays.

approaches may be broadly classified into single and multichannel cases, depending on whether the speech is collected from a single or multiple microphones.

Enhancement techniques can be grouped into two categories. In one category, attempts are made to cancel the effects of degrading components, and in the other category, attempts are made to enhance the speech components. In the first case, the emphasis is on improving the overall SNR of the degraded speech [57,60,62]. In this case more attention is given to the low SNR regions of speech. When attempting to reduce the effects of degradation in these regions, the natural characteristics of speech are affected, sometimes causing significant distortions. In the second case, the objective is to enhance the speech signal wherever possible, so that the resulting speech is perceived as less noisy and less reverberant, and thus increase the comfort level for listening. This is achieved by identifying and enhancing the high SNR regions [31,63,64].

Knowledge of either the vocal tract system (spectral) features or the excitation source information may be used for speech enhancement. Many of the existing enhancement methods are based on spectral features [57,60,62,67,68]. For instance, one approach is to estimate the spectral features of noise, and subtract these from the spectrum of the degraded speech signal [57]. Changes in the spectral characteristics may introduce audible distortions. Alternatively, manipulation of the excitation source alone may introduce minimal distortion, as spectral components are not affected.

In some earlier methods of using the excitation source information for speech enhancement, LP residual is used to identify and enhance high SNR regions [31,63,64]. These methods were developed for speech enhancement for a single channel case, in which the enhancement is achieved mainly with respect to background noise. Only partial success was achieved in reducing the effects of reverberation. One way to deal with reverberation is to identify different regions in the degraded speech, such as regions with high Signal-to-Reverberant component Ratio (SRR), low SRR and only reverberations and enhance only the high SRR regions [31].

The enhancement of speech may also be achieved using the knowledge of GC events. This is because the high SNR regions of a speech signal are due to the GC events. GC events are known to be robust against environmental degradations [30]. The locations of GC events along the time scale do not change due to degradations. Enhancement against reverberation may be more effective if signals from several microphones are used. A summary of the discussion related to the enhancement of speech is given in Table 2.4.

Table 2.4: Summary of the review of enhancement of speech.

- Estimating the degrading components and minimizing there effect in the degraded speech.
- Difficult to estimate time-varying degradation components.
- Speech specific knowledge may be used for the identification and enhancement of speech.

2.5 Enhancement of Speech in Multispeaker Environment

In multispeaker environments like meetings and discussions, several speakers will be speaking simultaneously. The signal collected by a microphone in such conditions is a mixture of speech from several speakers. Several methods have been proposed for enhancement of speech in a multispeaker environment [69–74]. These methods may be broadly classified into two categories, namely, single channel and multichannel cases. The single channel method is commonly termed as cochannel speaker separation. The implicit assumption in cochannel speaker separation is that there are only two speakers, and between them one is the desired one. In the multichannel case signals from all the microphones are processed to enhance speech of the desired speaker. This approach seems to be inspired by the binaural processing present in humans [74]. In the multichannel case speech of two or more speakers may be enhanced using signals from multiple microphones [73].

Several pitch-based algorithms have been proposed for cochannel speaker separation [69–71]. The assumption made in these studies is that pitch of the desired speaker and that of the interfering speaker are quite distinct, and the pitch contours are resolvable. The speech energy of a particular speaker is concentrated at his/her pitch harmonic frequencies. If the spectrum is sampled at the desired speaker's pitch harmonics, most of the energy of the spectrum samples would correspond to that speaker's voice. After obtaining harmonic amplitudes, the time domain waveform is reproduced using the synthesis algorithm. Harmonic Magnitude Suppression (HMS) technique for speech separation was proposed in [75]. Enhancement of speech of the desired speaker was achieved by estimating the interfering speech spectra and subtracting the same from the combined speech spectra by spectral subtraction approach. Lee and Childers [70] proposed a Minimum-Cross-Entropy Spectral Analysis (MCESA) approach for cochannel speaker separation. The MCESA is an information-theoretic method that simultaneously estimates the power spectrum of one or more independent signals, when a prior estimate of each is available. Quatieri and Danisewicz have proposed a method based on sinusoidal modeling of speech [76]. A least squares estimate algorithm was used to determine the sinusoidal components of each of the speakers, and the speech of the desired speaker was synthesized using the corresponding sinusoidal components. Morgan *et al* [71] have proposed a method for cochannel speaker separation termed as Harmonic Enhancement and Suppression (HES). The pitch of the stronger speaker was estimated first, and it was used for recovering his/her harmonics and formants. The weaker speaker information was obtained after suppressing the stronger speaker harmonics and formants information from the cochannel signal.

A method for enhancing speech of a speaker, while attenuating speech from other speakers using an array of microphones was proposed in [72]. A class of nonlinear processes using a microphone array was proposed, which emphasizes the wanted speech signal relative to the unwanted signals from other locations. The unwanted signals were attenuated and distorted, while the wanted speech signal was unaffected. When the unwanted signal is speech, the distortion makes it less intelligible. The problem of multispeaker speech enhancement in a multichannel case is also termed as Blind Source Separation (BSS). BSS consists of retrieving the source signals without using any *a priori* information about mixing of the signals. It exploits only the information carried by the received signals themselves, hence the term *blind*. Neural network models and learning algorithms for blind signal separation and deconvolution of signals are discussed in [77]. A method for multichannel signal separation using a dynamical recurrent network is proposed in [78]. Estimation of speech embedded in a reverberant environment with multiple sources of noises is proposed in [74,79]. The objective of this work is to make a specific speech signal more intelligible than the available microphone signals. An attempt is made to enhance the signal nearest to the microphones, which is the signal with high energy. This is achieved by mimicking the inner ear, through the use of a bank of self-adaptive band-pass wavelet filters, tracking of the fundamental frequency and by masking some parts of the speech with low energy.

In most of the existing methods knowledge of pitch is used for deriving the information related to each speaker. But reliable estimation of pitch in a multispeaker environment is a difficult task. A summary of the discussion related to the enhancement of speech in multispeaker environment is given in Table 2.5.

2.6 Detection of VOP Events

A method is proposed based on the assumption that the VOP events are characterized by the appearance of rapidly increasing resonance peaks in the amplitude spectrum [80]. A method for detection of VOP events is developed using zero-crossings, energy

Table 2.5: Summary of the review of enhancement of speech in multispeaker environment.

- Estimation of pitch and enhancement of speech desired speaker using the knowledge of the pitch.
- Difficult to estimate pitch in mutlispeaker environment.
- Knowledge of GC events may be used for the enhancement of speech of the desired speaker.

profile and pitch information [81]. The difficulty in using zero-crossings and energy profile for detecting VOP events lies in setting appropriate thresholds. A method based on wavelet transforms is developed for the detection of VOP events [82]. In this method, product function of the wavelet and the energy profile is used for detecting VOP events [82]. A method using energy derivative is proposed for the detection of VOP events [33]. For a given sound unit, the energy and its first derivative are obtained. The instants of maximum energy derivative are hypothesized as VOP events. The limitation of this method is that for some sound units like aspirated sounds and fricatives, the peak in the energy derivative may occur at the onset of aspiration or frication which is ahead of the VOP event. A neural network based approach is also proposed for the detection of the VOP events [83]. The acoustic cues, namely, signal energy, LP residual energy and spectral flatness are used as features in the algorithm. A multilayer perceptron network is trained using the features extracted from these cues to detect the VOP events. The assumption is that there will be significant changes in these acoustic cues in the regions before and after the VOP event. But, for some sound units like semivowels, nasals and aspirated sounds, the change may not be significant.

In all the existing methods, the VOP events are detected by extracting the vocal tract system features at the frame level, which results in poor resolution. A summary of the discussion related to the detection of VOP events is given in Table 2.6.

Table 2.6: Summary of the review of detection of VOP events.

- In the existing methods vocal tract system features are used and block processing is employed.
- Detected VOP events will have poor resolution.
- Excitation source information may also used for the detection of the VOP events.

2.7 Detection of End-points

The need for accurately detecting the end points of a speech utterance is important in many applications like isolated word recognition, connected digit recognition and text-dependent speaker verification [34, 84–90]. In all these applications detection of end-points is the first step for selecting the speech regions in the given utterance. Once the end points are located, feature vectors are extracted from the signal present between these points and used for further processing.

The performance of the system in which end-points detection is used as the first stage, depends critically on the accuracy of detection of the end-points [34, 90]. The computation process is minimum if the end-points are accurately detected. Hence there is a need for an algorithm for accurate detection of end-points. There are many algorithms proposed in the literature for detection of end-points [91–94]. All of them are based mainly on the energy of the speech utterance, and the decision for end-points is made using multiple thresholds. However, deriving appropriate thresholds is difficult under noisy conditions. Some algorithms also use the knowledge of pitch along with energy for end-points detection [95]. A summary of the discussion related to the detection of end-points is given in Table 2.7.

Table 2.7: Summary of the review of detection of end-points.

- Existing methods use knowledge of energy and pitch.
- Extraction of energy and pitch is still a difficult task under degraded conditions.
- VOP events may be used for detection of end-points.

2.8 Summary

We have discussed some issues in the analysis of speech related to the present work. Most of the existing analysis methods (except [27, 43]) employ block processing. The goal of this work is to show that the proposed event-based approach is useful in many applications. A summary of the review of different issues discussed in this chapter is given in Table 2.8.

Table 2.8: Summary of the review of some issues in the analysis of speech.

Speech analysis task	Review of existing methods	Proposed method
Detection of the GC events	Employ block processing except for [27, 43]	Methods proposed in [27, 43] exploit event nature and are discussed
Extraction of pitch	Performance will be poor for degraded conditions	Knowledge of the GC events is used, which are robust to degradations
Estimation of time-delay	Employ block processing and use spectral features	Event-based approach and uses source information derived from GC events
Enhancement of speech	Employ block processing and use spectral features	Based on source information derived from GC events
Detection of the VOP events	Employ block processing and use spectral features representing vocal tract information	Exploits event nature of vowel onset
Detection of end-points	Employ block processing and energy threshold	Knowledge of VOP events

Chapter 3

GLOTTAL CLOSURE EVENT FOR SPEECH ANALYSIS

In Chapter 1, the event nature of speech production and its potential use for an event-based analysis of speech was discussed. The GC and VOP events were chosen to discuss the issues related to the proposed event-based approach. In the previous chapter we reviewed some of the issues in the analysis of speech, which is performed mostly by block processing. Since the objective of this work is to propose methods using the event-based approach, the first step is to detect the events in speech signal so that they can be used as anchor points for further analysis. This chapter discusses the issues involved in the detection of GC events. The GC is an event at the microlevel, around which significant information about excitation source as well as vocal tract system is present. The quasiperiodic nature of occurrence of GC events provides an important perceptual feature for speech, namely pitch. Pitch and associated variations contain important information about speech, speaker and language. Therefore GC events may be used as anchor points for analysis of speech.

This chapter is organized as follows: Various issues involved in the detection of GC events are discussed in Section 3.1. Acoustic cues derived from LP residual that are useful for the detection of GC events are discussed in Section 3.2. In Section 3.3, manual detection of GC events using the proposed acoustic cues is explained. Section 3.4 discusses a method for accurate detection of GC events. There are applications where approximate information about GC events is sufficient for analysis. Section 3.5 discusses the realization of approximate GC event information. The discussion related to the detection of GC events is summarized in Section 3.6.

3.1 Issues in the Detection of GC Events

Although the source of excitation for voiced speech is a sequence of glottal pulses, the significant excitation of the vocal tract system, to a first approximation can be considered to occur at discrete instants of time, called the GC events. Due to this, ideally within each pitch period, the instant prior to the maximum amplitude of the speech signal may correspond to the GC event. However responses due to successive excitations overlap, forming a composite signal, which makes the detection of the GC event difficult in the speech signal. For instance, segments of speech signals of vowel /i/ spoken by different speakers are shown in Figure 3.1. The question (**?**) marks indicate the regions of ambiguity for marking the GC events.

As it is difficult to identify the GC events directly from the speech signal, the other mostly used approach is to inverse filter the speech signal. The parameters of the inverse filter are obtained by LP analysis [96]. A brief discussion on LP analysis is given in Appendix–B. In LP analysis, the voiced speech is assumed to be the output of an all-pole filter. Hence, the prediction will be good at all places except the GC events, due to which the output of the inverse filter shows large error around the GC events. Figure 3.2 shows the LP residuals for the speech segments shown in Figure 3.1.

Although the LP residual contains information pertaining to the excitation, identification of GC events directly from the LP residual is not recommended due to the following problems [27]: LP analysis assumes an all-pole model for representing the combined effect of impulse response of the vocal tract system and the glottal pulse shape. The all-pole model implicitly assumes a minimum phase characteristic of the speech signal. If this is not valid, the phase response of the vocal tract system is not compensated exactly by the inverse filter. Phase compensation will also be affected when formants and their bandwidths are not estimated accurately. Effect of uncompensated phase on LP residual is not known. Moreover, the inverse filter does not

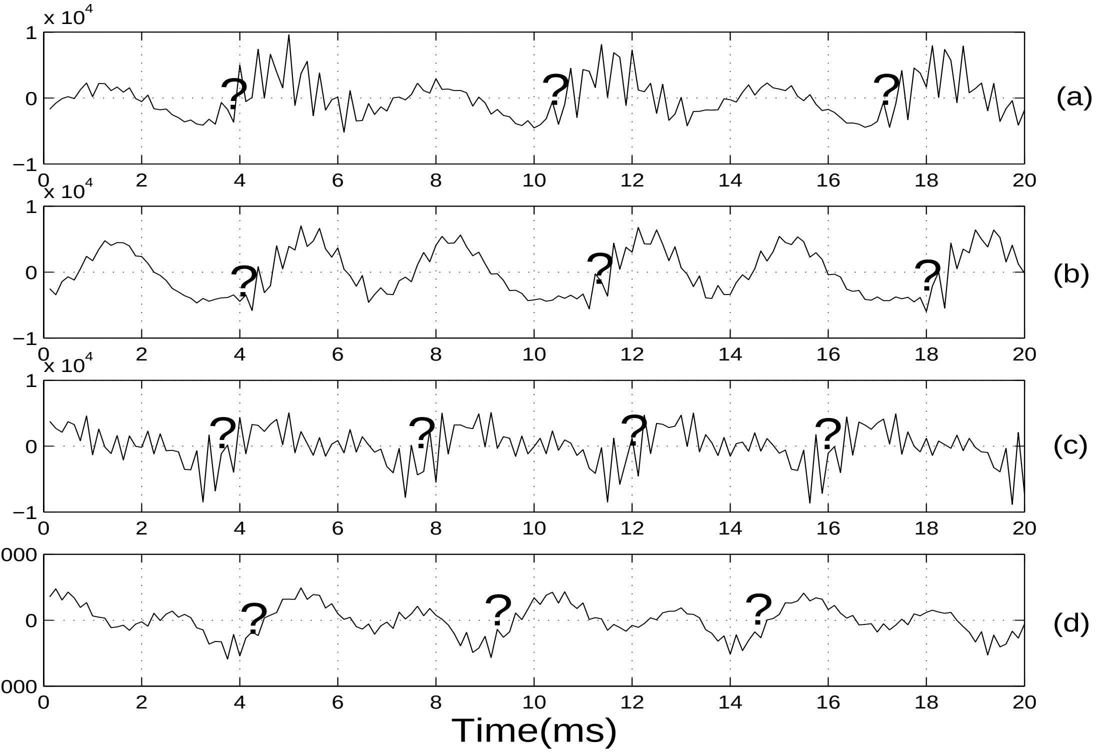

Figure 3.1: Speech segments of vowel /i/ of two male ((a) and (b)) and two female ((c) and (d)) speakers. The question (**?**) marks indicate the regions of ambiguity for marking the GC events.

compensate for zeros which may be introduced due to the finite duration of glottal pulse or the nasal coupling. These factors cause multiple peaks of either polarity to occur in the LP residual, and make the estimation of the epochs from the LP residual difficult. The presence of multiple peaks of either polarity around the GC events is shown by question (**?**) marks in Figure 3.2.

There are sounds like nasals and voiced stop consonants where it is more difficult to detect the GC events from the LP residual. In the case of nasals like /m/, due to poor modeling by LP analysis, the inverse filter does not compensate for zeros introduced due to the nasal coupling. As a result, peaks of either polarity occur around the GC events. This makes it difficult to detect the GC events unambiguously from the LP residual. This is illustrated for nasal /m/, both for male and female speakers in Figure 3.3. In the low voiced stop consonants like /b/, it is even difficult to define the GC events, let alone detect it. Figure 3.4 shows speech signals and the corresponding LP residuals for voiced consonant /b/. Due to loading of the vocal tract system, the

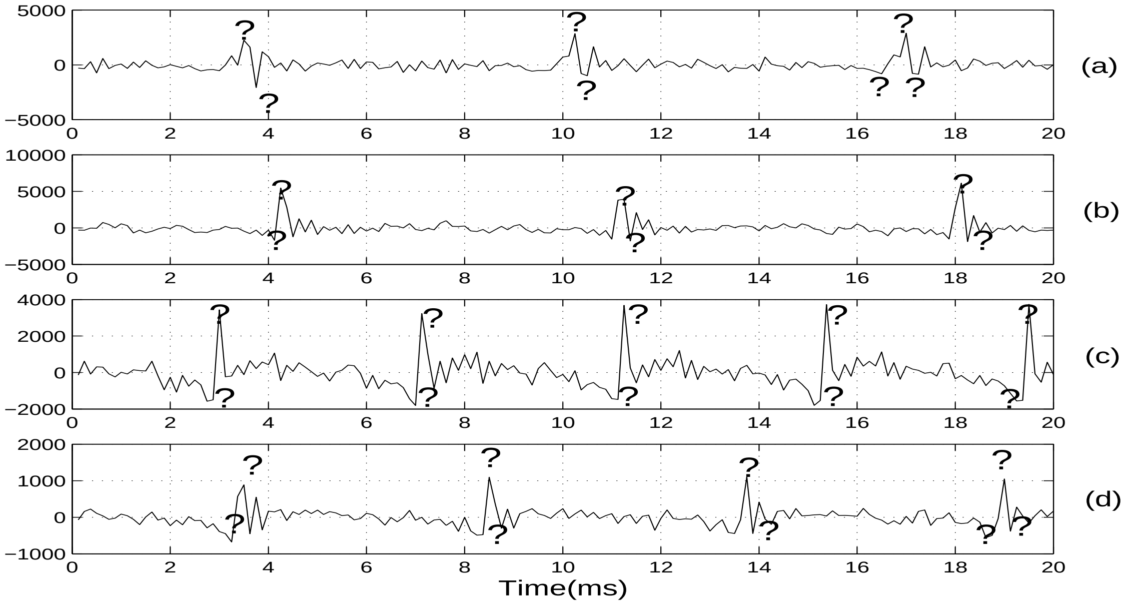

Figure 3.2: LP residuals of the speech segments of vowel /i/ shown in Figure 3.1. The question (**?**) marks indicate the regions of ambiguity encountered for marking the GC events.

strength of excitation at the GC events is comparable to that at other places.

Further the ambiguity for the detection of the GC events is more in the case of female speakers. This may be attributed to the following factors: The following factors may be attributed for this: As the pitch frequency is higher, the assumption of quasistationary for analysis of speech over segments of 10-30 ms is no longer valid. Hence the estimation of characteristics of the vocal tract will be poor. The vocal folds will be vibrating at a faster rate, and hence the closed phase interval is minimum, and even nonexistent in some cases. To make this happen the suction pressure with which the vocal folds will be closing is low. Hence the strength of excitation will be weak, which in turn produces low error at the GC events in the LP residual.

3.2 Acoustic Cues for the Detection of GC Events

Multiple peaks of either polarity are present around the GC events, thus causing ambiguity in the detection of GC events directly from the residual. Hence some acoustic cues derived from the LP residual in which the evidence about GC events is less ambiguous, are essential for detecting the GC events. This section discusses the acoustic

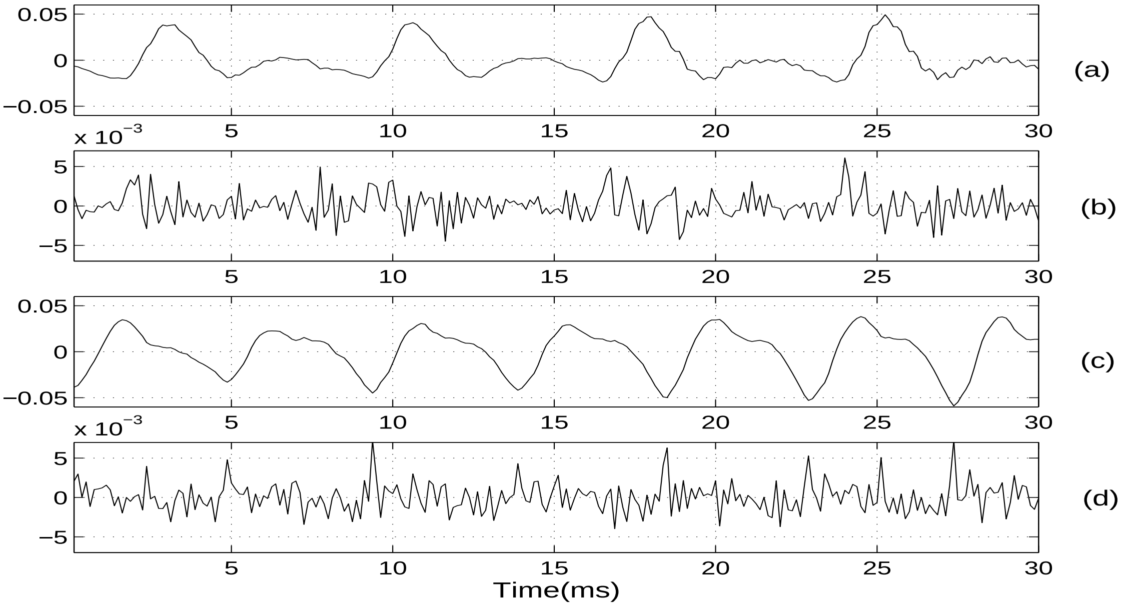

Figure 3.3: Speech segments and LP residuals of nasal /m/ of male ((a) and (b)) and female ((c) and (d)) speakers.

cues, namely, magnitude of LP residual, energy of LP residual, LP residual of low pass filtered speech and Hilbert envelope of LP residual, which are useful for the detection of GC events.

3.2.1 Magnitude of LP Residual

The magnitude of LP residual approximately represents the strength of excitation and hence may be used as an acoustic cue for detecting the GC events. Speech segment of vowel /u/, the corresponding LP residual and magnitude of LP residual are shown in Figure 3.5. The ambiguity due to peaks of either polarity is minimized. The region of GC events may be detected by considering the peaks in the magnitude plot.

3.2.2 Energy of LP Residual

A smoothed version of magnitude of LP residual may be obtained by computing the energy of the LP residual. Energy is computed by considering frames of smaller size, typically, 1 ms, with a shift of one sample. Peaks in the energy plot approximately indicate the location of the GC events. Figure 3.6 shows a segment of vowel /u/,

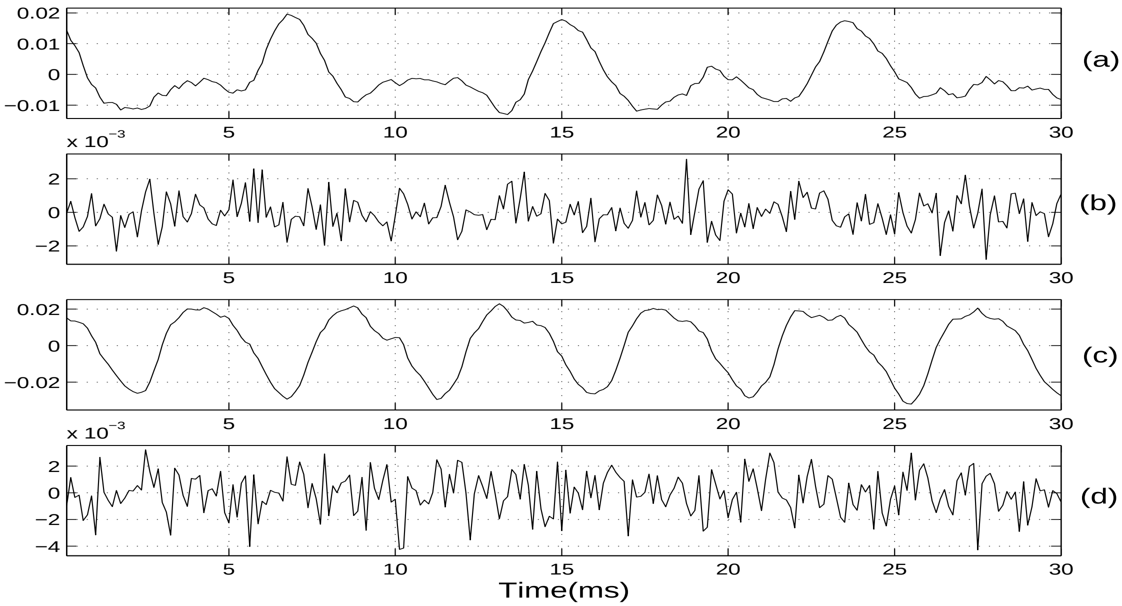

Figure 3.4: Speech segments and LP residuals of voiced stop consonant /b/ of male ((a) and (b)), and female ((c) and (d)) speakers.

corresponding residual and the energy plot. The ambiguity in the energy plot is less as it is a smoothed version of the magnitude. However, the resolution is still poor for the detection of GC events. This is because energy is computed over a frame which only indicates a region over which the GC event is present.

3.2.3 LP Residual of Low Pass Filtered Speech

In each pitch period, the region around GC event is known to be high SNR region. The ambiguity for the detection of GC events may be reduced by eliminating the variations present in low SNR regions. As the high SNR components in speech are present upto 2 kHz, the speech signal may be low pass filtered with a cut-off frequency of 2 kHz, and the LP residual may be computed from the low pass filtered speech. Figure 3.7 shows a segment of speech of vowel /u/, its LP residual and the LP residual computed from the speech low pass filtered using a cut-off frequency of 2 kHz. It can be seen that the ambiguity for detection of GC instants is less in the LP residual of the low pass filtered speech, when compared to the LP residual of original speech.

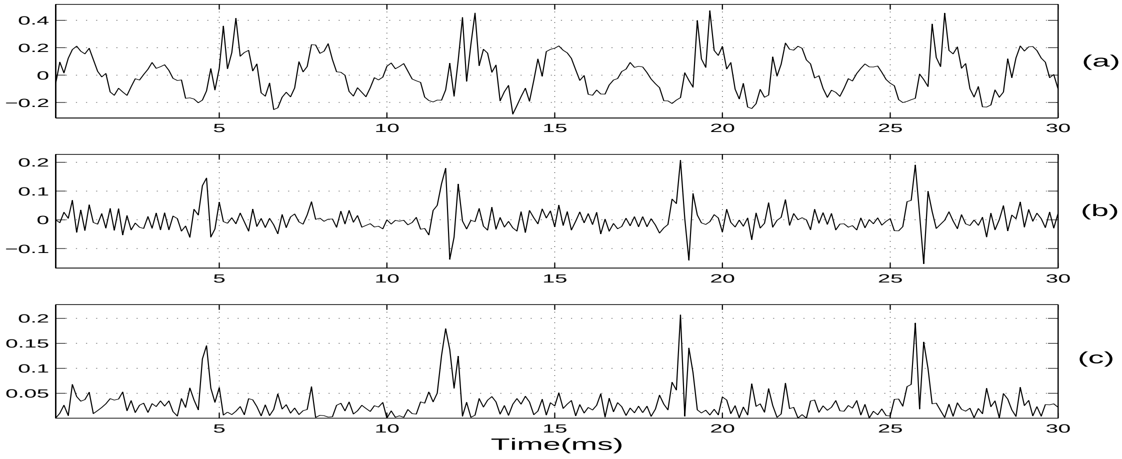

Figure 3.5: (a) Speech segment of vowel /u/, corresponding (b) LP residual and (c) magnitude of LP residual.

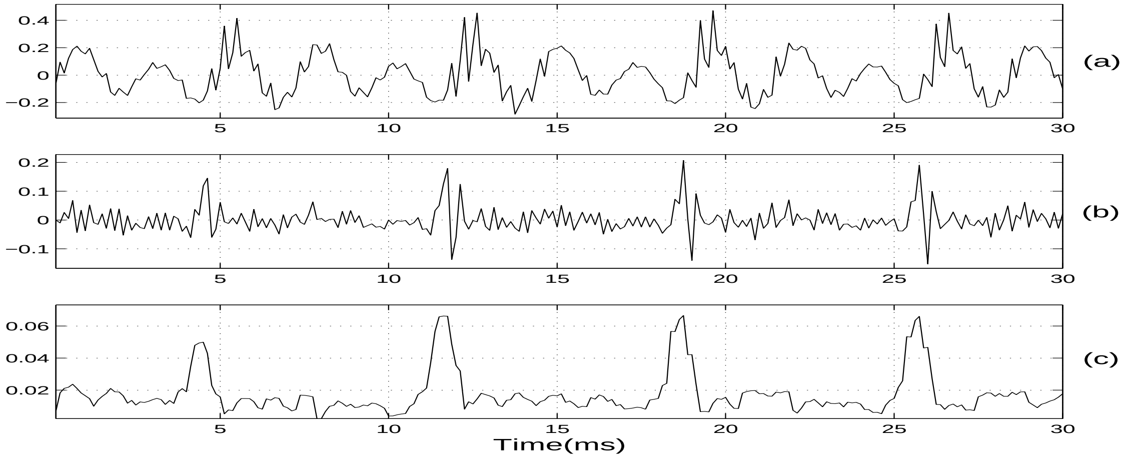

Figure 3.6: (a) Speech segment of vowel /u/, corresponding (b) residual and (c) energy.

3.2.4 Hilbert Envelope of LP Residual

A better method to detect GC events is to exploit the property that the GC events are impulse-like excitations, and the strength of excitation in voiced speech is large around the GC event. This can be seen by computing the energy in short (1 ms) intervals of the residual. Ideally it is desirable to derive an impulse-like signal around the GC event. A close approximation to this is possible by using Hilbert envelope of the LP residual, instead of the energy in short intervals of time. Even though the real and imaginary parts of an analytic signal (related through the Hilbert transform) have positive and negative samples, the Hilbert envelope of the signal is a positive

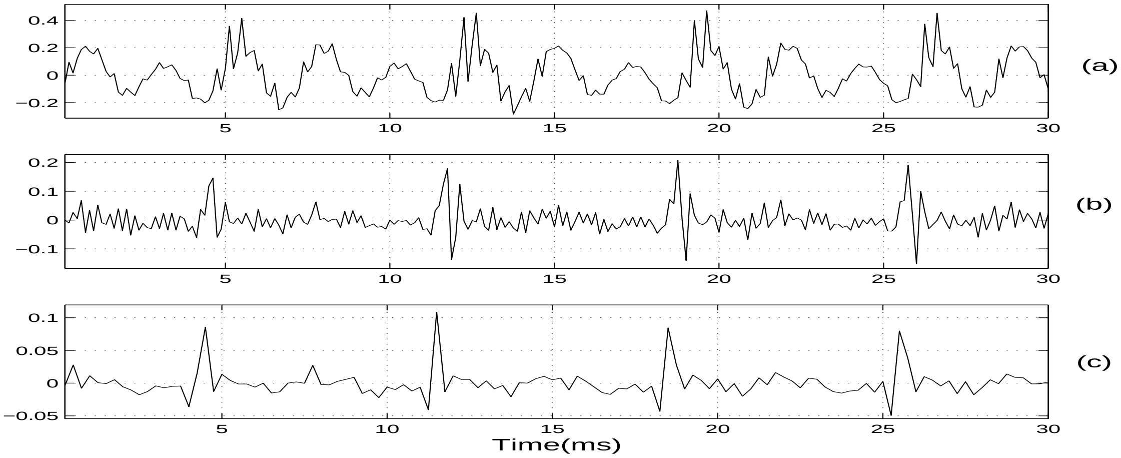

Figure 3.7: (a) Speech segment of vowel /u/, corresponding (b) residual, and (c) residual of low pass filtered speech (cut-off frequency of 2 kHz).

function, giving the envelope of the signal [97]. For example, the Hilbert envelope of a unit sample sequence or its derivative has a peak at the same instant. Thus the properties of Hilbert envelope can be exploited to derive the impulse-like characteristics of the GC events. The Hilbert envelope $h_e(n)$ of the LP residual $e(n)$ is defined as follows [27, 30, 97]:

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{3.1}$$

where $e_h(n)$ is the Hilbert transform of $e(n)$, and is given by [27]

$$e_h(n) = \left\{ \begin{array}{rl} IDFT[-jE(\omega)], & 0 < \omega < \pi \\ IDFT[jE(\omega)], & 0 > \omega > -\pi \\ 0 & \omega = 0, \pi \end{array} \right\} \tag{3.2}$$

where IDFT is the Inverse Discrete Fourier Transform, and $E(\omega)$ is the discrete Fourier transform of $e(n)$. A discussion on the Hilbert transform relations is given in Appendix-C.

Figure 3.8 shows a segment of LP residual for vowel /u/, its Hilbert transform and the Hilbert envelope. The peaks in the Hilbert envelope indicate the epoch locations. Also, as there is no smoothing involved in eliminating multiple peaks around the epochs, the GC events are detected with high resolution.

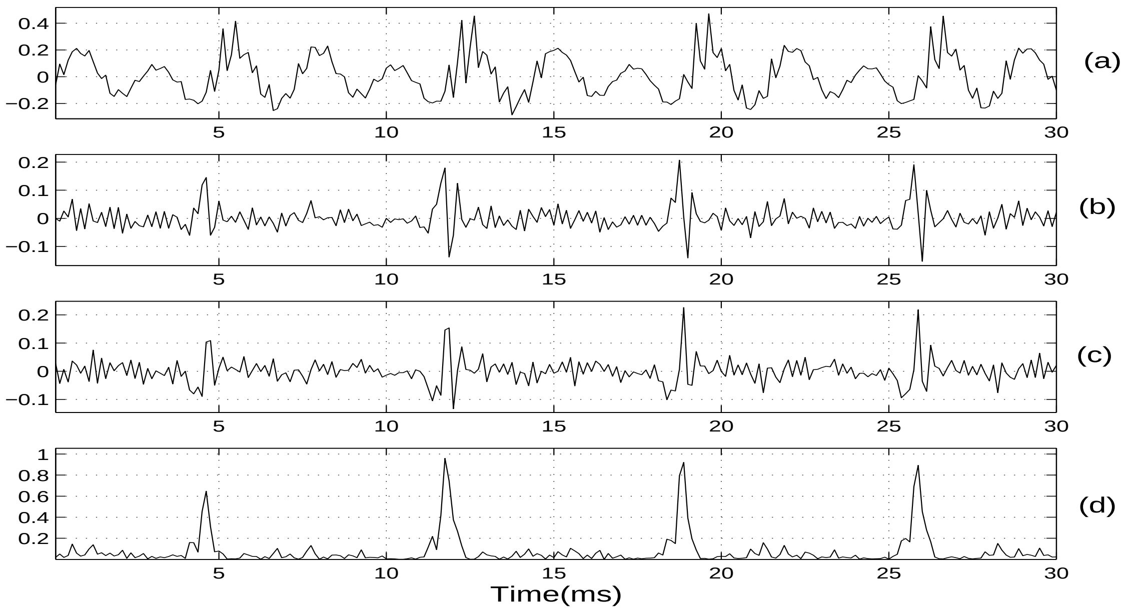

Figure 3.8: (a) Speech segment of vowel /u/, corresponding (b) residual, (c) Hilbert transform of the LP residual and (d) Hilbert envelope of the LP residual.

3.2.5 Summary of the Acoustic Cues

The magnitude and energy of LP residual indicate only the region over which the GC event may be present, but the GC event is associated with an instant. The LP residual of the low pass filtered speech will also provide poor resolution due to downsampling. The Hilbert envelope of LP residual provides high resolution compared to other acoustic cues in most of the cases. Thus, acoustic cues such as magnitude of LP residual, energy of LP residual and LP residual of low pass filtered speech may be used to identify the approximate region of the GC events, and the Hilbert envelope of the LP residual may be used to mark the GC event.

3.3 Manual Detection of GC Events

The objectives of manual detection of the GC events are: (1) to understand the difficulties involved in the detection of GC events, (2) to observe the signal characteristics around the GC events and (3) to identify the acoustic cues which may be useful for automatic detection of the GC events. Manually, the GC events are detected as the instants at which most of the cues show peak in each pitch period. In particular, magnitude of LP residual, energy of LP residual and LP residual of low pass filtered speech are used for initial identification of the region of the GC events. Finally, the GC events are marked by referring to the peaks in the Hilbert envelope of the LP residual in these regions.

Speech segment of vowel /i/, its LP residual and the proposed acoustic cues derived from the LP residual are shown in Figure 3.9. The GC events are marked by referring to the acoustic cues. It is interesting to note that even though information about the GC events are manifested well in all the cues, the resolution in the case of Hilbert envelope of LP residual is high. Figures 3.10 and 3.11 show speech segments of nasal /m/, their residual and the acoustic cues computed for male and female speakers, respectively. In this case, the manifestation of epoch information in the magnitude of LP residual is poor due to poor modeling. Energy of LP residual and LP residual of the low pass filtered speech indicate the region over which the GC events are present. The GC events are marked by referring to the Hilbert envelope of LP residual.

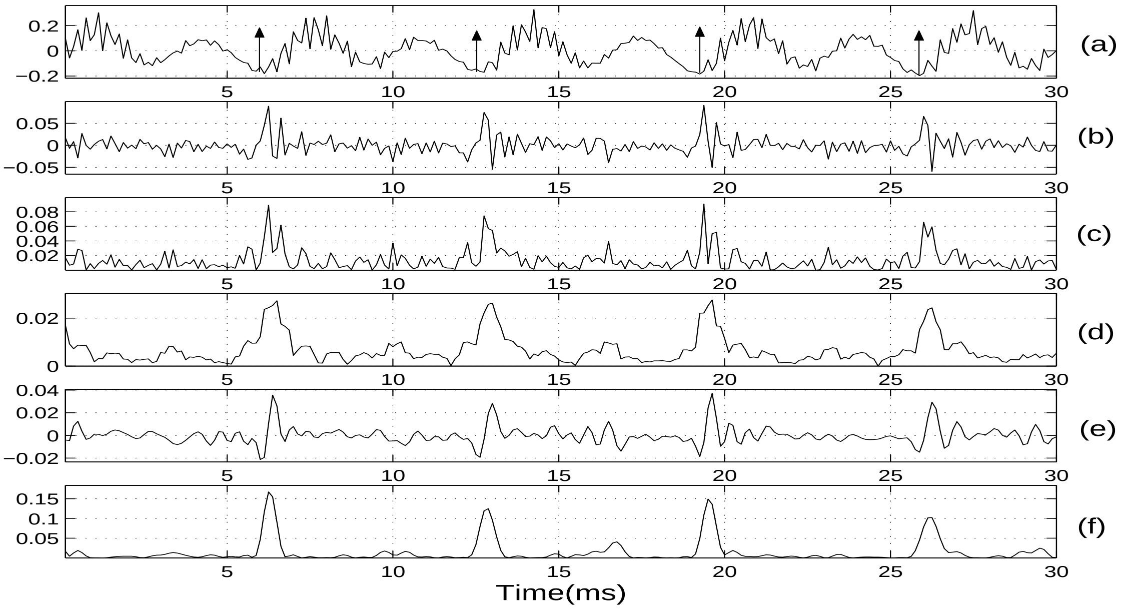

Figure 3.9: (a) Speech segment of vowel /i/ with manually marked GC events (shown by ↑), and its (b) LP residual, (c) magnitude of the LP residual, (d) energy of the LP residual, (e) LP residual of the low pass filtered speech and (f) Hilbert envelope of the LP residual.

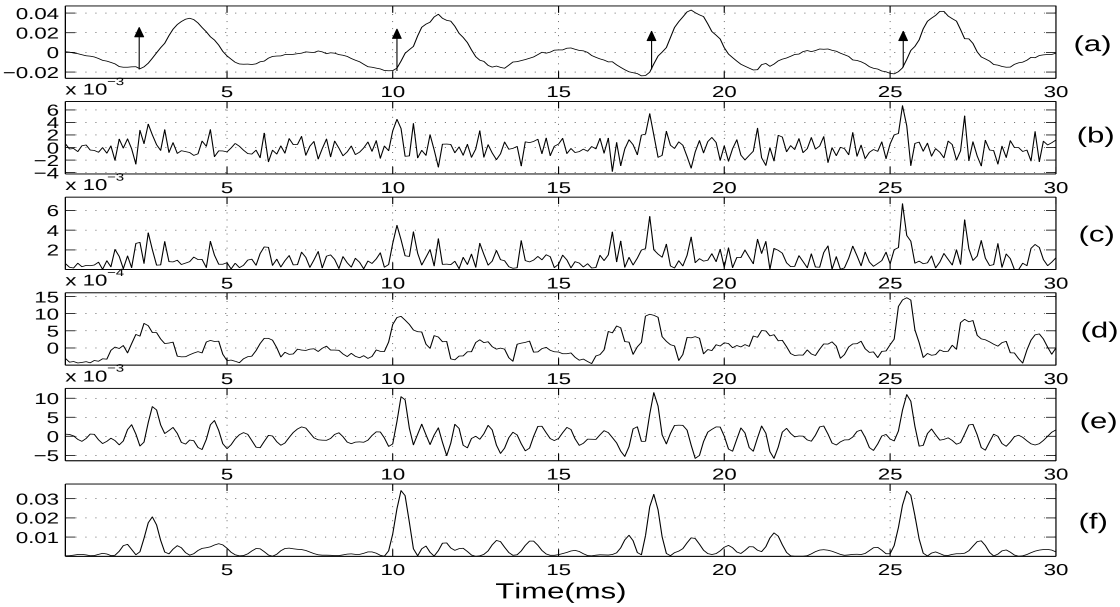

Figure 3.10: (a) Speech segment of nasal /m/ of male speaker with manually marked GC events (shown by ↑), and its (b) LP residual, (c) magnitude of the LP residual, (d) energy of the LP residual, (e) LP residual of the low pass filtered speech and (f) Hilbert envelope of the LP residual.

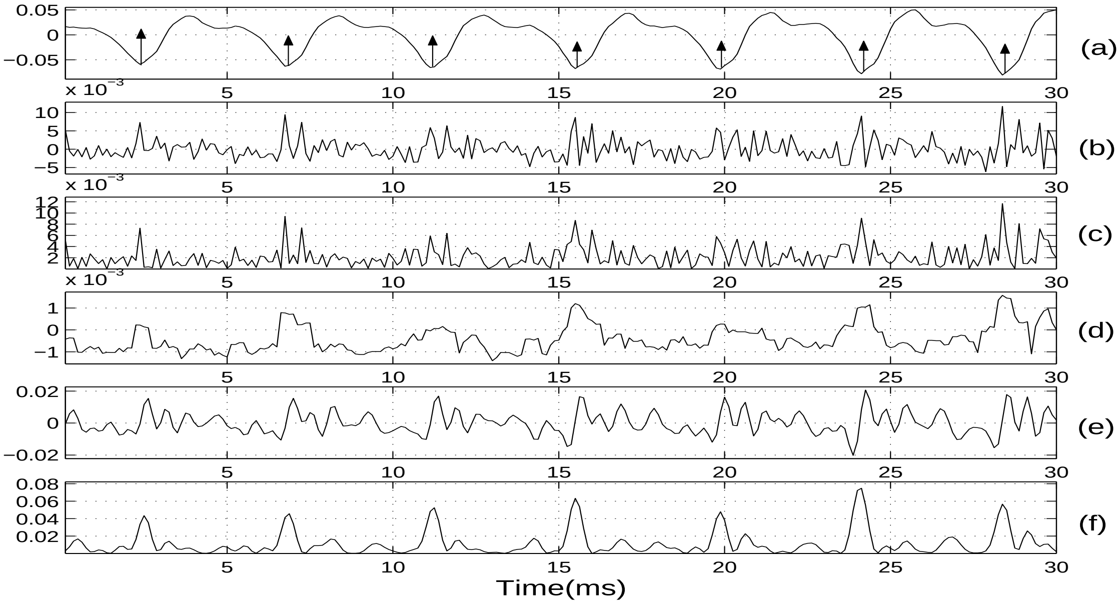

Figure 3.11: (a) Speech segment of nasal /m/ of female speaker with manually marked GC events (shown by ↑), and its (b) LP residual, (c) magnitude of the LP residual, (d) energy of the LP residual, (e) LP residual of the low pass filtered speech and (f) Hilbert envelope of the LP residual.

3.4 Automatic Detection of the GC Events

This section discusses a method for automatically detecting the GC events from voiced speech using the group delay functions proposed in [42, 43]. The method is based on global phase characteristics of minimum phase signals. The average slope of the unwrapped phase of short-time Fourier Transform (FT) of LP residual is computed as a function of time. This average slope obtained as a function of time is termed as the phase slope function. Instants where the phase slope function makes a positive zero crossing are identified as the GC events.

Consider a unit sample sequence delayed by τ samples. The FT of the sequence is $exp(-j\omega\tau)$. The FT phase function is $\phi(\omega) = -\omega\tau$ and its negative derivative is $-\phi'(\omega) = \tau$. Thus the phase function has a constant slope which corresponds to the delay of the unit sample in the time domain. Let us assume an analysis window enclosing the unit sample. As the window is moved to the right or left, the delay of the unit sample changes with respect to the position of the window. The average value of the negative derivative of the FT phase (group delay function) varies linearly with the position of the window. The instant at which the phase slope function crosses zero is identified as the delay of the unit sample in time domain.

Now consider a delayed damped sinusoid. The average value of the derivative of the phase (phase slope) is equal to the delay of the window. As the analysis window is moved, the phase slope value varies linearly with time. In general a minimum phase signal starting at time $t = 0$ has the property that its average value of the unwrapped FT phase spectrum is zero. If the signal is delayed, then the average slope of the phase spectrum is proportional to this delay. This is the basis for the proposed method for the detection of the GC events.

If $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of the windowed signal $x(n)$ and $nx(n)$, respectively, then the group delay $(-\phi'(\omega))$ is given by [98]

$$-\phi'(\omega) = \tau(\omega) = \frac{X_R Y_R + X_I Y_I}{X_R^2 + X_I^2} \tag{3.3}$$

where $X_R + jX_I = X(\omega)$ and $Y_R + jY_I = Y(\omega)$. Isolated peaks in $\tau(\omega)$ are removed by using a three-point median filter. The average value of the smoothed $\tau(w)$ is computed. The resulting phase slope function is computed by moving the analysis window by one sample at a time. The positive zero crossing instants of the phase slope function correspond to the instants of significant excitation. The steps involved in the detection of GC events are illustrated for a segment of vowel /i/ in Figure 3.12. The algorithm for determining the GC events is given in Table 3.1.

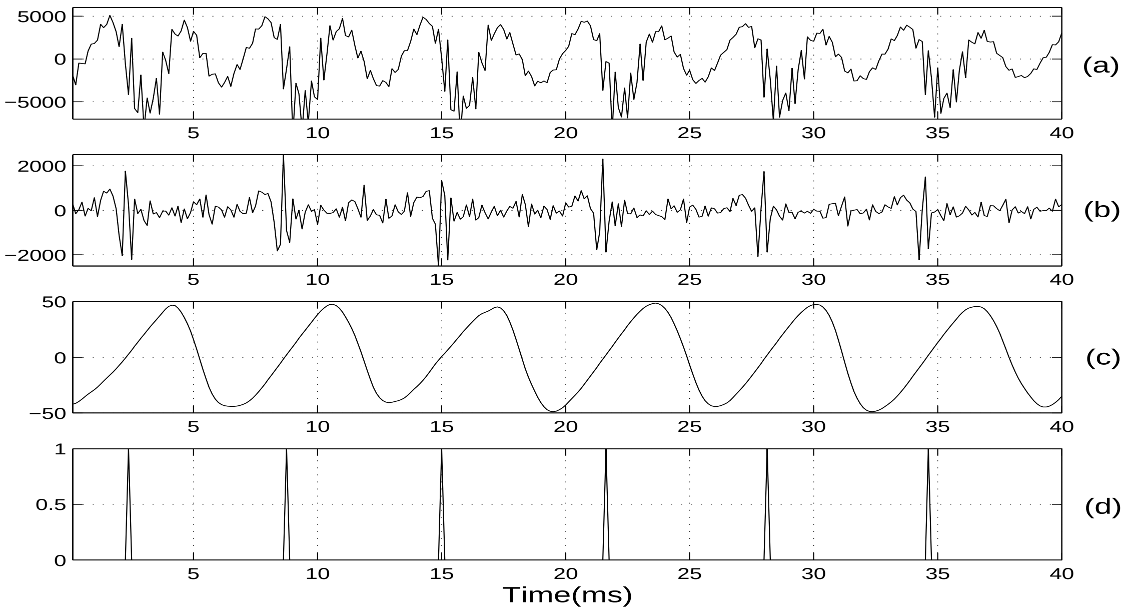

Figure 3.12: (a) Waveform of speech segment of vowel /i/, its (b) LP residual, (c) phase slope function and (d) extracted GC events.

In the present study, after determining the instants of significant excitation, some of the spurious instants are eliminated by using the algorithm given in Table 3.2. The spurious instants correspond to noise excitations in nonspeech region, excitations like onset of burst in unvoiced speech, and secondary excitations (instants of glottal opening) within a pitch period in voiced speech. Figure 3.13 shows the utterance of unvoiced aspirated velar consonant vowel /kha/, its LP residual, phase slope function, extracted instants and the instants after removing the spurious ones.

The algorithm is found to be robust, and is capable of detecting the GC events accurately even in continuous speech with some degradation. For instance, a continuous speech signal and the detected GC events are shown in Figure 3.14. All the GC events are detected accurately. The detected GC events for a speech signal degraded by background noise and reverberation are shown in Figure 3.15. Even under degradation the algorithm detects the GC events accurately, and there are no missing GC

Table 3.1: Algorithm for extracting the GC events.

1. Preemphasize the speech signal.
2. Compute the LP residual using a frame size of 10 ms, frame shift of 5 ms and 10^{th} order LP analysis.
3. Compute the group delay $(-\phi'(w))$ for each frame of the residual signal of size 10 ms and frame shift of one sample using the relation
   $-\phi'(\omega) = \tau(\omega) = \frac{X_R Y_R + X_I Y_I}{X_R^2 + X_I^2}$
   where $X_R + jX_I = X(\omega)$ and $Y_R + jY_I = Y(\omega)$
   $X(\omega)$ is the Fourier transform of $x(n)$ whose group delay is required and $Y(\omega)$ is the Fourier transform of $nx(n)$.
   $n = 0, 1, 2, \cdots, N$ and $N - 1$ is length of $x(n)$.
4. Smooth the group delay function using a median filter of order 3 for removing the unwanted peaks.
5. Find the average group delay for each frame which is the required phase slope function.
6. Smooth the phase slope function with a Hamming window of order 8.
7. Identify the positive zero crossings as GC events.

events.

Table 3.2: Algorithm for removing the spurious instants.

1. Eliminating spurious instants in nonspeech region:
   – Compute frame energies of the speech signal.
   – Eliminate instants in the frames having energy
   less than 30 dB of the maximum frame energy.
2. Eliminating spurious instants in speech region:
   LEVEL-I:
   – Compute the strength of instants using
   Hilbert Envelope of the LP residual. The amplitude of
   Hilbert envelope at the given instant is its strength
   – Eliminate the present instant, if its strength is less than
   the strengths of both previous as well as next instants, and its
   value is less than 0.25 times the maximum strength.
   LEVEL-II:
   – Compute the epoch intervals, which is the time difference
   between successive GC events.
   – Compute average value of the epoch intervals.
   – Eliminate the present instant, if its interval with respect to
   previous as well as next instant is less than 0.7 times
   the average epoch interval.

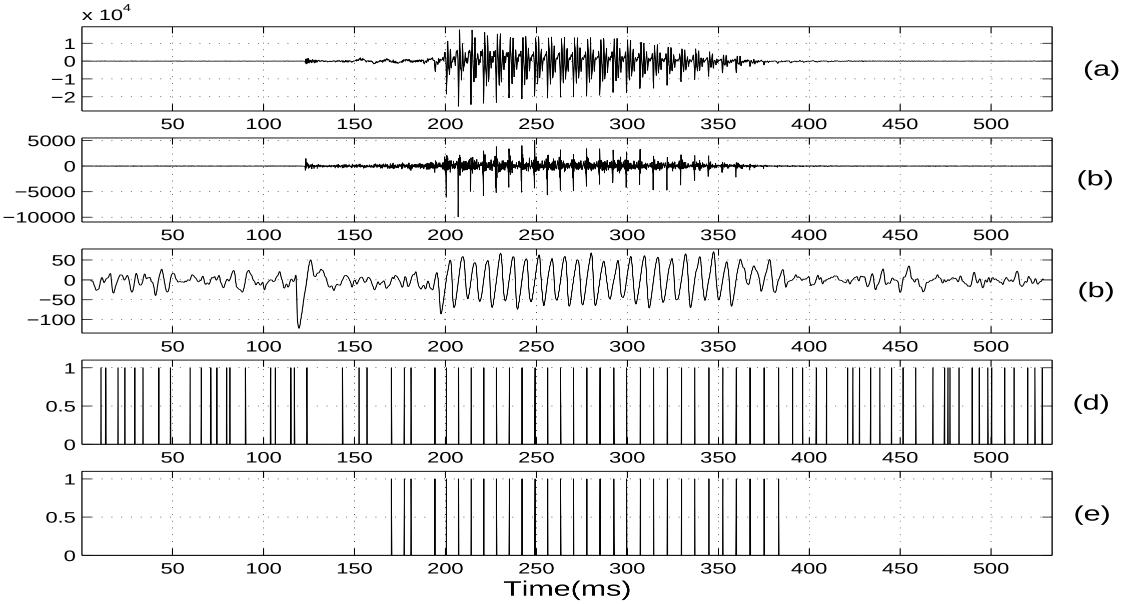

Figure 3.13: (a) Speech signal of /kha/ and its, (b) LP residual, (c) phase slope function, (d) extracted instants of significant excitation and (e) GC events after eliminating the spurious ones.

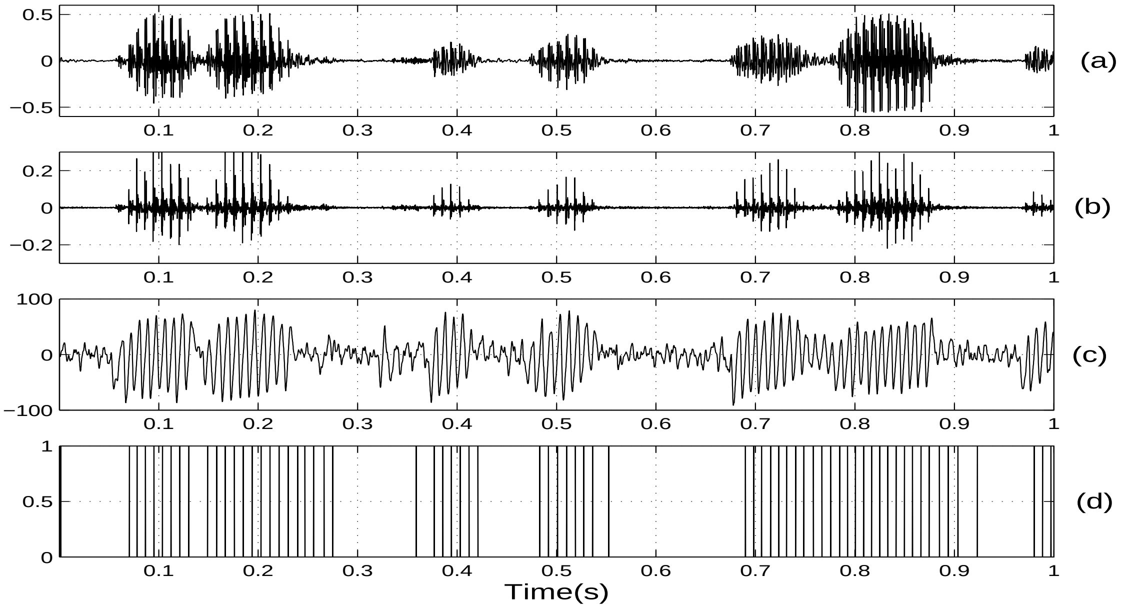

Figure 3.14: (a) Continuous speech signal and its (b) LP residual, (c) phase slope function and (d) GC events after eliminating the spurious ones.

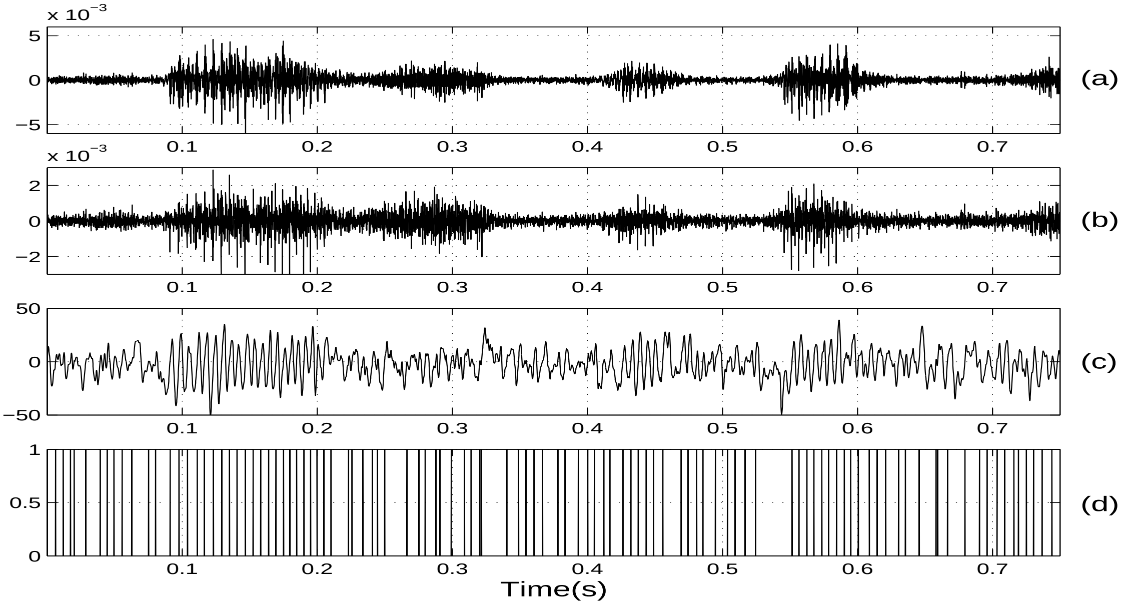

Figure 3.15: (a) Degraded speech signal and its (b) LP residual, (c) phase slope function and (d) GC events after eliminating the spurious ones.

3.5 Approximate Information of GC Events

In the previous section a method for accurate detection of GC events was discussed. However, there are many applications in which even approximate location of GC events, but extracted in a computationally efficient manner, may be sufficient. Among the acoustic cues proposed for the detection of GC events, Hilbert envelope of LP residual provides better resolution, and the peaks in the Hilbert envelope indicate the approximate locations of GC events. For instance, for a segment of vowel /i/, the GC events detected by group delay based approach and the GC events from Hilbert envelope of LP residual are shown in Figure 3.16. For comparison, the GC events detected by the group delay based approach are weighted by their strength. As it can be seen from the figure, the peaks of the Hilbert envelope correspond to the region of GC events.

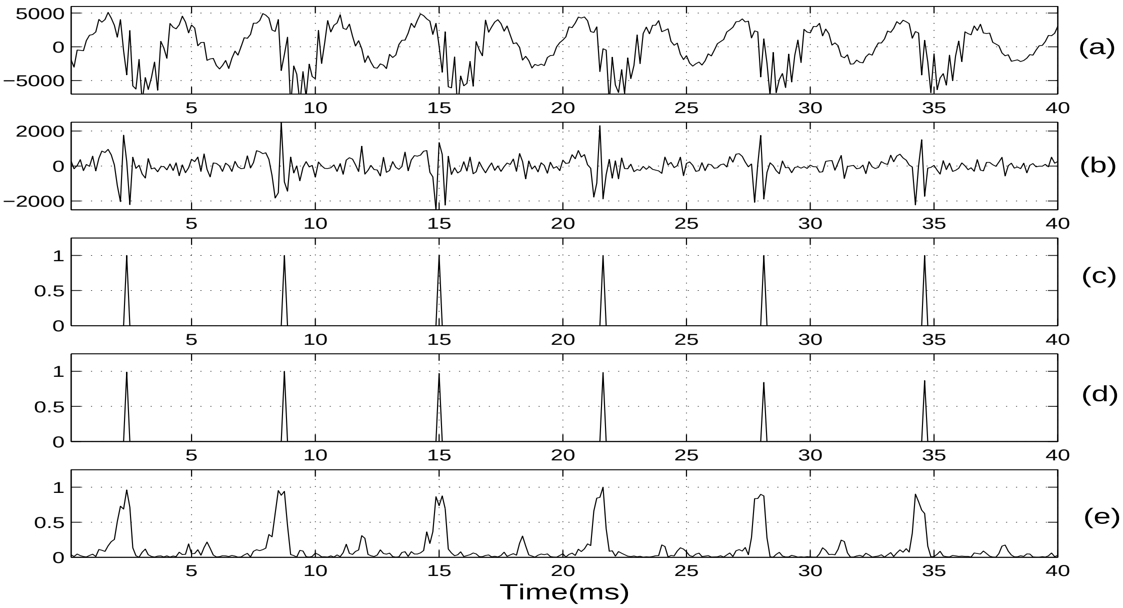

Figure 3.16: (a) Segment of vowel /i/, (b) LP residual, (c) GC events detected by the group delay based approach, (d) GC events weighted by their strength and (e) Hilbert envelope of the LP residual.

The Hilbert envelope of LP residual is robust, as it detects the GC events even for degraded speech. A speech signal, the detected GC events using the group delay approach and Hilbert envelope of LP residual are shown in Figure 3.17. The Hilbert envelope of LP residual indicates that the GC events appear near the correct locations determined by the group delay based approach. The detected GC events for degraded speech are shown in Figure 3.18. It is interesting to note that even in the case of degradation, the GC events are detected in the Hilbert envelope of the LP residual with high resolution.

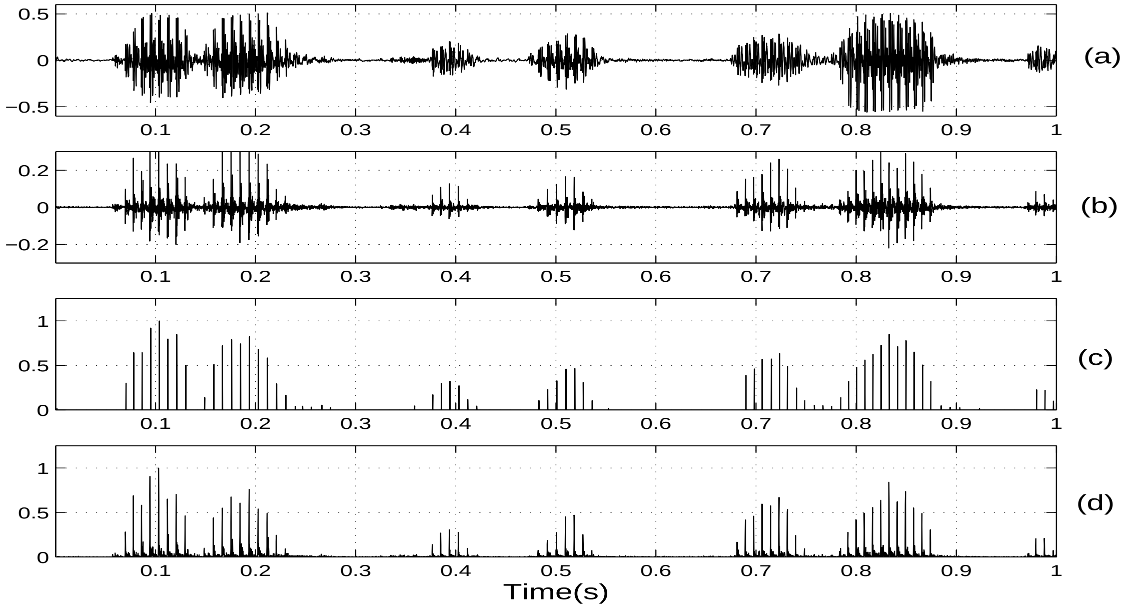

Figure 3.17: (a) Continuous speech signal, (b) LP residual, (c) GC events detected by the group delay based approach and (d) Hilbert envelope of the LP residual.

As will be discussed in the following chapters, approximate information of GC events is sufficient for several applications like the extraction of pitch, estimation of time-delay, enhancement of degraded speech and enhancement of speech in multi-speaker environment. It is interesting to note that various properties of the excitation source and the vocal tract system can be studied knowing even the approximate information of GC events. In the present work, we use the group delay based approach in applications where accurate location of GC events is needed, and Hilbert envelope of LP residual for applications where approximate locations of GC events are sufficient.

3.6 Summary

In this chapter, issues involved in the detection of GC events were discussed. Some acoustic cues derived from LP residual of speech signal were examined. Manual mark-

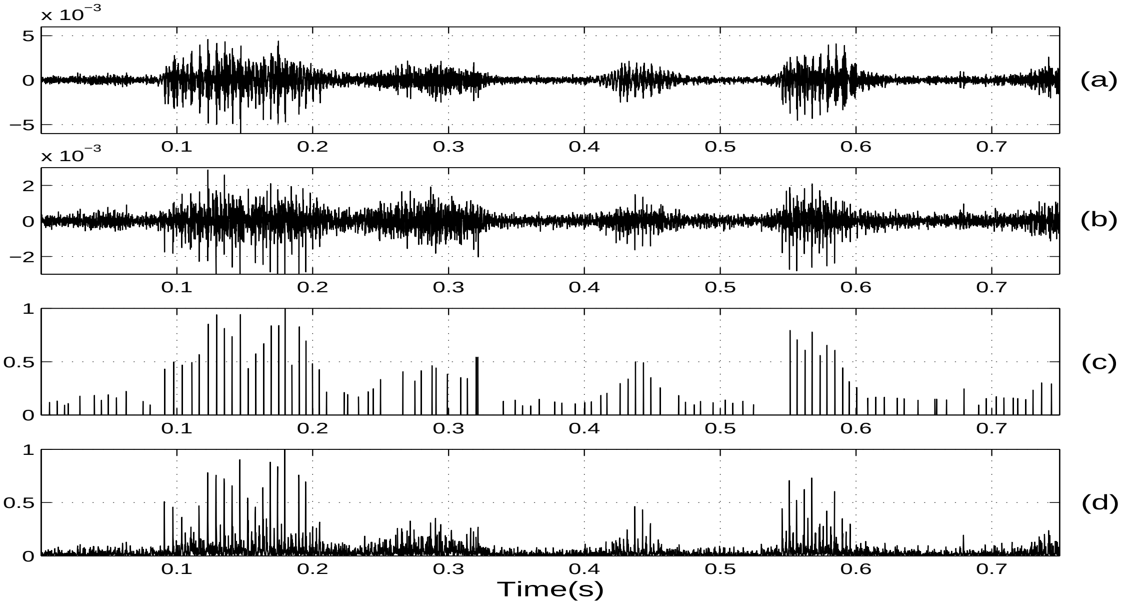

Figure 3.18: (a) Degraded speech signal, (b) LP residual, (c) GC events detected by the group delay based approach and (d) Hilbert envelope of the LP residual.

ing of GC events using the proposed acoustic cues was performed to evaluate the proposed acoustic cues. Among the different acoustic cues proposed, Hilbert envelope of LP residual detects GC events with highest resolution. A method for automatic detection of GC events based on the property of minimum phase signals and group delay functions was discussed. Finally a method for the realization of the approximate epoch information was explained. Summary of the various issues discussed in this chapter is given in Table 3.3.

In the following two chapters, we discuss some of the applications of GC events.

Table 3.3: Summary of the issues discussed with respect to detection of GC events.

GC Event for Speech Analysis

- Issues in the detection of GC events
  - During speech production, responses due to successive GC events overlap to form a composite signal and hence detection of GC events directly from the speech signal is difficult.
  - Peaks of either polarity are present in LP residual and hence unambiguous detection of the GC events from LP residual is difficult.
  - The detection of GC events from LP residual is also difficult due to the low strength of glottal excitation and poor modeling of the vocal tract system in nasals and voiced stop consonants, especially in female speakers.
- Acoustic cues for detection of GC events
  - Magnitude, energy and LP residual of low pass filtered speech gives information about the region of GC events.
  - Hilbert envelope of LP residual provides high resolution for detection of GC events.
- Manual detection of GC events
  - Instants at which most of the acoustic cues show maximum value in a pitch period.
- Automatic detection of GC events
  - Positive zero-crossings in phase slope function derived by the group delay analysis on LP residual, are detected as GC events.
- Approximate epoch information
  - Hilbert envelope of LP residual may be used in tasks where approximate information about GC events is sufficient.

Chapter 4

APPLICATIONS OF GC EVENTS FOR SINGLE CHANNEL CASE

In the previous chapter, issues involved in detection of GC events and methods for the detection of GC events were discussed. In this chapter and the following chapter, we discuss some applications of GC events. This chapter deals with two applications of GC events in which the speech data is collected with a single microphone (single channel). They are extraction of pitch and enhancement of degraded speech.

4.1 Introduction

A method for extraction of pitch in adverse conditions is proposed. Real environment, in which the degradation is due to several unpredictable sources like background noise, reverberation and channel noise, is treated as adverse condition. The proposed method is based on the knowledge of GC events. Hilbert envelope of LP residual gives information about the location of GC events. Autocorrelation analysis is performed on Hilbert envelope of LP residual. The properties of Hilbert envelope of LP residual are exploited for extraction of pitch from the autocorrelation sequence.

A method for enhancement of degraded speech collected over a single channel is proposed. The proposed method is suitable for speech collected from a severely degraded channel. The degraded speech is processed by LP analysis for deriving Hilbert envelope of LP residual. The property of Hilbert envelope of LP residual in the autocorrelation sequence is exploited to derive a weight function. LP residual of the degraded speech is multiplied with the weight function to enhance the excitation regions of the speech. Speech signal synthesized using the modified LP residual is found to be perceptually enhanced significantly.

This chapter is organized as follows: In Section 4.2 a method is proposed for extraction of pitch using Hilbert envelope of LP residual. In Section 4.3 a method is proposed for the enhancement of speech collected over a single channel. A summary of the applications discussed in this chapter is given in Section 4.4.

4.2 Extraction of Pitch in Adverse Conditions

Even though several algorithms have been proposed in the literature for extraction of pitch, cepstrum and SIFT algorithms stood over time as good, simple and efficient methods for the estimation of pitch. However, these methods are suitable mainly for clean speech. Performance of these methods deteriorates significantly as the degradation increases. Hence, new methods for extraction of pitch are needed. A method is proposed for the extraction of pitch in adverse conditions, and it is based on the information of GC events. The proposed algorithm employs autocorrelation analysis and its results are compared with the results of the SIFT algorithm.

4.2.1 Pitch Extraction using GC Event Information

One approach for extraction of pitch is to detect the peaks near GC events in the Hilbert envelope and compute the time difference of successive peaks. But peak picking, especially in the case of degraded speech, is difficult. Therefore autocorrelation analysis of Hilbert envelope of LP residual is proposed here. Although the autocorrelation of a voiced speech segment generally displays a peak at the pitch period, the peaks due to formant structure of the signal are also often present. The autocorrelation of LP residual shows a peak at the pitch period without any significant influence of peaks corresponding to the formants. The ease with which this peak can be detected depends on the prominence of the peak, which in turn depends on the phase values of the signal around the GC events. The ambiguity due to the phase can be minimized using Hilbert envelope of LP residual.

A segment of voiced speech, its LP residual, Hilbert envelope of LP residual and the corresponding autocorrelation sequences are shown in Figure 4.1. Since the Hilbert envelope is positive, the mean of the segment is subtracted before computing the autocorrelation. The LP residual is computed from differenced speech (sampled at 8 kHz) by LP analysis using a frame size of 20 ms, a frame shift of 5 ms and an LP order of 10. The Hilbert envelope of LP residual is also processed using frames of 20 ms with a shift of 5 ms to extract pitch. In the autocorrelation sequence, the first major peak in the range of 2.5 to 12.5 ms after the central peak is detected. The distance of the first major peak from the central peak is marked as the pitch period. The pitch periods from the previous and next frames are also computed. If the pitch period of the present frame is within ± 0.25 ms (2 samples at 8 kHz) of either of the adjacent periods, then the pitch value is retained for validation in the next stage, else it is discarded.

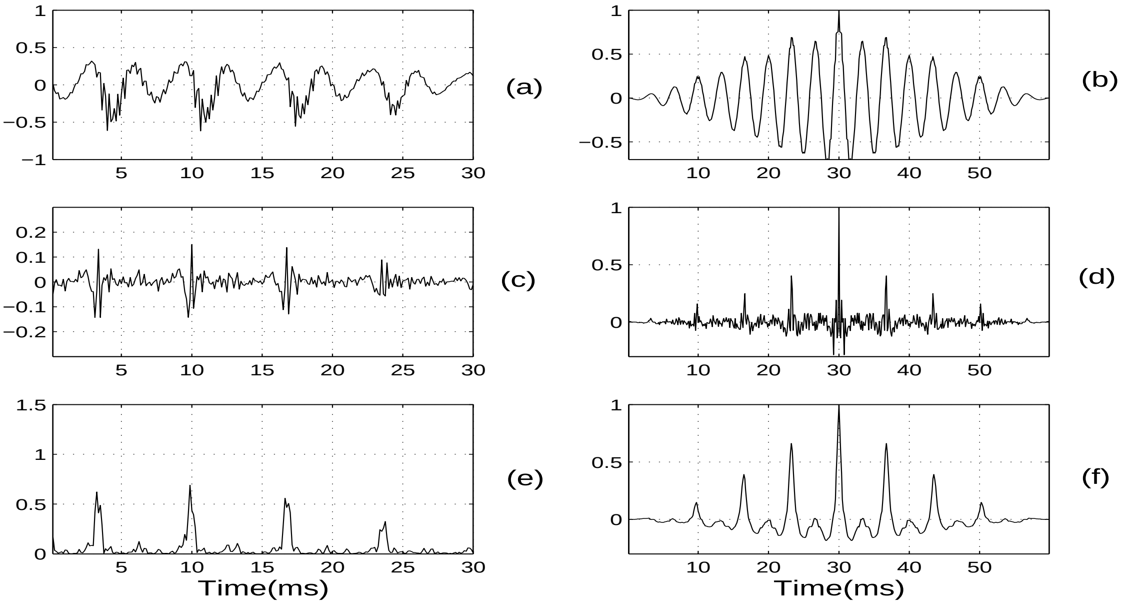

Figure 4.1: (a) Segment of voiced speech and its (b) autocorrelation sequence. (c) Segment of LP residual and its (d) autocorrelation sequence. (e) Segment of Hilbert envelope of the LP residual and its (f) autocorrelation sequence.

Another property of Hilbert envelope of LP residual is the similarity of behavior of the samples around the first major peak in the autocorrelation sequence of the adjacent frames for voiced speech. This similarity can be measured by comparing samples in a region of 2.5 ms on either side of the first major peak of the present frame, with the samples from the previous or the next frame. This is measured using the correlation coefficient (c) [99], which is given by

$$c = \frac{\sum |(x - \bar{x})||(y - \bar{y})|}{\sqrt{\sum (x - \bar{x})^2}\sqrt{\sum (y - \bar{y})^2}} \tag{4.1}$$

where x and y represent samples around the first major peak in the current frame and the previous or the next frame, respectively, and $\bar{x}$ and $\bar{y}$ represent their mean. If the correlation coefficient is more than 0.7, then the pitch value of the present frame is accepted, else it is set to zero. The proposed algorithm for extraction of pitch is given in Table 4.1. A segment of the voiced speech from an isolated utterance, the values of correlation coefficients (5 point median filtered) and pitch values extracted by the proposed method and the SIFT algorithm are shown in Figure 4.2. In the case of isolated utterance, there is not much variation in the pitch, as the speech is well articulated. Hence both the methods perform equally well.

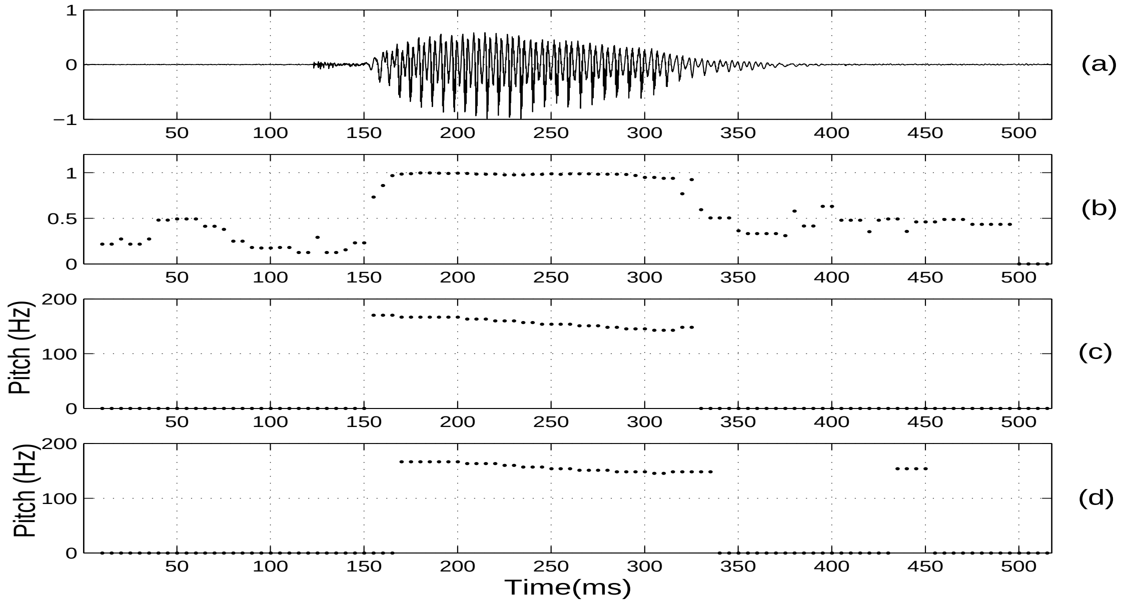

Figure 4.2: (a) Speech of isolated utterance /*ki*/. (b) Values of correlation coefficients. (c) Pitch values from the proposed algorithm. (d) Pitch values from the SIFT algorithm.

Table 4.1: Proposed algorithm for the extraction of pitch

1. Preemphasize the input speech signal.
2. Compute LP residual using frame size = 20 ms, frame shift = 5 ms and LP order = 10.
3. Compute Hilbert envelope of the LP residual.
4. Perform autocorrelation on the Hilbert envelope of the LP residual using frame size = 20 ms and frame shift = 5 ms.
5. In the autocorrelation sequence find the first major peak after the central peak, in the range 2.5 to 12.5 ms and find its distance from the central peak.
6. Find the similarity between the small segment (2.5 ms on either side of the first major peak) of the present frame with the corresponding segment from the previous or the next frame using the correlation coefficient given by
   $c = \frac{\sum |(x-\bar{x})||(y-\bar{y})|}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$
7. Smooth the correlation coefficient values with a 5 point median filter.
8. If the distance of the first major peak is approximately same as that of the previous or the next frame (±0.25 ms), retain this value for further validation, else set the pitch value as zero.
9. If the distance value is nonzero and the similarity measure is greater than or equal to 0.7, then declare the distance as the pitch, else set the pitch value to zero.

4.2.2 Pitch in Continuous Speech

In continuous speech the pitch may vary over a large range. A segment of continuous speech is taken from a broadcast news database. The values of correlation coefficients computed from the autocorrelation of Hilbert envelope of LP residual, and the pitch contours obtained by the proposed method and the SIFT algorithm are shown in Figure 4.3. The proposed algorithm is able to preserve the variations in the extracted pitch values. In case of SIFT algorithm the performance degrades slightly in regions where there is sudden change in the pitch values. The poor performance may be attributed to the less prominence of the first major peak in the autocorrelation of the LP residual. While singing, there will be large variations in the values of pitch. A segment of speech extracted from a song and the extracted pitch contours by the proposed method and the SIFT algorithm are shown in Figure 4.4. The pitch contour by the proposed algorithm preserves the variations in the pitch better, compared to the SIFT algorithm.

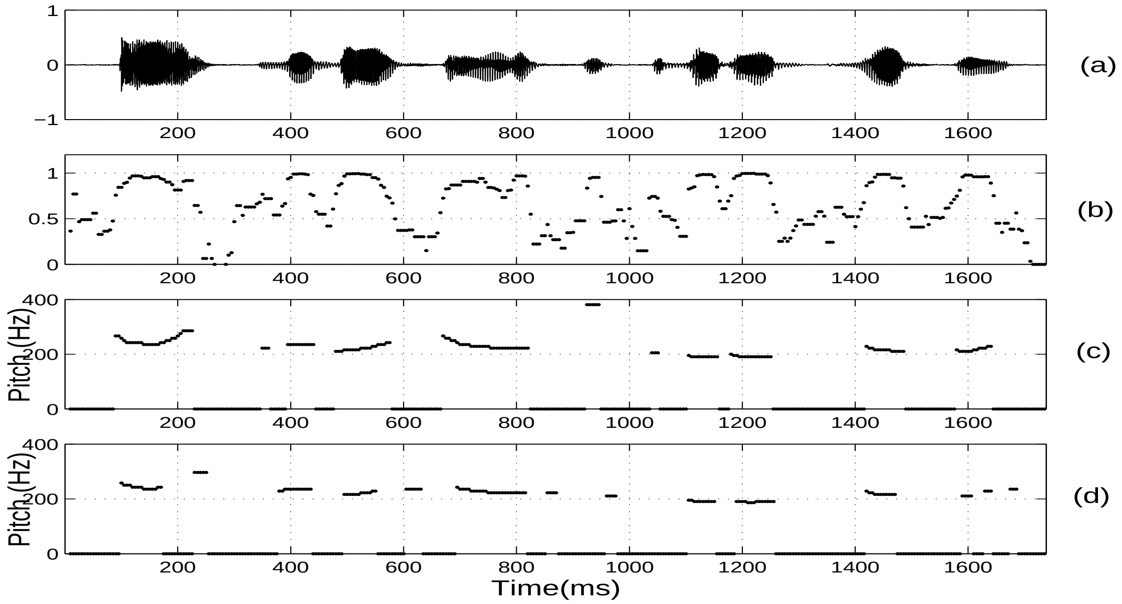

Figure 4.3: (a) Segment of continuous speech extracted from a broadcast news database. (b) Values of correlation coefficients. (c) Pitch values from the proposed algorithm. (d) Pitch values from the SIFT algorithm.

4.2.3 Pitch in Adverse Conditions

Practically there will be situations in which speech signal may be degraded by the presence of background noise, reverberation and channel noise. In such conditions humans are still able to perceive speech. Processing speech to extract pitch in such conditions is a challenging task. A segment of degraded speech, its LP residual and Hilbert envelope of LP residual and their autocorrelation sequences are shown in Figure 4.5. Since, in the Hilbert envelope of LP residual, the information of the GC events is preserved better compared to the LP residual, the autocorrelation analysis of Hilbert envelope brings out pitch information clearly. This is evident in the autocorrelation sequences shown in Figure 4.5.

A segment of continuous speech degraded by the background noise and reverberation, and the corresponding pitch contours by the proposed method and the SIFT algorithm, are shown in Figure 4.6, for a male speaker. The pitch contour extracted by the proposed method appears to be smoother compared to that extracted by the SIFT algorithm. The main effect of degradation is on the phase of LP residual. This in turn affects the prominence of the peak in the autocorrelation sequence of the LP residual. In the proposed method the effect of this phase is reduced using Hilbert envelope of

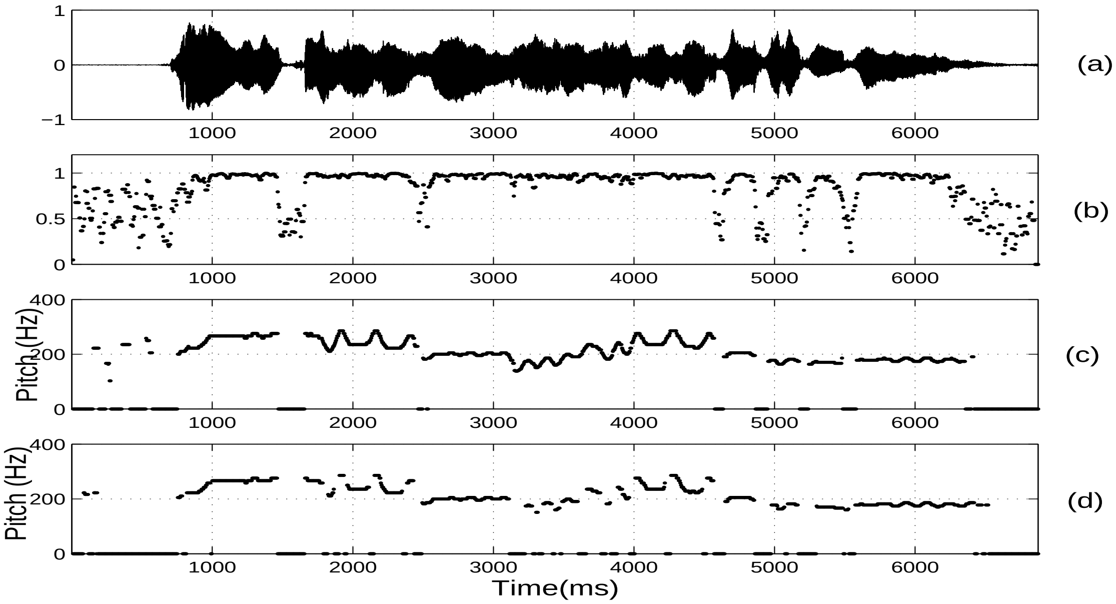

Figure 4.4: (a) Segment of continuous speech extracted from a song. (b) Values of correlation coefficients. (c) Pitch values from the proposed algorithm. (d) Pitch values from the SIFT algorithm.

LP residual.

A segment of speech signal degraded by background noise and reverberation and collected from another acoustical environment is shown in Figure 4.7, for a female speaker. The pitch contours extracted by the proposed method and the SIFT algorithm are shown in Figure 4.7. The pitch contour shape is smoother for the proposed method. This example illustrates the robustness of the proposed method.

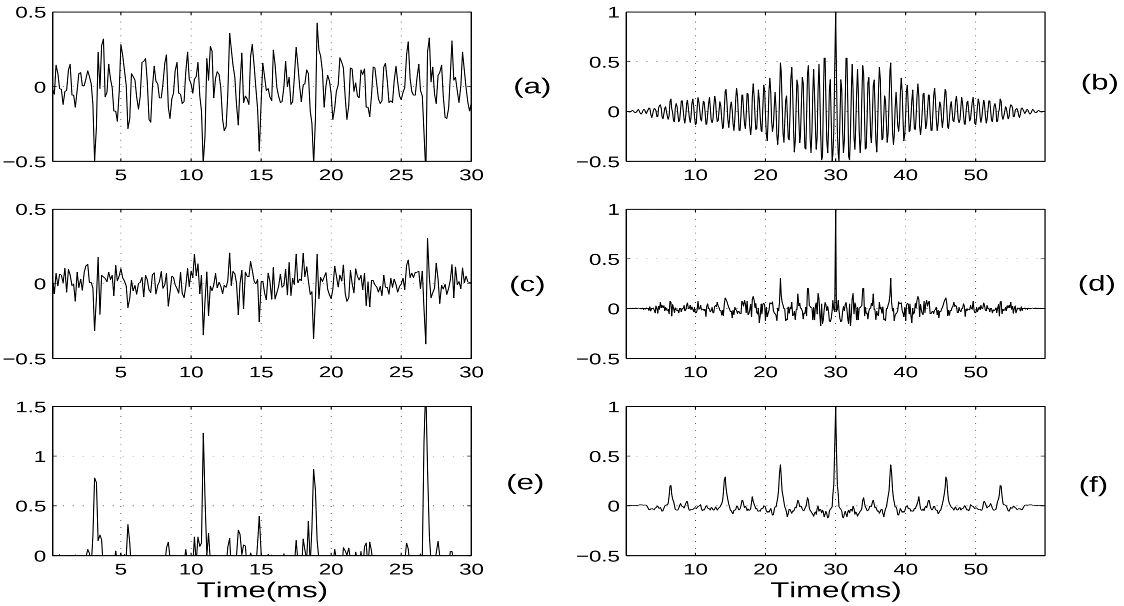

Figure 4.5: (a) Segment of degraded speech and its (b) autocorrelation sequence. (c) Segment of LP residual and its (d) autocorrelation sequence. (e) Segment of Hilbert envelope of the LP residual and its (f) autocorrelation sequence.

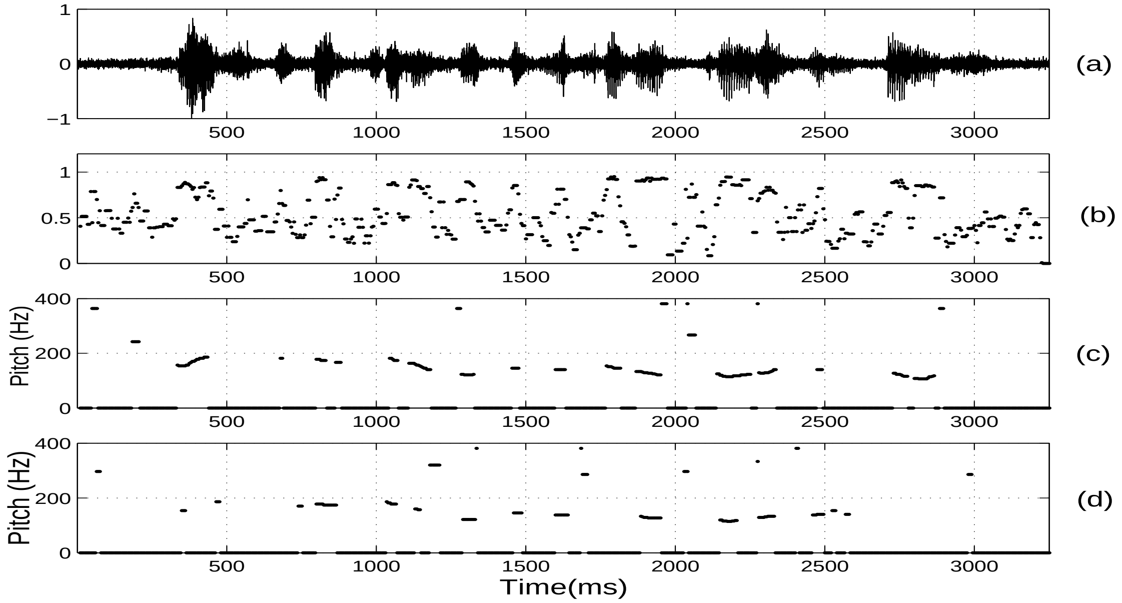

Figure 4.6: (a) Segment of degraded speech of a male speaker affected mainly by background noise and room reverberation. (b) Values of correlation coefficients. (c) Pitch values computed from the proposed algorithm. (d) Pitch values computed from the SIFT algorithm.

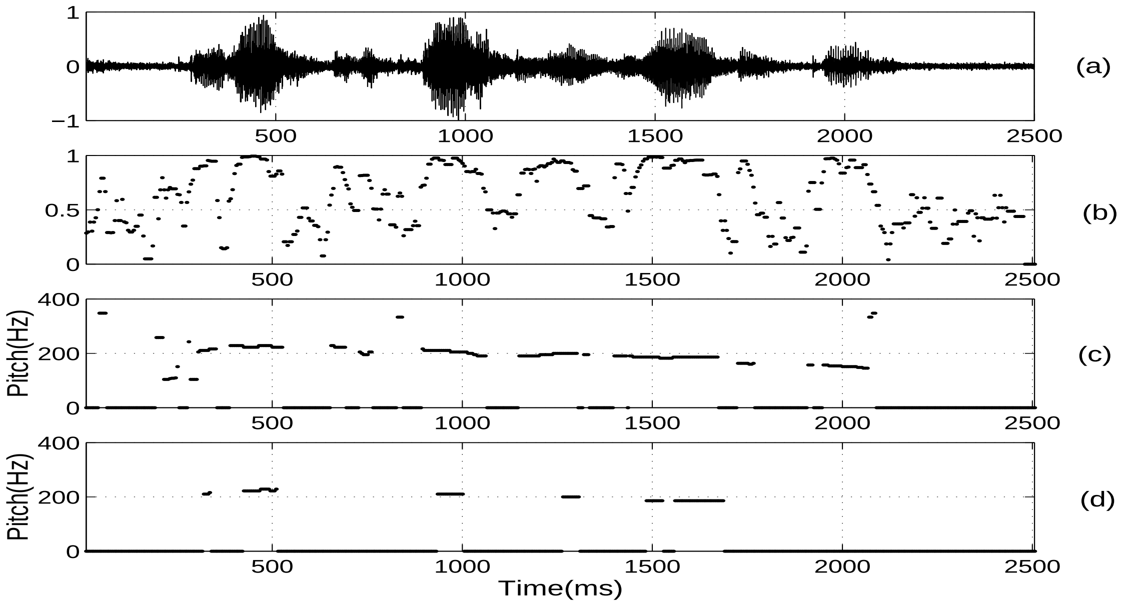

Figure 4.7: (a) Segment of speech signal of a female speaker affected by background noise and reverberation. (b) Values of correlation coefficients. (c) Pitch values computed from the proposed algorithm. (d) Pitch values computed from the SIFT algorithm.

4.3 Speech Enhancement in Single Channel Case

Perceiving information from speech signals collected over severely degraded channels is a difficult task. To increase the comfort level of listening, one can process the speech signal to reduce noise in the nonspeech regions. This is possible if we are able to identify the speech regions. Noise characteristics may be estimated and subtracted from the degraded speech signal. However, in real environments, noise characteristics vary significantly over time. Hence reliable estimation of noise characteristics is a difficult task. Alternatively, characteristics of speech may be exploited to process the degraded speech. One advantage of using the knowledge of speech is that the characteristics of speech are more predictable compared to that of the noise components [31].

4.3.1 Speech Enhancement Method

The speech-specific knowledge from the vocal tract system, the excitation source or both may be used for enhancement. In this study, we use the knowledge of the excitation source. Hilbert envelope of LP residual containing information about the excitation source is derived from the speech signal. One property of Hilbert envelope of LP residual of the speech signal collected over a severely degraded channel is that the correlation among the samples is high in speech regions and low in nonspeech regions. Autocorrelation analysis may be performed on the Hilbert envelope of the LP residual to estimate the amount of correlation among the samples. For illustration, a 30 ms frame of Hilbert envelope of LP residual computed from a high voiced segment of degraded speech, and its autocorrelation sequence are shown in Figures 4.8(a) and (b), respectively. The strength of the first peak (after the central peak) in the autocorrelation sequence is an indication of the level of correlation in the frame, which is high in this case. Hilbert envelope of LP residual of a 30 ms frame of weak voiced speech and its autocorrelation values are shown in Figures 4.8(c) and (d), respectively. The strength of peak is relatively lower in this case. Similarly, autocorrelation analysis performed for Hilbert envelope of LP residual of a 30 ms frame of nonspeech is also shown in Figures 4.8(e) and (f), respectively. Thus the autocorrelation analysis performed on the frames of Hilbert envelope of LP residual for every sample shift gives an indication of the level of speech at each sample in the degraded signal.

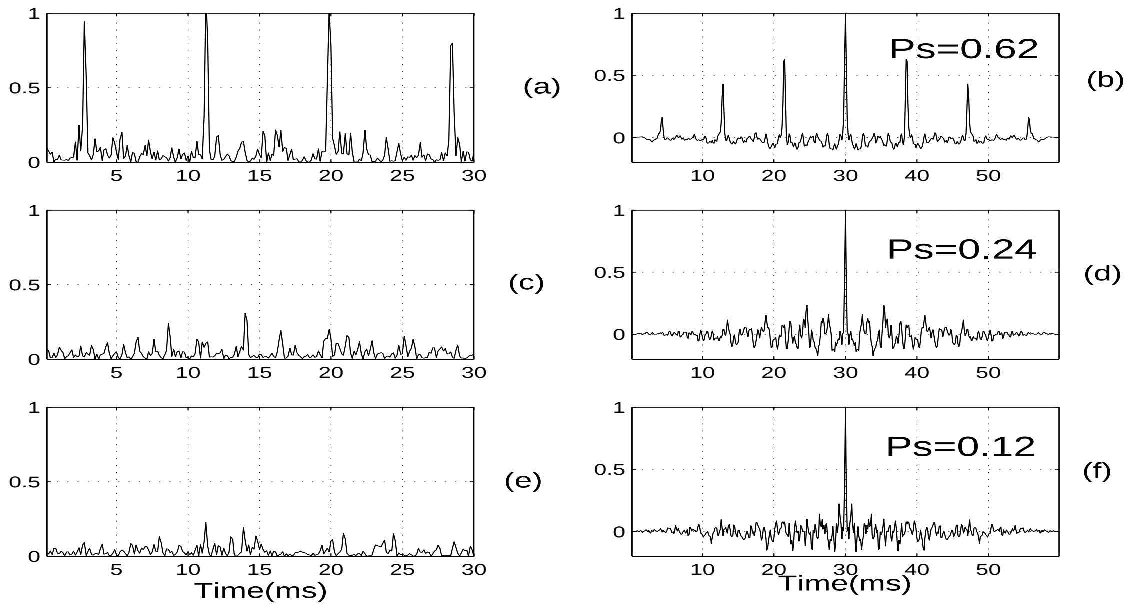

Figure 4.8: Hilbert envelope of the LP residual of a 30 ms (a) high voiced frame and its (b) autocorrelation sequence, (c) low voiced frame and its (d) autocorrelation sequence, (e) nonspeech frame and its (f) autocorrelation sequence. P_s indicates normalized first peak strength.

A 10^{th} order LP analysis is performed on the degraded speech signal (see Figure 4.9(a)) sampled at 8 kHz, to obtain the LP residual. Hilbert envelope of the LP residual is computed. The autocorrelation is performed on the Hilbert envelope of the LP residual using frames of size 30 ms and frame shift as one sample. For each frame, the strength of the first peak of the autocorrelation sequence, normalized with respect to the central peak, is noted. The normalized peak strength of the autocorrelation sequence, computed for Hilbert envelope of LP residual is shown in Figure 4.9(b). High values in the normalized peak strength indicate the speech regions. The normalized peak strength sequence is suitably processed using a 500 point Hamming window and the smoothed sequence is shown in Figure 4.9(c). A weight function is derived from the smoothed sequence using a nonlinear mapping function in such a way that the samples corresponding to the speech regions are enhanced relative to the samples in the nonspeech regions. The nonlinear mapping function is given by

$$P_m = \frac{1}{1 + e^{-(P_s - \theta)/\tau}} + \alpha \tag{4.2}$$

where, P_m is value of the weight function value, P_s is the smoothed peak strength value (normalized in the range 0-1), $\theta = 0.2$, $\tau = 0.04$ are the slope parameters and $\alpha = 0.05$ is the offset which is the minimum value of the weight function. The weight function derived using the mapping function is shown in Figure 4.9(d).

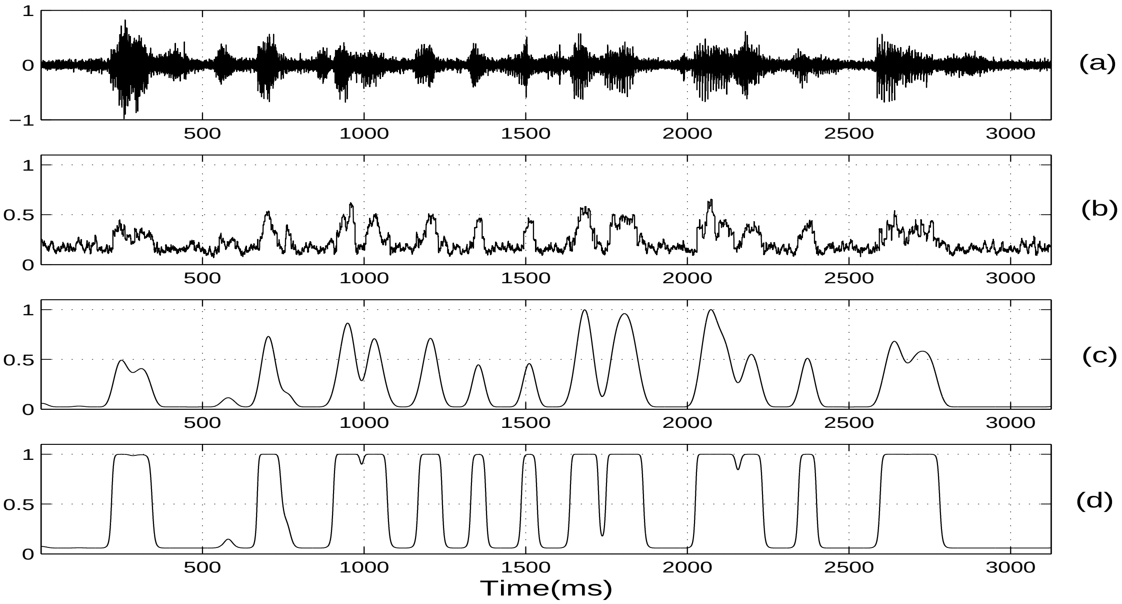

Figure 4.9: (a) Degraded speech signal, (b) normalized peak strengths, (c) smoothed peak strengths and (d) weight function to enhance the speech regions.

The LP residual of the degraded speech signal is processed using the weight function to produce the modified LP residual. As the samples of the LP residual signal are less correlated compared to the samples of the speech signal, modifying the LP residual may introduce less distortion in the synthesized signal. The enhanced speech signal is synthesized by exciting the time-varying filter using the modified LP residual. The parameters of the filter are derived from the degraded speech signal. The proposed algorithm is summarized in Table 4.2.

4.3.2 Experimental Results

A segment of speech signal collected over a severely degraded channel is shown in Figure 4.11(a). The LP residual computed using a 10^{th} order LP analysis is shown in Figure 4.11(b). The Hilbert envelope of the LP residual is processed as discussed in the previous section, to derive the weight function which is shown in Figure 4.11(d).

Table 4.2: Proposed algorithm for enhancing the speech signal collected over a single channel.

1. Preemphasize the degraded speech signal.
2. Compute LP residual (frame size = 20 ms, frame shift = 10 ms and LP order = 10).
3. Compute Hilbert envelope of the LP residual.
4. Using Hilbert envelope of the LP residual find the normalized peak strength values (frame size = 30 ms for every sample shift).
5. Smooth the normalized peak strength values.
6. Compute the weight function using $P_m = \frac{1}{1+e^{-(P_s-\theta)/\tau}} + \alpha$
7. Modify the LP residual using the weight function.
8. Synthesize speech from the modified LP residual.

The LP residual is processed using the weight function and the modified LP residual is shown in Figure 4.11(e). The excitation in the speech regions are enhanced in the modified LP residual.

The speech signal synthesized using the modified LP residual is shown in Figure 4.11(f). The speech regions are enhanced in the synthesized speech. The narrowband spectrograms of the degraded and the corresponding enhanced speech signals are shown in Figure 4.11. From the narrowband spectrograms we can infer that the energy of the frequency components in the speech regions are unaltered and are attenuated significantly in the nonspeech regions. The degraded signal and the corresponding enhanced speech signal obtained by the proposed method are available for listening at http://speech.cs.iitm.ernet.in/Main/result/enhance.html.

4.4 Summary

In this chapter a method for extraction of pitch in adverse conditions was proposed using the information about the GC events. Hilbert envelope of LP residual was used to represent the GC events. Pitch was extracted by performing autocorrelation analysis on the mean subtracted Hilbert envelope frames. A method was proposed for enhancing speech collected over a single channel. The proposed method exploits the knowledge of the excitation source of speech production to identify the speech regions. The speech regions were enhanced by emphasizing the excitation in the speech regions of the LP residual. The synthesized speech signal was found to be significantly enhanced perceptually compared to the degraded speech signal. A summary of various issues discussed in this chapter is given in Table 4.3.

In the next chapter we discuss some more applications of GC events in which the speech data is collected over multiple microphones.

Table 4.3: Summary of the discussion with respect to applications of GC events for single channel case.

Extraction of Pitch in Adverse Conditions

- Issues involved in the extraction of pitch
  - Performance of existing algorithms degrades under adverse conditions.
- Proposed method for extraction of pitch
  - Uses information about GC events.
  - Autocorrelation analysis on Hilbert envelope of LP residual.
  - Since the effect of phase is minimized in the Hilbert envelope, the proposed method is suitable for adverse conditions also.

Speech Enhancement in Single Channel Case

- Issues involved in the enhancement of speech
  - Difficult to perceive information from speech collected over a severely degraded channel.
  - The degraded speech needs to be processed for perceptual enhancement.
  - Existing methods estimate the noise characteristics and subtract the same from the degraded speech. But estimation of noise characteristics is difficult as it varies over time.
- Proposed method for enhancement of speech
  - Uses speech-specific knowledge to identify speech regions.
  - Autocorrelation analysis on Hilbert envelope of LP residual to derive a weight function.
  - Multiply the LP residual with the weight function and synthesize speech from the modified LP residual.

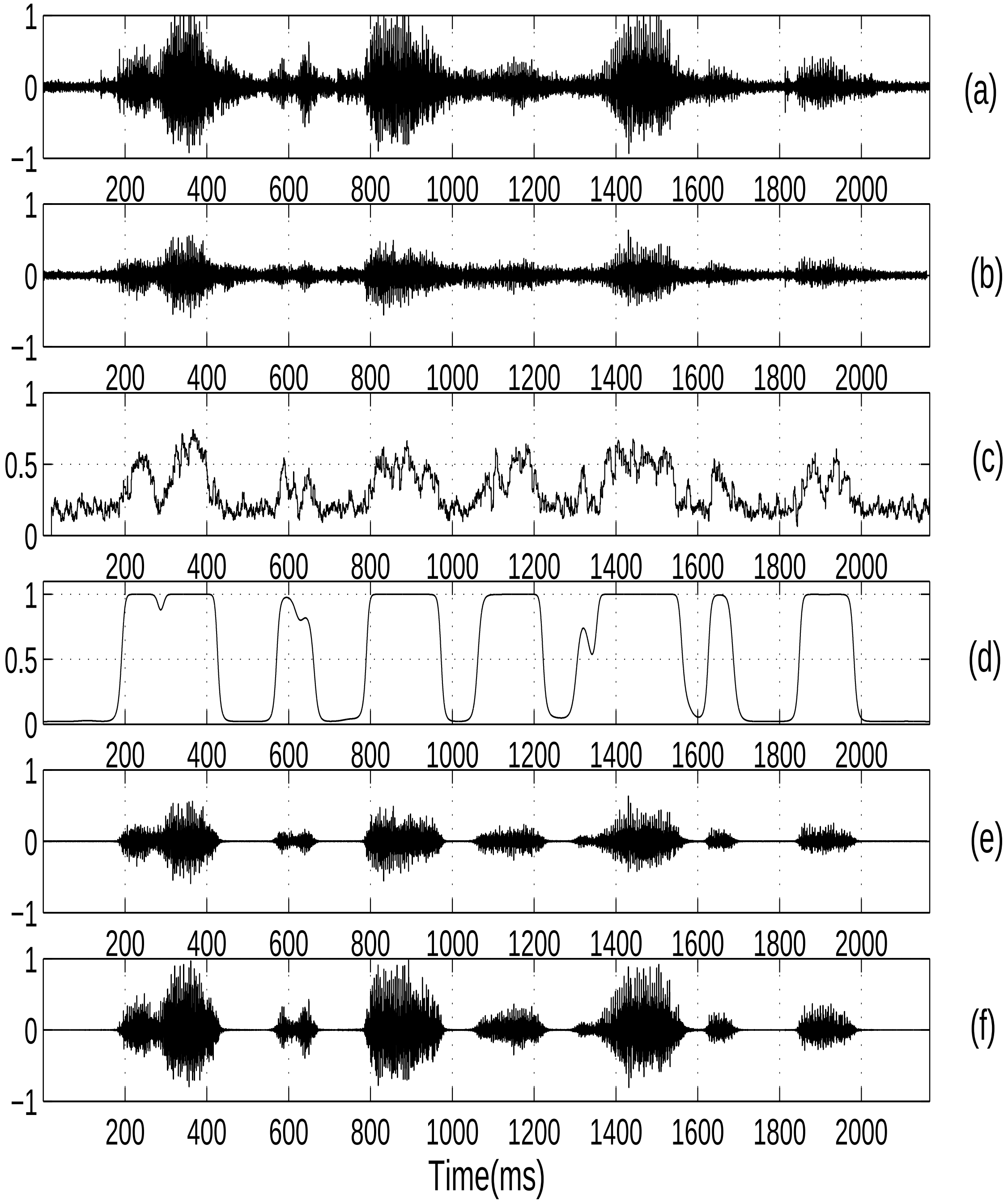

Figure 4.10: (a) Degraded speech signal, and its (b) LP residual, (c) normalized peak strength values, (d) weight function, (e) modified LP residual and (f) enhanced speech signal.

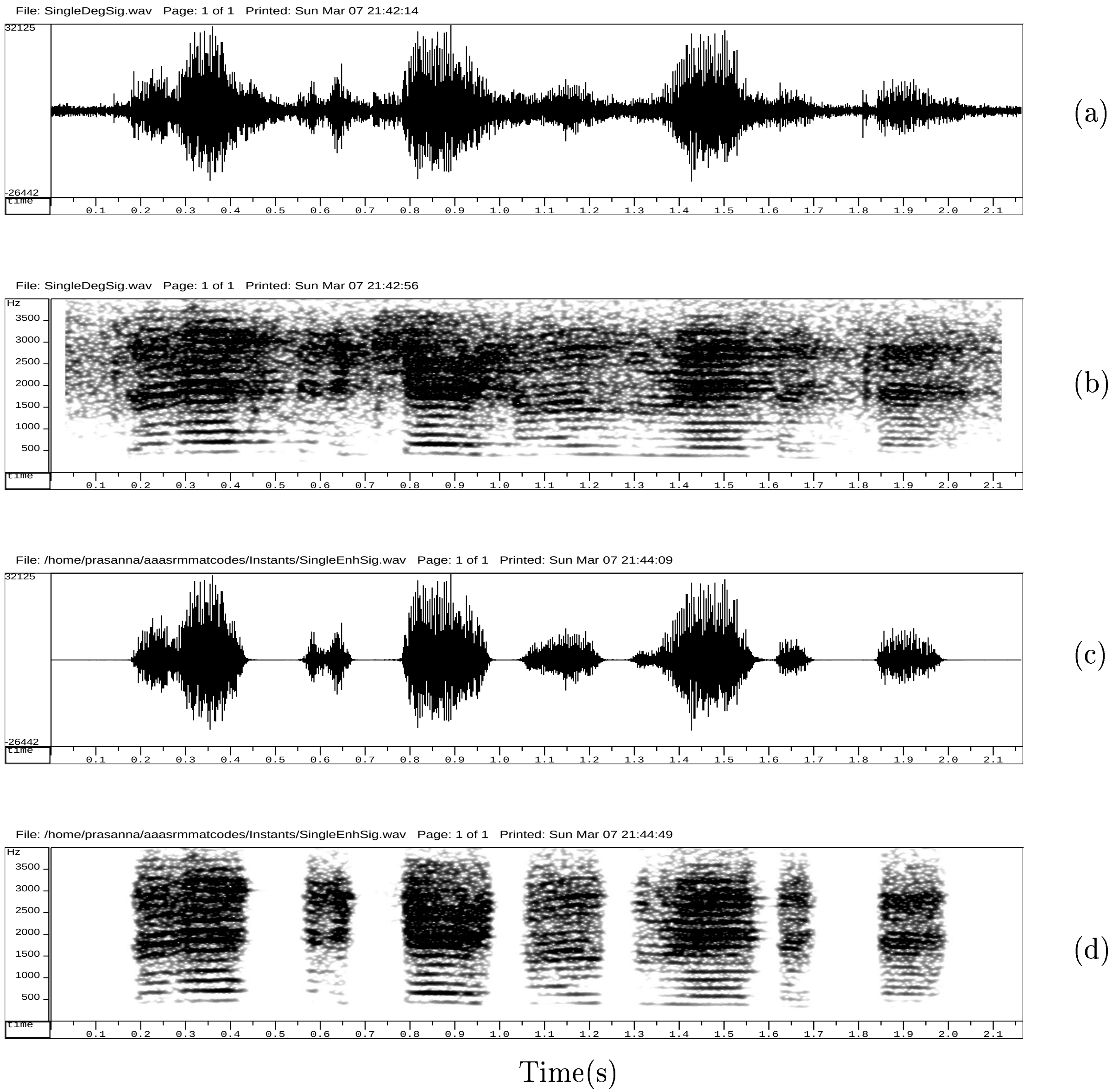

Figure 4.11: (a) Degraded speech signal and its (b) narrowband spectrogram. (c) Enhanced speech signal and its (d) narrowband spectrogram. The waveforms (a) and (c) are same as (a) and (f), respectively.

Chapter 5

APPLICATIONS OF GC EVENTS FOR MULTICHANNEL CASE

In the previous chapter we discussed two applications namely, extraction of pitch and enhancement of speech in single channel case. We explained how the knowledge of GC events is useful in these applications. In this chapter we discuss three more applications namely, time-delay estimation, speech enhancement in multichannel case and speech enhancement in a multispeaker environment. In all these applications, speech from the acoustical environment is collected simultaneously using multiple microphones. We explain how the information of GC events is useful in these applications.

5.1 Introduction

A method of estimating time-delay between the speech signals collected at two microphone locations using the knowledge of GC events is presented. For time-delay estimation, speech signals are normally processed using short-time spectral information (magnitude or phase or both). The spectral features are affected by degradations in speech caused by noise and reverberation. Features corresponding to the excitation source of speech production mechanism are robust to such degradations. By that we mean the relative spacing between GC events will not be affected by the degradations. Also the excitation source features are important mainly from the perception of point of view. The time-delay estimate can be obtained using the source features extracted even from short segments (50-100 ms) of speech from a pair of microphones. The proposed method for time-delay estimation is found to perform better than the Generalized Cross-Correlation (GCC) approach.

A method is proposed for enhancing speech signal corrupted by background noise and reverberation using the knowledge of the excitation source. The speech signal is collected using a set of spatially distributed microphones. The first step in the procedure for speech enhancement in multichannel case is the estimation of time-delay between a pair of microphones, which is obtained/computed using the information about GC events. Addition of speech signals from several microphones, after compensating for the delays, will give enhancement mainly against the background noise. The coherently-added signal can be processed further for achieving enhancement against reverberation. A weight function to highlight the high SNR regions is derived from the excitation source information. The residual of the coherently-added speech signal is multiplied with the weight function to enhance the high SNR regions. The weighted residual is used to excite the time-varying all-pole filter to obtain an enhanced speech signal. Performance of the proposed method is illustrated through spectrograms, subjective and objective evaluations.

We also propose a method for enhancing the speech of an individual speaker from the speech of multiple speakers using the knowledge of excitation source of speech production. Speech in a multispeaker environment is collected simultaneously over two spatially distributed microphones. The time-delay of the speech collected by a pair of spatially separated microphones is different for each speaker. The time-delay at a pair of microphones due to each speaker is estimated using the information around GC events in the excitation source of voiced speech. The estimated time-delays are used to reinforce the excitation information present in the residual signal, obtained after removing the significant resonances of the vocal tract system. In the reinforced signal corresponding to one speaker, the excitation information in the signals add up coherently, and the excitation information of the other speakers add up incoherently. This property is exploited to derive a weight function that enhances the characteristics of excitation of one speaker relative to other speakers. The weight function is used to modify the excitation residual signal derived from the degraded speech signal and the modified excitation signal is reused to synthesize the speech for the desired speaker. The proposed method of enhancement is demonstrated through waveforms and listening tests.

This chapter is organized as follows: A method for estimation of time-delay using Hilbert envelope of LP residual is presented in Section 5.2. In Section 5.3 we discuss about the enhancement of speech in multichannel case. Enhancement of speech degraded by speech from other speakers is a challenging task and a method for the same is proposed in Section 5.4. Section 5.5 summarizes the applications discussed in this chapter.

5.2 Time-Delay Estimation

Most of the existing methods for time-delay estimation rely on the spectral characteristics of speech signal [47–50, 52–54, 100, 101]. The spectrum of the received signal depends on how the waveform is modified due to distance, noise and reverberation. Therefore the spectra of the signals obtained at two different microphone locations differ significantly. Compensating for the spectral side effects or enhancement of the spectral components of speech have met with limited success, as there still will be a lack of coherence in the filtered or spectral compensated signals from different microphones.

We propose a method that relies on some features of the excitation source of voiced speech for estimating the time-delays [2, 27, 31, 37, 42, 43, 102, 103]. The method is based on exploiting the characteristics of excitation source especially for voiced speech. The excitation source for voiced speech consists of impulse-like excitation around GC events. The impulse-like excitation is robust to degradation in the sense that the relative spacing of the epochs due to direct sound remains unchanged at different microphone locations. On the other hand, the impulse-like excitation due to reflected sound occurs at random locations at the microphones. The impulse-like excitation characteristics are captured using Hilbert envelope of LP residual of voiced speech. The Hilbert envelopes can be added coherently to reinforce the direct sound and reduce relatively the effects of noise and reverberation. Thus the proposed method is better than the previous efforts because: (1) The vocal tract influence which changes more rapidly is removed, (2) the Hilbert envelope tends to emphasize the instants of significant excitation and (3) the instants of significant excitation from the Hilbert envelopes of different microphones can be added coherently to get a more reliable output. For coherent addition of the Hilbert envelopes, time-delay between two microphone signals needs to be estimated.

5.2.1 Significance of GC Events for Time-Delay Estimation

In the proposed method for time-delay estimation, the production characteristics of speech are exploited to extract the relevant information from the degraded speech signal received at a microphone. In speech, the response of the vocal tract system is superimposed on a sequence of glottal excitation pulses. Since the waveform is affected by the transmission medium, noise and the response of the room, the received speech signal contains information about the vocal tract system corrupted by different types of degradations at different microphones. It is difficult to determine the characteristics of these degradations to compensate for their effects by processing the received signal.

The instants of significant excitation in a voiced segment are unique and their locations along the time scale do not vary with the transfer characteristics of the medium and the microphones [102]. Noise and reverberation components show significant amplitudes in the extracted excitation component at instants other than the epochs due to direct sound. The identification of the epochs due to direct sound is difficult, due to the presence of reverberation component in the speech signal. It is important to note that the effect of reverberation is different in different regions of a voiced segment [31]. For example, in the vowel region of a typical syllable-like unit the initial high energy pitch periods are less affected by reverberation compared to the pitch periods that occur later.

Figure 5.1 shows the signals received by a close speaking microphone (*mic-0*) and 3 other microphones (say, *mic-1*, *mic-2* and *mic-3*) placed in an office room of dimension 3m×4m×3m with a reverberation time of about 200 ms. The waveforms are clearly different from each other, and from the clean speech waveform obtained with a close speaking microphone. The figure also shows the short-time (50 ms frame shown by dashed lines) spectra for each of the segments, to illustrate the differences in the short-time spectral envelopes. Figure 5.1 shows the epoch locations for all the four signals. It is obvious from the clean speech case (Figure 5.1(b)) that if the epoch locations can be derived from the received signals, the problem of time-delay estimation is not only trivial, but also the resulting estimation will be accurate. But the spurious epochs due to noise and reverberation make it difficult to use the epoch locations directly for time-delay estimation.

A better method to estimate the time-delay is to exploit the property that the strength of excitation in voiced speech is large around the GC event. Figure 5.2 shows the LP residual, its Hilbert transform and the Hilbert envelope for a segment of the speech signal at the close speaking microphone (*mic-0*) and also for a segment of the degraded speech signal at *mic-1.* The figure clearly illustrates the important property of Hilbert envelope of a voiced speech segment, namely, the peak of the envelope occurs around the GC event within each pitch period. Even for the degraded speech signal at *mic-1*, the Hilbert envelope shows the largest peak around the GC event within each pitch period. This important property of Hilbert envelope forms the basis for the proposed method for estimating time-delay. While the amplitude of the Hilbert envelope is high at the GC event, the amplitudes of the Hilbert envelope will also be high at the epochs of the reflected sound in the reverberant speech. But these epochs will be located at random instants. In the next section, we will show how the Hilbert envelopes of the LP residual signals from different microphones can be used to estimate the time-delay for each pair of microphones.

5.2.2 Time-Delay Estimation using GC Event Information

The instants corresponding to the direct signal will be *coherent* at different microphone positions. On the other hand, the instants corresponding to the reverberation components will be at random locations along the time scale. This can be seen from Figure 5.3, where the Hilbert envelopes for signals from the three microphone positions are time aligned and displayed. The effect of coherence of the direct components can be seen when we add the delay-compensated Hilbert envelope signals from the three microphones. It is important to note that the coherent addition in Figure 5.3(e) pro-

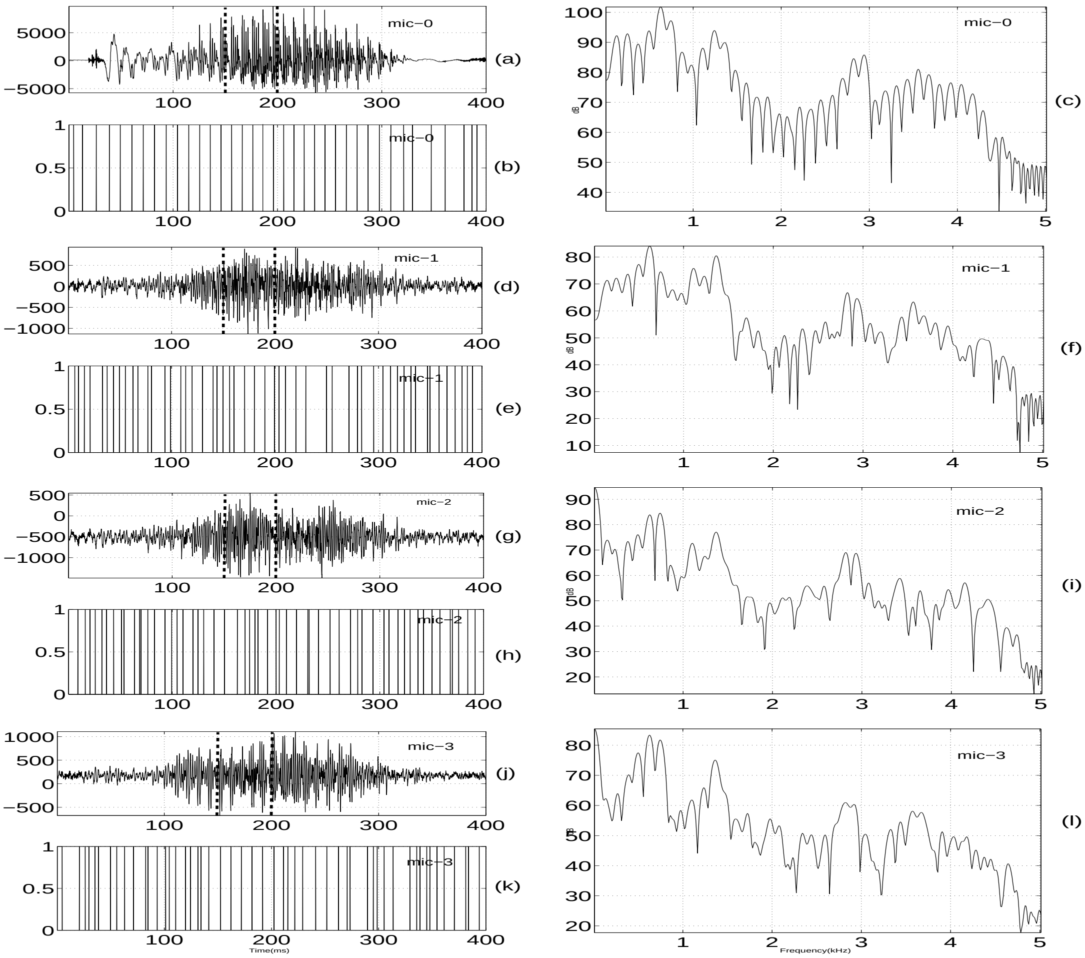

Figure 5.1: Nature of speech signals at four different microphone locations (*mic-0*, *mic-1*, *mic-2* and *mic-3*). Figures (a), (d), (g) and (j) are waveforms of the speech segments at the four microphone locations. Figures (b), (e), (h) and (k) are the extracted instants of significant excitation corresponding to the four speech segments. Figures (c), (f), (i) and (l) are the short-time spectra for the portions marked in the speech segments.

duces significant peaks at the epochs, whereas the incoherent addition in Figure 5.3(d) produces several peaks at random locations.

For coherent addition, one needs the values of the time-delays. We propose a cross-correlation method to determine the time-delays. Consider a frame of 50 ms from one of the microphones, say *mic-1* and compute the cross-correlation of Hilbert envelope of the LP residual of this frame and the corresponding frame of 50 ms from the second microphone, say *mic-2*. The cross-correlation of two sequences $x(n)$ and $y(n)$ is given by

$$r_{yx}(l) = \sum_{-\infty}^{\infty} y(n)x(n-l) \tag{5.1}$$

The location of the peak in the cross-correlation corresponds to the delay. The time-

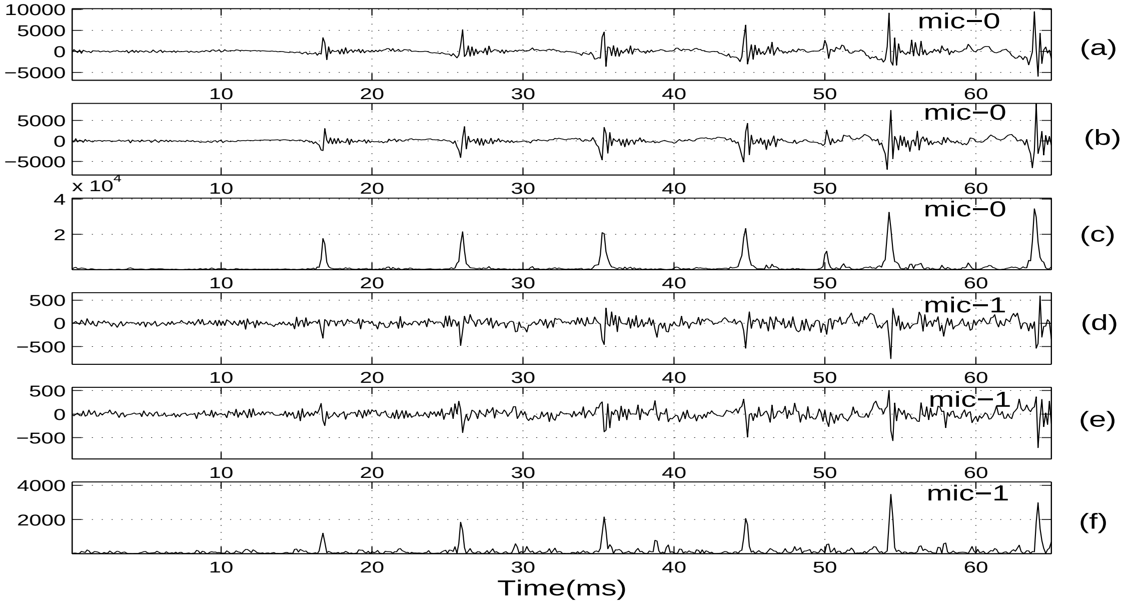

Figure 5.2: Illustration of the characteristics of the Hilbert envelope. Figures (a), (b) and (c) are the LP residual, its Hilbert transform and the Hilbert envelope for the speech signal at *mic-0*. Figures (d), (e) and (f) are the LP residual, its Hilbert transform and the Hilbert envelope for the speech signal at *mic-1*.

delay to be estimated is assumed to be much less ($< 10\%$) than the size of the frame (50 ms in this case) being considered. Figure 5.4 shows the cross-correlation function of the Hilbert envelopes of segments of the two microphone signals. The delay is indicated in number of samples from the center sample number, which is 400 in this case.

The time-delay for each frame of 50 ms is computed with a shift of 10 ms between successive frames, and the result is plotted in Figure 5.5(b). Note that the estimation results in random delays, mostly for unvoiced segments. This is indicated by segments corresponding to the low energy regions of the Hilbert envelope, as shown in Figure 5.5(a). The energy of the Hilbert envelope is obtained for each frame of 50 ms by computing the mean squared values of the amplitudes of the envelope within the frame. The normalized energy plot in Figure 5.5(a) is obtained by computing the energy for each frame shifted by 10 ms and normalizing the energy values by dividing them with the maximum value over the segment. The regions of the low normalized values, say values below 0.25 in the Figure 5.5(a), correspond mostly to silence or noise or unvoiced or low voicing regions. The low voicing regions are those regions where

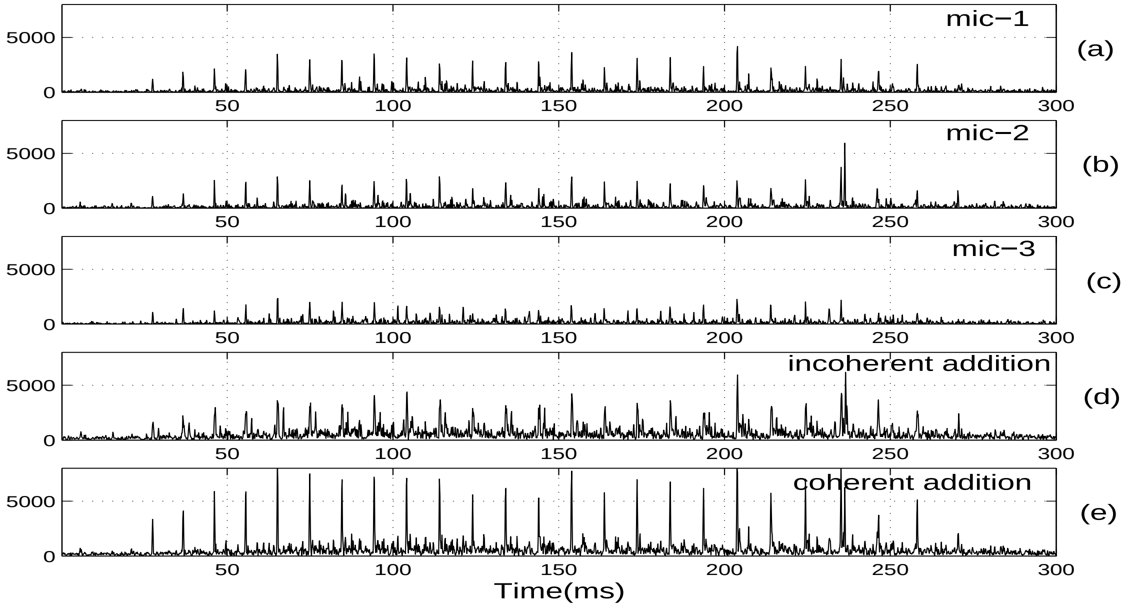

Figure 5.3: Effect of coherent and incoherent additions of the Hilbert envelopes. (a) Hilbert envelope for the signal at *mic-1*. (b) Hilbert envelope for the signal at *mic-2*. (c) Hilbert envelope for the signal at *mic-3*. (d) Result of incoherent addition of the Hilbert envelopes. (e) Result of coherent addition of the Hilbert envelopes.

the strength of excitation around the glottal closure is not high. This is also indicated by the lower values of the Hilbert envelope relative to the values in the high voicing regions.

The estimation of the time-delay gets better when we consider longer frame sizes as shown in Figures 5.5(c) and (d), for frame sizes of 200 ms and 500 ms, respectively. The improvement in the delay estimate is indicated by fewer spurious or random delays compared to the case of 50 ms frame. But using longer segments for delay estimation may make it difficult to keep track of a moving source/speaker.

Figures 5.6 and 5.7 illustrate the performance of time-delay estimation using the Hilbert envelope of the LP residual for two different types of degradation. The time-delay estimation for different levels of reverberation are obtained by placing the microphones at different distances from the speaker. Note that for longer distances, the SNR decreases due to the constant background noise. In Figure 5.7 the estimated time-delays are plotted when the microphone is placed close to a room air conditioner and a fan. A constant value of the time-delay is obtained for successive frames in the regions of high energy values in all the plots in Figures 5.6 and 5.7.

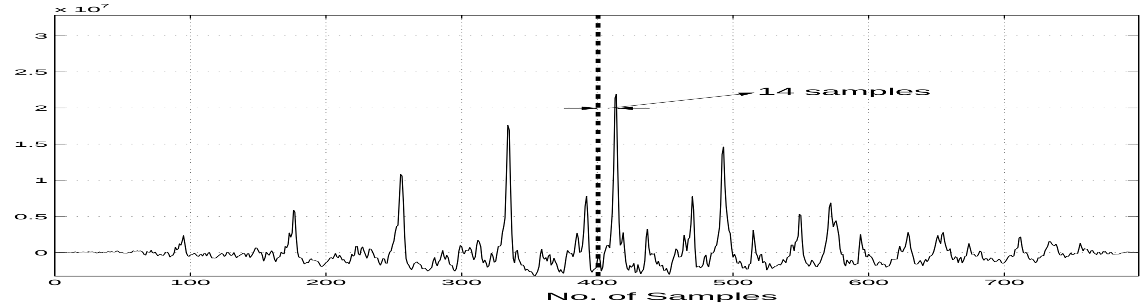

Figure 5.4: Cross-correlation function of the Hilbert envelopes of frames of size 50 ms (400 samples), corresponding to the signals at *mic-1* and *mic-2*. The time-delay estimated between *mic-1* and *mic-2* signals is 14 samples.

5.2.3 Comparison with the GCC method

The GCC ($R_{x_1x_2}(\tau)$) is computed as the inverse Fourier transform of the cross-spectrum $X_1(\omega)X_2^*(\omega)$ of the received signals, scaled by a weighting function $W(\omega)$ [53]. That is,

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} W(\omega)X_1(\omega)X_2^*(\omega)e^{j\omega\tau}d\omega \tag{5.2}$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the microphone signals $x_1(t)$ and $x_2(t)$. The weight function is chosen as $W(\omega) = |X_1(\omega)X_2^*(\omega)|^{-1}$. This corresponds to the use of phase transform for cross-correlation.

Since accurate estimation of the time-delays with smaller frame size helps in tracking a moving source, we consider a frame of 50 ms size with a shift of 10 ms to compare the performance of time-delay estimation by the proposed method and by the GCC method. Note that in the Hilbert envelope plot in Figure 5.2, there are large values around the instants of glottal closure for the voiced segments, followed by small values within each pitch period. Therefore, variance of sample values of the Hilbert envelope in the voiced region will be large, compared to that in unvoiced segments where the sample values are relatively more uniformly distributed, thus contributing to low variance. The variance of the sample values of the Hilbert envelopes is computed for each frame of 50 ms shifted by 10 ms. Plots of the standard deviation of samples of the Hilbert envelope against estimated time-delays for frames of 50 ms with a shift of 10 ms are shown in Figure 5.8. The figure shows the plots for different pairs of microphones

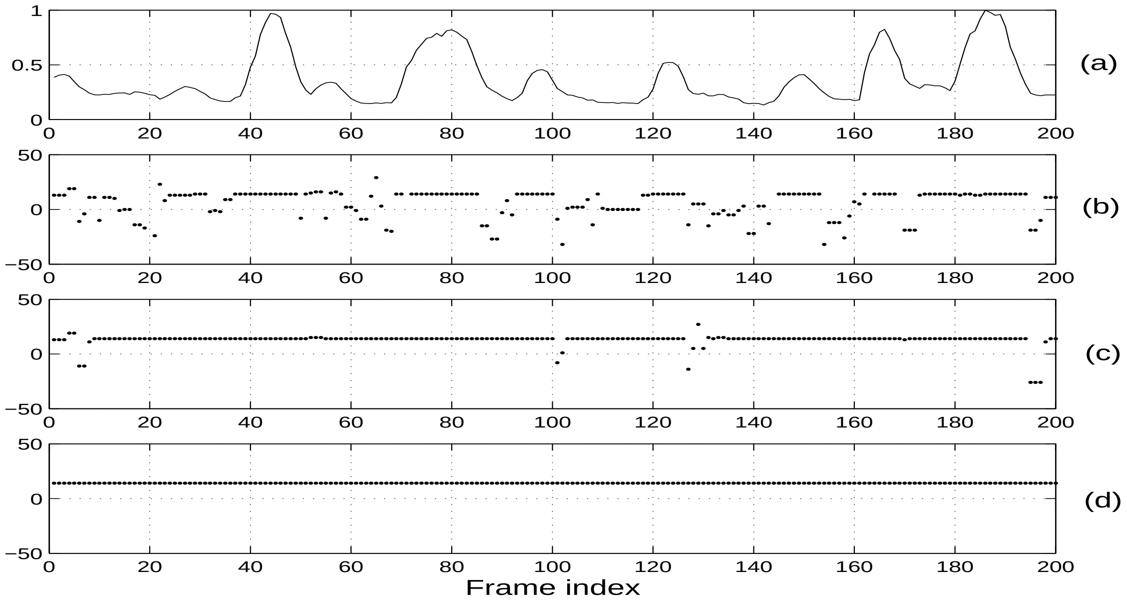

Figure 5.5: Characteristics of the estimated time-delay for different frame sizes. (a) Normalized Hilbert envelope energy. Time-delays from analysis frames of size (b) 50 ms, (c) 200 ms, and (d) 500 ms, each with a shift of 10 ms.

to illustrate the effects of degradations as the signals at different microphones are not of the same quality. One important point to be noted from these plots is that, even in the low voiced regions (low standard deviation), the proposed method estimates the delays accurately, whereas the GCC shows significant variations in the estimated delays. Ideally all the points should lie along a vertical line at the delay value. So the spread of points from the vertical line indicates degradation in the performance of the method.

An objective measure for comparison of the performance could be the ratio (r) of the number of points around the time-delay within ±1 sample deviation to the total number of points above a certain threshold of the value of the standard deviation of the samples of the Hilbert envelope. Since lower values of the standard deviation correspond mostly to nonvoiced regions, we can ignore the values below 0.25 for computing this ratio. The values of r for the different cases are shown in Figure 5.8. The larger the value of r, the better is the method for estimating the time-delay. From these illustrations we can infer that the proposed method is superior to the GCC method.

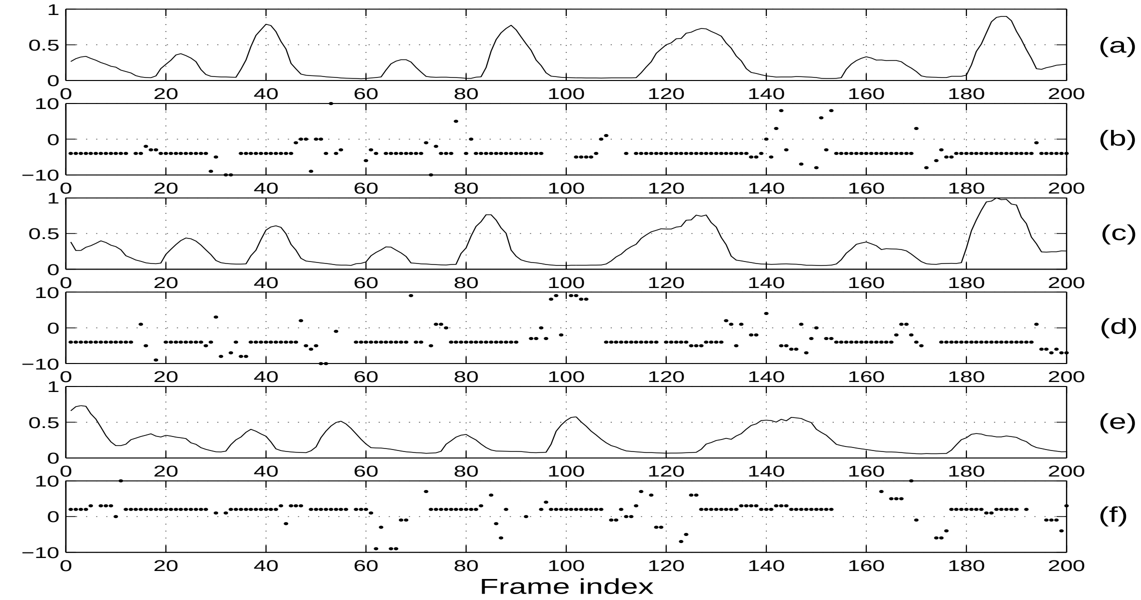

Figure 5.6: Characteristics of the estimated time-delay for different levels of reverberation obtained by collecting speech data at distances (b) 2 feet, (d) 4 feet and (f) 6 feet from the speaker. The corresponding normalized Hilbert envelope energy plots are shown for each case in (a), (c) and (e), respectively.

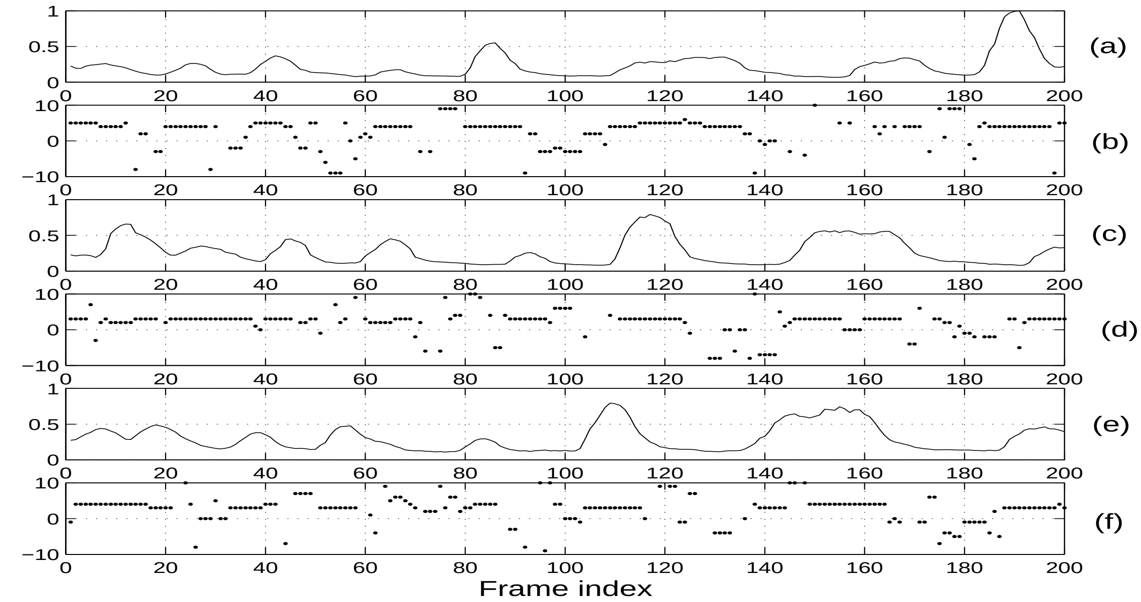

Figure 5.7: Characteristics of the estimated time-delay for speech degraded by different types of noises, namely, (b) air-conditioning noise, (d) fan noise and (f) both air-conditioning and fan noise. The microphones were placed close to the noise source, and at a distance of 6 feet from the speaker. The microphones were placed close to the noise source. The normalized Hilbert envelope energy plots are shown for each case in (a), (c) and (e), respectively.

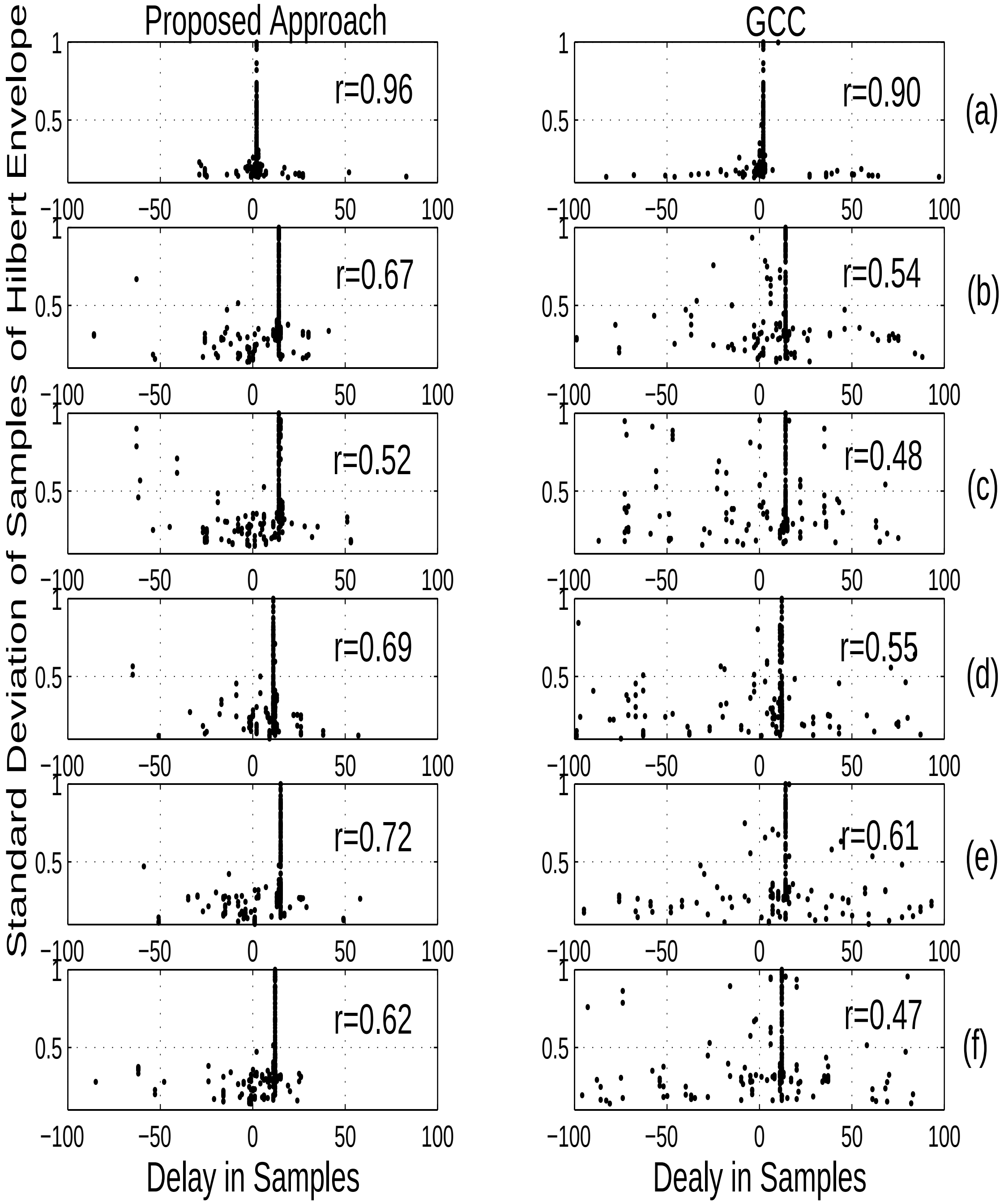

Figure 5.8: Quantitative comparison of the proposed time-delay estimation method with GCC approach. Standard deviation of samples of the Hilbert envelope vs estimated time-delay (in samples) are shown for different cases. (a) *mic-1* and *mic-3* signals. (b) *mic-1* and *mic-2* signals. (c) *mic-2* and *mic-11* signals. (d) *mic-1* and *mic-14* signals. (e) *mic-1* and *mic-15* signals. (f) *mic-2* and *mic-9* signals.

5.3 Speech Enhancement in Multichannel case

When speech is transmitted in an acoustical environment like in an office room, it will be degraded by background noise and reverberation [31,55–66,104–107]. Multichannel case is more effective for enhancement compared to the single channel case, but requires estimation of time-delays [58]. One simple method for enhancement in multichannel case is addition of the speech signals, after compensating for their delays. Coherent addition of speech signals from different microphones will provide enhancement mainly against background noise. The improvement in enhancement is directly related to the number of microphones used. For achieving significant enhancement, especially due to reverberation, additional processing of the microphone signals is required. In this work a method is proposed for enhancement using the GC event information, which helps in reducing the effects of reverberation significantly.

5.3.1 Significance of GC Events for Speech Enhancement

A segment of clean speech is shown in Figure 5.9(a), which is collected over a microphone placed close to the speaker. In this case the amount of degradation is negligible, and hence it will have high values for SNR. A segment of speech collected over a microphone placed at a distance of about 9 feet in an office room of about 3m × 4m × 3m with a reverberation time of about 200 ms, is shown in Figure 5.9(b). Clearly this speech signal is different compared to the clean speech signal in Figure 5.9(a). The speech signal collected over a distance from the speaker may be characterized by the following model.

$$x_d(n) = s(n) + z(n) + \sum_{i=1}^{N} b_i s(n - n_i) \tag{5.3}$$

where $x_d(n)$ is the degraded signal, $s(n)$ is the signal component, $z(n)$ is the background noise component, b_i is the relative amplitude of the reflection arriving after a delay of n_i samples and N is the number of reflections.

The background noise component is independent of speech, whereas the reverberation component is dependent on the previous speech samples. The value of SNR at a given instant in the speech signal depends on the values of the degrading component. High SNR regions are places where the effect of degrading component is low. In other

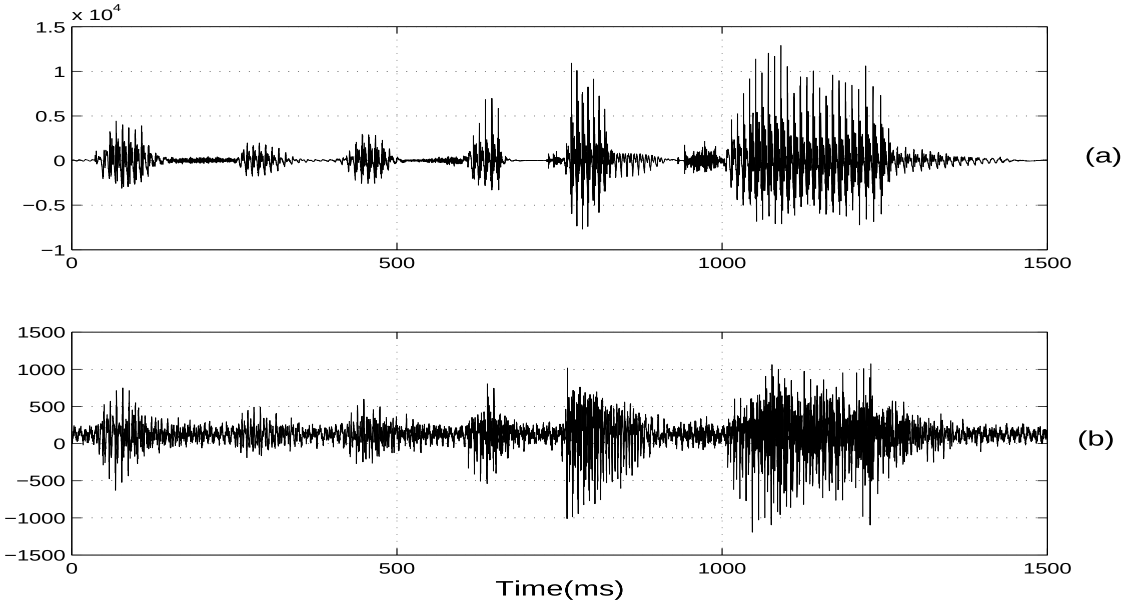

Figure 5.9: (a) Speech signal from a close speaking microphone and (b) degraded speech from a microphone placed at a distance of 9 feet from the speaker.

places the degrading component may be comparable to the the signal component, and hence these are low SNR regions. Generally the regions immediately after the onset of syllables correspond to high SNR regions. Similarly within each pitch period, regions immediately after the GC event are more likely to be high SNR regions [31].

A segment of clean speech and its LP residual are shown in Figures 5.10(a) and 5.10(b), respectively. Figures 5.10(c) and 5.10(d) show the corresponding speech and the LP residual for the degraded case. When speech is degraded, the LP residual contains random noise and reflected epochs, along with the regular epochs. The effects of the reflected epochs are not predictable. If they occur in phase with the speech epochs, then they will reinforce the strength of the excitation, otherwise they will reduce the strength of the signal. When the reflected epochs arrive in between two speech epochs, they cause reverberation. One way to minimize this effect is to enhance the signal at the speech epochs relative to the signal at the reflected epochs. This can be achieved by deriving a suitable weight function, which, when multiplied with the LP residual of the degraded speech, enhances the signal at the speech epochs. To derive a weight function for enhancing the high SNR regions, Hilbert envelope of the LP residual is used as the excitation source information. Figure 5.11 shows the LP residuals and the Hilbert envelopes for clean and degraded speech signals.

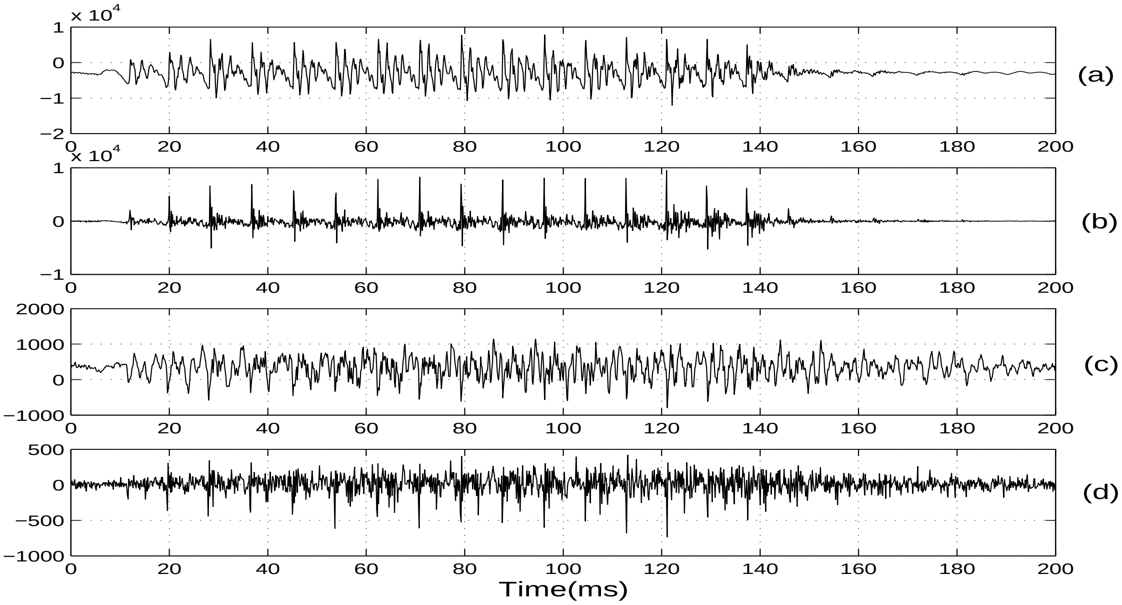

Figure 5.10: (a)-(b) Segments of clean speech and its LP residual, (c)-(d) corresponding segments of degraded speech (*mic-1*) and its LP residual.

5.3.2 Speech Enhancement using GC Event Information

Speech was collected from 14 spatially distributed microphones placed in an office room of dimension 3m × 4m × 3m with a reverberation time of about 200 ms. The delay between every pair of microphones is computed using the excitation source information as explained in the previous section. The coherently-added signal, obtained after compensating for their delays is given by

$$s_{e1}(n) = \frac{1}{N}[s_1(n) + s_2(n-\tau_{12}) + ... + s_N(n-\tau_{1N})] \tag{5.4}$$

The coherent addition reinforces speech components and thus reduces the effect of the background noise. However, the reverberant component is still present in the resulting signal. The degree of enhancement achieved at this level depends on the number of microphones used in the coherent addition. For instance, the enhanced speech signals and their narrowband spectrograms, when signals from 2, 5, 10 and 14 microphones are added, are shown in Figure 5.12. The clean speech and the degraded speech from *mic-1* are also shown in the figure. As can be seen from the narrowband spectrograms,

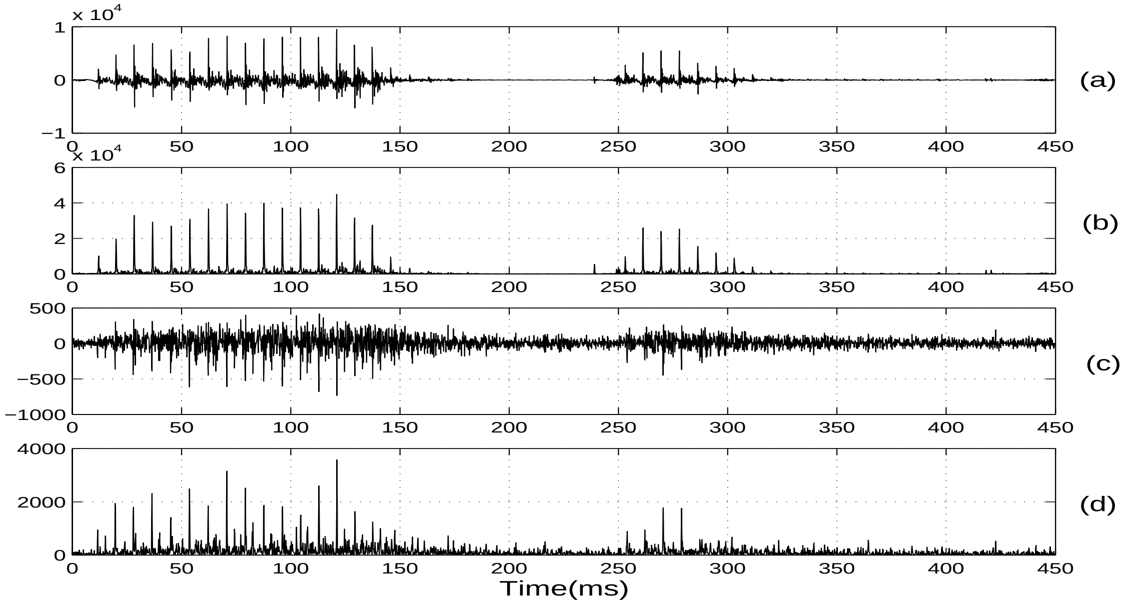

Figure 5.11: (a)-(b) LP residual and its Hilbert envelope for clean speech, (c)-(d) LP residual and its Hilbert envelope for degraded speech (*mic-1*)

there is a decrease in the background noise as we increase the number of microphones. It is interesting to note that in the case where signals from 14 microphones are added, even though the effect of background noise is reduced, the reverberation tails are still present in the speech regions. This can be observed by comparing the spectrogram, especially at low frequencies, with that of the clean speech shown in Figure 5.12(a). The presence of reverberation tails is also clearly visible in the waveforms (compare Figure 5.12(a) and Figure 5.12(f)). It is necessary to process the coherently-added speech signal further to achieve enhancement with respect to reverberation.

For each of the microphone signals, LP residual and Hilbert envelope of the LP residual are computed. The Hilbert envelopes for *mic-1*, *mic-2* and *mic-3* are shown in Figures 5.13(a), 5.13(b) and 5.13(c), respectively. The coherent addition of the Hilbert envelopes (Figure 5.13(d)) reinforces the epoch information, whereas the incoherent addition will spread the epoch information (Figure 5.13(e)).

The coherently-added Hilbert envelope exhibits several interesting features. The deviation among the samples of the Hilbert envelopes is high in the voiced speech regions. Typically, voiced speech regions in continuous speech have a minimum duration of 50 ms. Hence, by considering a block of 50 ms duration and a shift of one

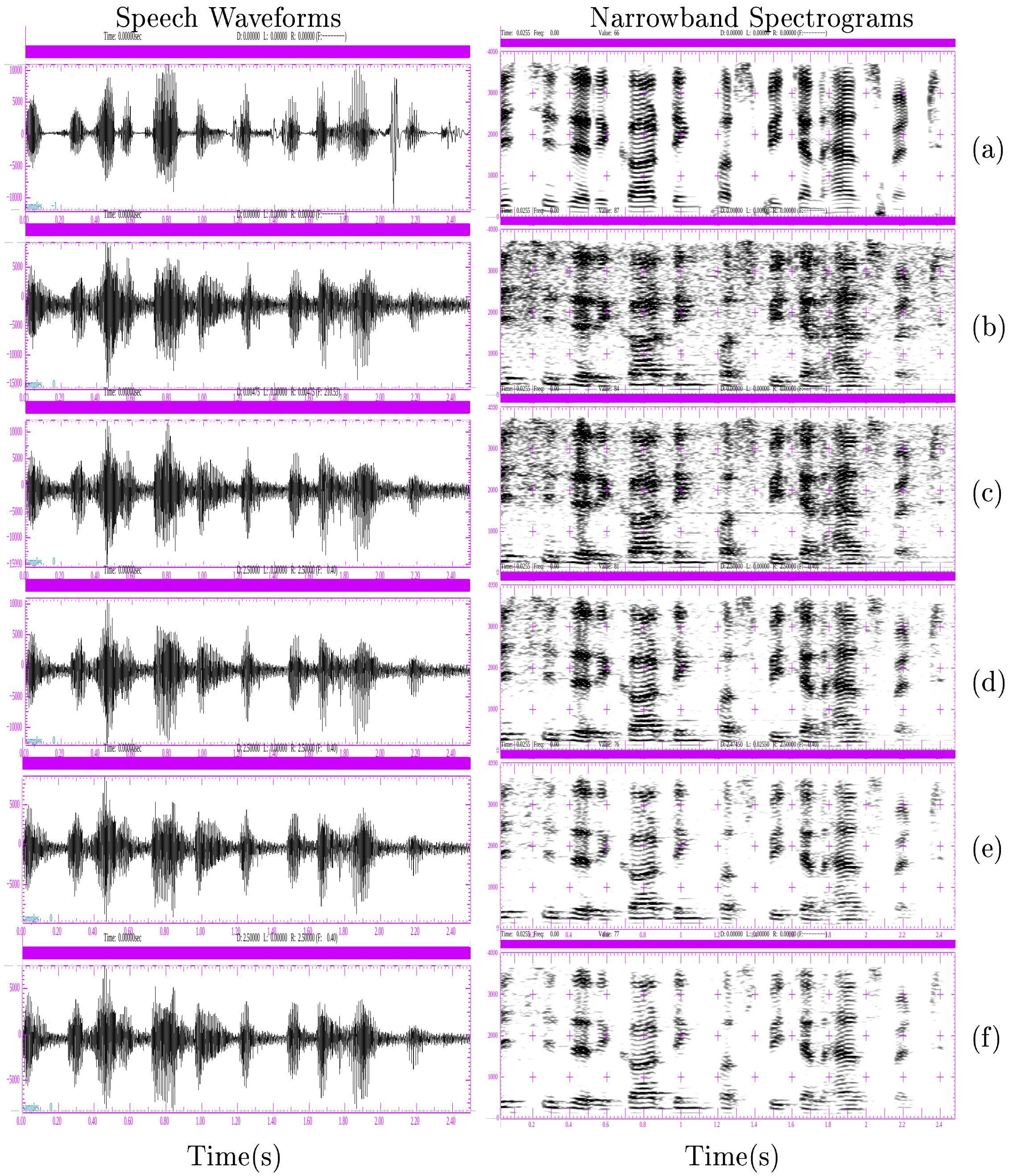

Figure 5.12: Speech and narrowband spectrograms for (a) Close speaking microphone, (b) Distant microphone (say, *mic-1*), (c)-(f) Coherently-added signals from 2, 5, 10 and 14 microphones.

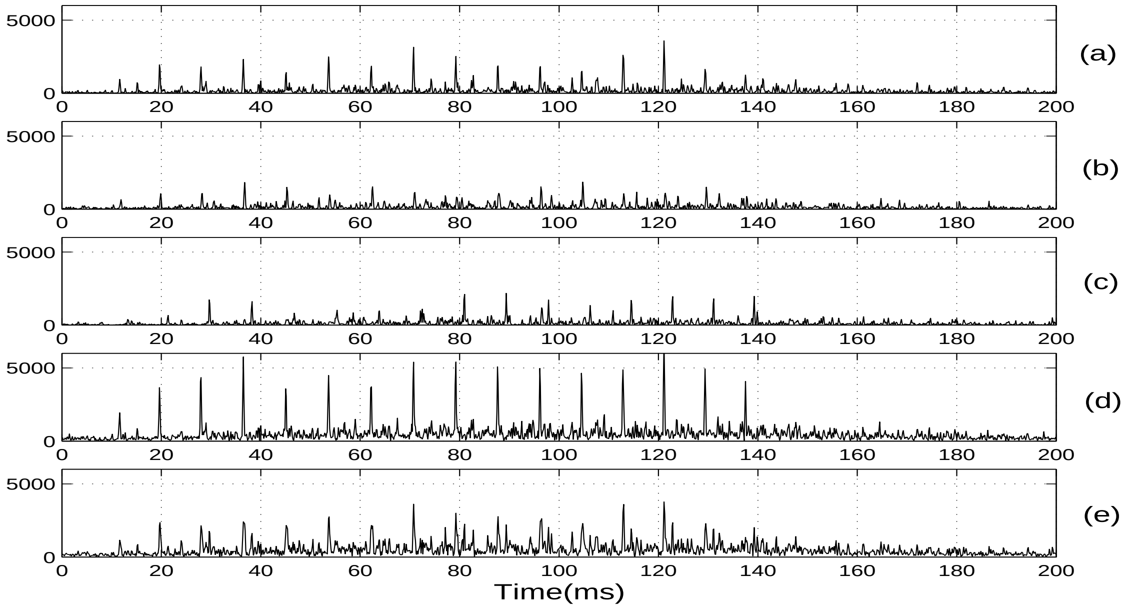

Figure 5.13: Hilbert envelope of the LP residual of a segment of (a) *mic-1* signal, (b) *mic-2* signal, (c) *mic-3* signal. Results of (d) Coherent and (e) incoherent addition, of the Hilbert envelopes of (a), (b) and (c).

sample, the mean and standard deviation of the coherently-added Hilbert envelope samples in each block are computed. The standard deviation values are normalized with the respective mean values as shown in Figure 5.14. In the normalized standard deviation plot, the deviation of Hilbert envelope samples is high in the speech regions. Further, the normalized standard deviation is high in the initial portions of the voiced speech regions and it decreases towards the end of the voiced regions. This is because the initial parts are high SNR regions. Towards the end of the voiced regions, the levels of degrading components increase and hence they correspond to low SNR regions. Another interesting property of the coherently-added Hilbert envelope is that, the samples in each pitch period around the epochs have large deviation compared to the samples away from the epoch. The mean, standard deviation and normalized standard deviation for a segment of coherently-added Hilbert envelope are shown in Figure 5.15, for a block size of 3 ms and a shift of one sample.

A weight function is derived by adding the two (long and short blocks) normalized standard deviation values as shown in Figure 5.16. The combined deviation plot is multiplied with the LP residual of the coherently-added signal. The residual of the coherently-added signal along with the modified residual using the combined deviation

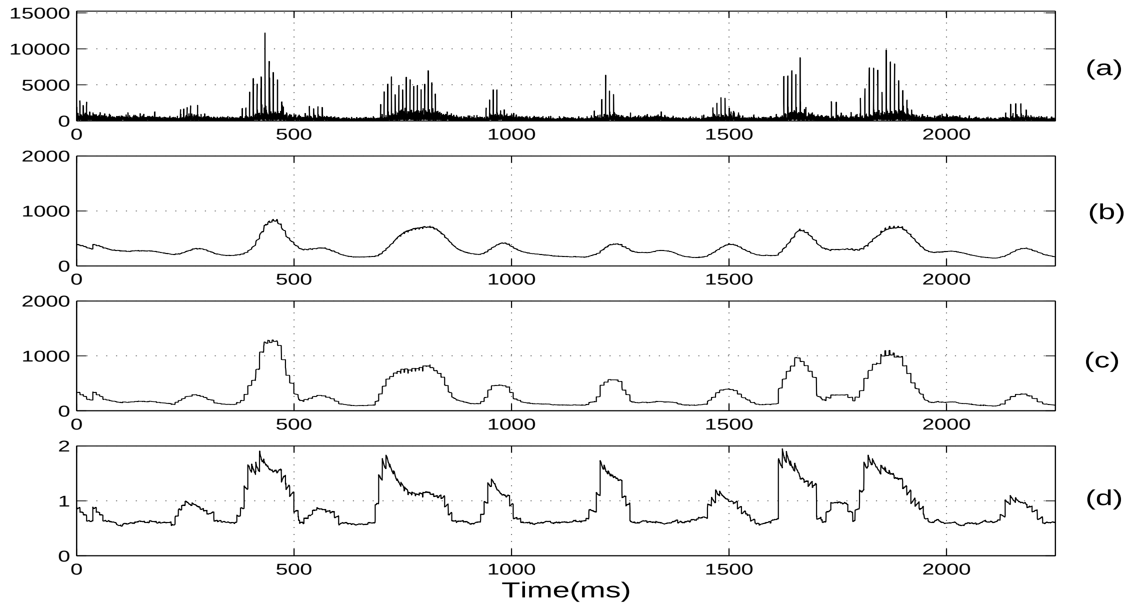

Figure 5.14: (a) Result of coherent addition of Hilbert envelopes of the LP residuals of *mic-1*, *mic-2* and *mic-3* signals, (b) mean values, (c) standard deviation values, and (d) normalized standard deviation values of the resultant Hilbert envelope, computed for every block of 50 ms size with one sample shift.

plot are shown in Figure 5.17. The excitation of speech components are significantly enhanced in the modified residual. The modified residual is used to excite the time varying all-pole filter derived from the coherently-added signal, to synthesize the enhanced speech signal. The clean speech, its degraded version, coherently-added signal from three microphones and the enhanced speech, along with their narrowband spectrograms are shown in Figure 5.18. From this figure, it can be seen that the speech signal is enhanced both with respect to background noise as well as reverberation. This will be further confirmed by subjective and objective evaluations described in the following sections. The degraded and the corresponding enhanced speech signals obtained by the proposed method are available for listening at
http://speech.cs.iitm.ernet.in/Main/result/enhance.html.

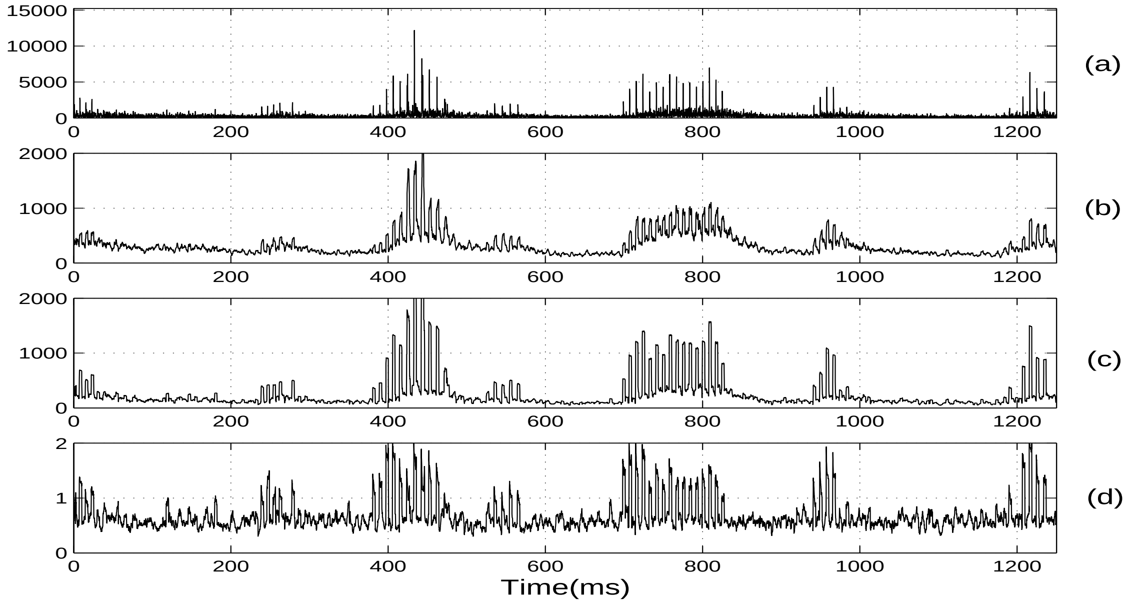

Figure 5.15: (a) Result of coherent addition of Hilbert envelopes of the LP residuals of *mic-1*, *mic-2* and *mic-3*, signals (b) mean values, (c) standard deviation values and (d) normalized standard deviation values with respect to mean for every block of 3 ms frame size and one sample shift for the coherently-added Hilbert envelope.

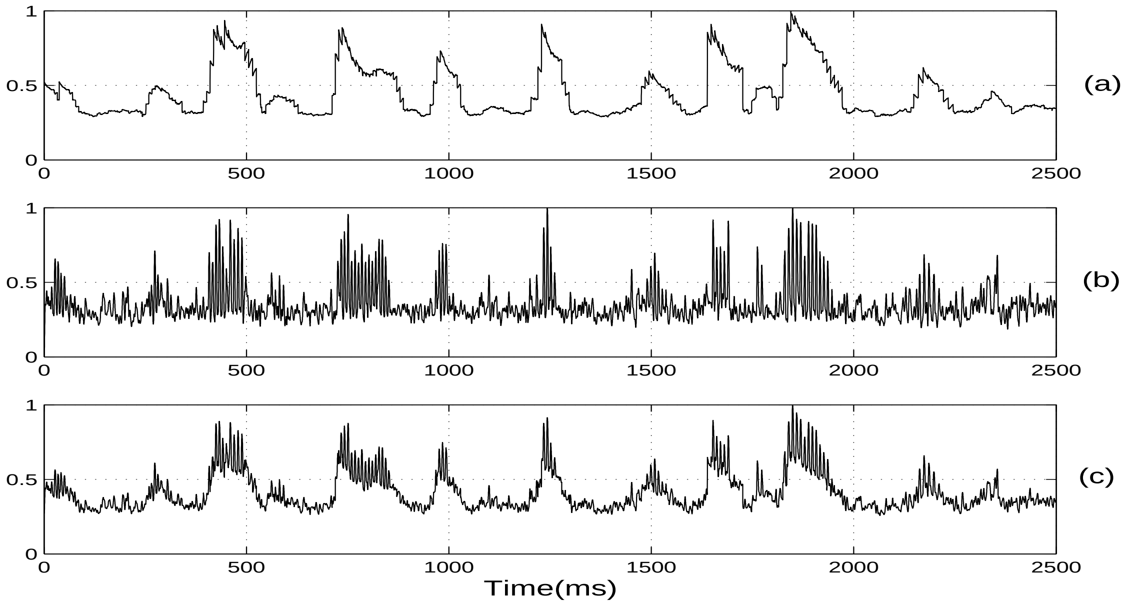

Figure 5.16: (a) Normalized standard deviation plot derived using block size of 50 ms and shift of 1 sample, (b) normalized standard deviation plot derived using block size of 3 ms and shift of 1 sample and (c) weight function obtained by adding (a) and (b).

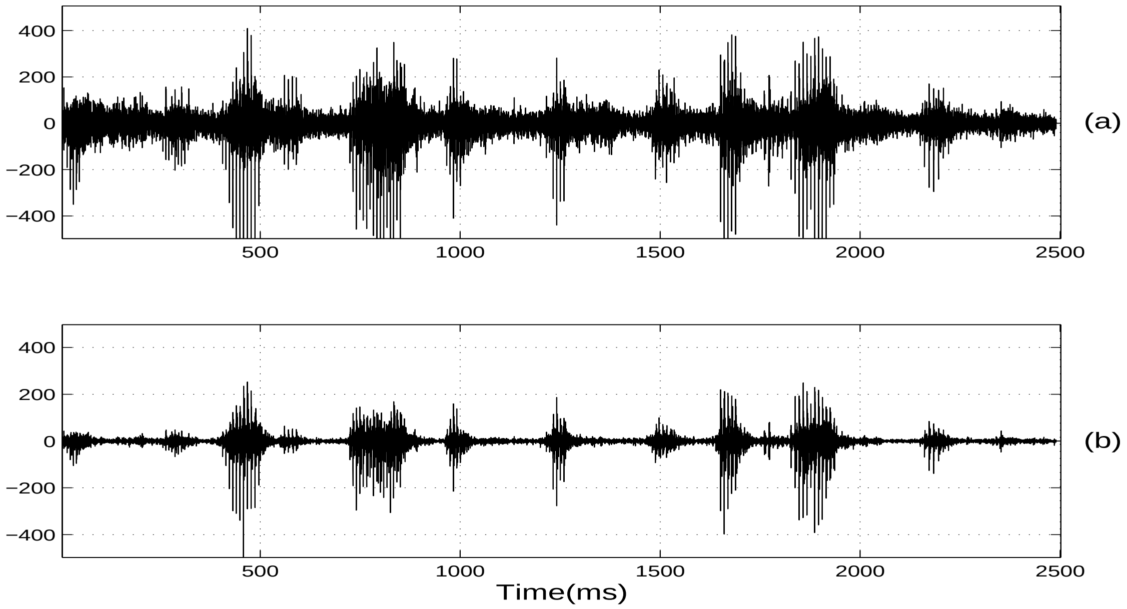

Figure 5.17: (a) LP residual of coherently-added speech signal from three microphones and (b) modified LP residual obtained by multiplying the LP residual in (a) with weight function.

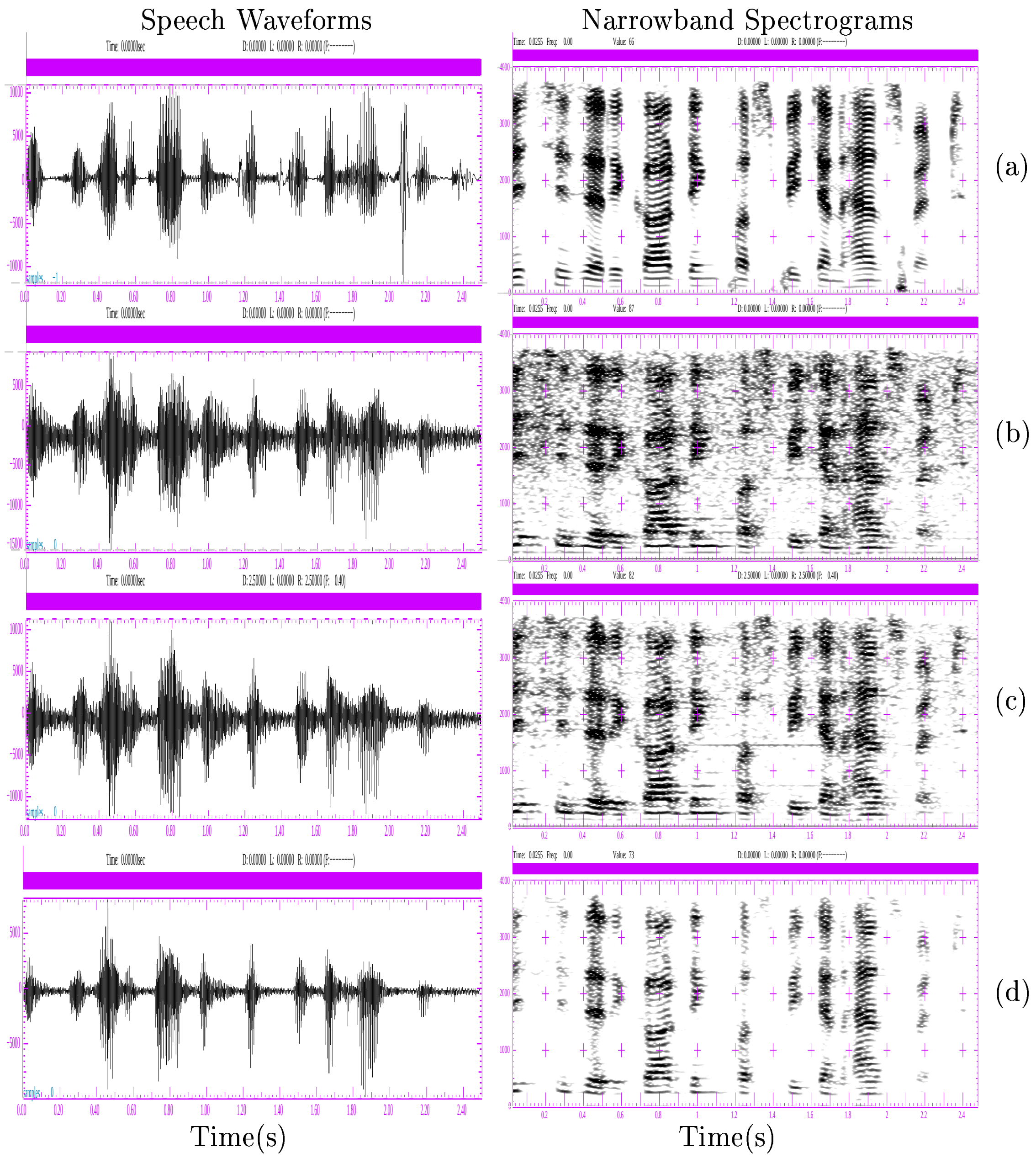

Figure 5.18: Speech and is narrowband spectrograms for (a) close speaking microphone, (b) distant microphone (say, *mic-1*), (c) coherently-added signals from 3 microphones, and (d) enhanced signal by the proposed method.

5.3.3 Performance Evaluation

Subjective Evaluation

Subjective evaluation is performed for assessing the quality of the enhanced speech. The subjective tests were conducted with the help of 10 research scholars in the age group of 21 to 35, who volunteered for the task. Each of the subjects were given a pilot test about the perception of different types of speech signals like clean speech, speech degraded by background noise and speech degraded by background noise and reverberation. They were explained the effect of background noise and reverberation. Once they were comfortable with judging, they were allowed to take the tests. The tests were conducted in the laboratory environment by playing the speech signals through headphones.

Two types of subjective tests were conducted. The first test was to judge the amount of background noise present in the given coherently-added signal by comparing with the degraded speech. Each of the coherently-added signal is processed further by the proposed method to achieve enhancement against reverberation. The objective of the second test was to judge the enhancement achieved against reverberation as compared to the corresponding coherently-added signal.

In the first test, the subjects were asked to judge the enhancement for background noise in each of the 13 different coherently-added signals, on a four point scaling as given in Table 5.1. Reference signals were provided for each point of scaling and the subjects were asked to rank each of the coherently-added signal to the nearest point. The histograms of rankings are shown in Figure 5.19. As shown in the histograms, the ranking for the speech signal increases as the number of microphones is increased for enhancement. This indicates that the enhancement against background noise depends on the number of microphones.

In the second test, the subjects have to judge the amount of reverberation present in each of the enhanced speech, on a four point scaling as shown in Table 5.2. The histograms of ranking for the 13 different enhanced signals are shown in Figure 5.20. The ranking increases rapidly with the number of microphones. Also, it is interesting to note that comparatively high ranking is achieved, using even 5 or 6 microphones. This shows the robustness of the proposed method. It also shows that we can achieve

Table 5.1: Ranking used for judging the quality of enhanced speech for background noise obtained by coherently adding the signals.

Point	Quality of Speech
1.	Sounds like degraded
2.	Sounds slightly better than degraded
3.	Sounds significantly better than degraded
4.	Sounds like clean speech

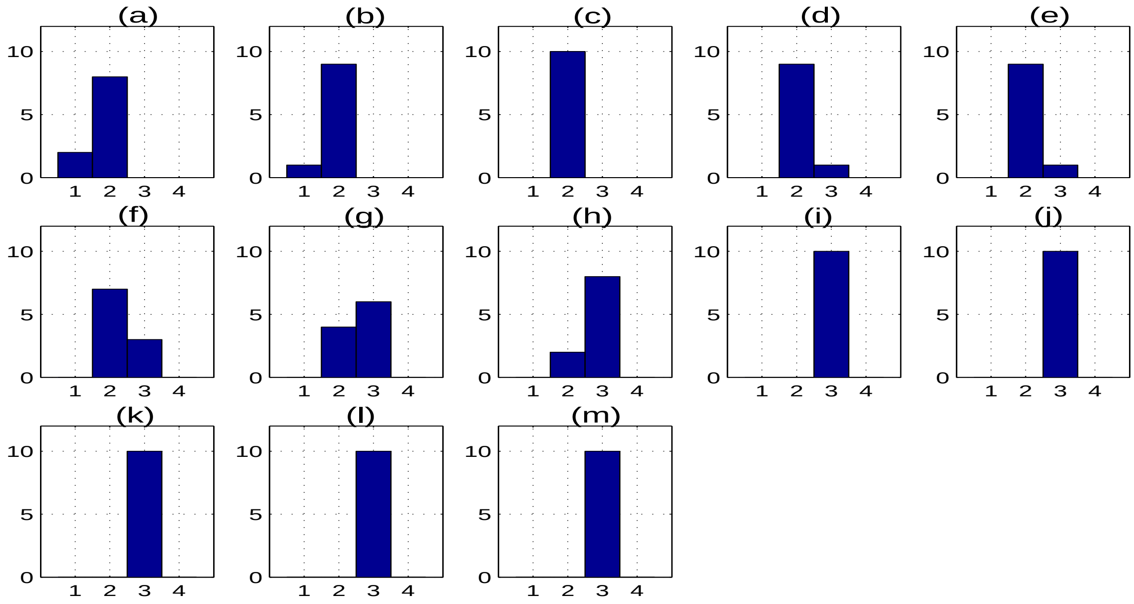

Figure 5.19: Histograms of the rankings obtained for the subjective tests conducted to assess the quality enhancement for background noise in case of the 13 coherently-added signals obtained by adding degraded speech signals from (a) 2, (b) 3, ... (m) 14 microphones. The ranking increases as the number of microphones is increased.

significant enhancement even with fewer microphones. It is also interesting to note from the histogram plots in Figures 5.19 and 5.20 that the proposed method indeed produces improvement over coherently-added speech signals.

Objective Evaluation

A new method is proposed for measuring the degree of enhancement objectively based on the normalized error (η) and is termed as *normalized error measure* (η_n) . The normalized error (η) is the ratio of the residual energy to the signal energy [38,96]. η_n is the total difference between η of the clean speech and η of the enhanced signal. The more the amount of enhancement achieved, the closer will be the enhanced signal

Table 5.2: Ranking used for judging the quality of enhanced speech obtained by processing coherently-added signals further by the proposed method.

Point	Quality of Speech
1.	Sounds like coherently-added
2.	Sounds slightly better than coherently-added
3.	Sounds significantly better than coherently-added
4.	Sounds like clean speech

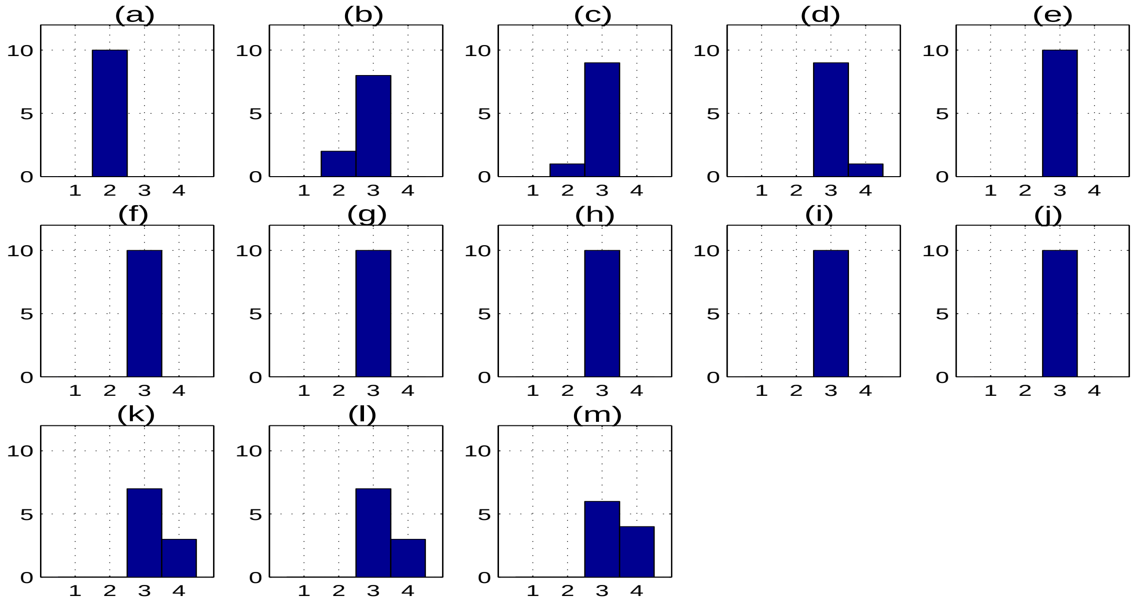

Figure 5.20: Histograms of the rankings obtained for the subjective tests conducted to assess the enhancement against reverberation in case of the 13 different enhanced signals ((a) 2, (b) 3, ... (m) 14 microphones) by the proposed method. The ranking increases rapidly and significant amount of enhancement is achieved using fewer microphones (5 or 6).

to the clean speech and hence lower will be the value of η_n.

The normalized error η for every frame of 5 ms with one sample shift is computed for each of the enhanced speech signals. η_n computed for each of the enhanced signals is shown in Table 5.3. η_n for the enhanced signal derived from a given number of microphones is always lower for the enhanced signal obtained by the proposed method compared to the coherently-added signal. It may appear that weighting the residual gives lower values for η_n in the proposed method. However, it is to be noted that the ratio of the LP residual energy is taken with respect to the corresponding enhanced speech signal. Spectral distance measures may not be useful in the present case, as the proposed method does not alter the spectral information.

Table 5.3: Normalized Error Measure (η_n) for coherently-added speech signals and enhanced speech signals by the proposed method.

Speech from (# *microphones*)	η_n for Coherently added signal	η_n for Enhanced speech by proposed method
2	4611	4500
3	4157	4011
4	4764	4654
5	3909	3801
6	4322	4182
7	3630	3498
8	4266	4127
9	3352	3184
10	3721	3608
11	3113	2946
12	3519	3377
13	3160	2991
14	3265	3091

5.4 Speech Enhancement in Multispeaker Environment

In a multispeaker environment the objective is to separate the speech component corresponding to each speaker, while retaining the quality and intelligibility as much as possible. The signal collected by a microphone in a multispeaker environment is a mixture of speech signals from several speakers. Processing speech for enhancement in such conditions is a challenging task, as the speech of the other speakers acts 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. The difficulty is further compounded by the fact that the spectral characteristics are modified by the response of the room and also by the background noise. The extent of degradation depends on the relative position of the microphone with respect to the speaker and also on the background noise. The primary causes of degradation are room reverberation, background noise and the distance of the speaker from the microphone.

Most of the speech message is carried through the voiced part of speech, especially when the microphone is far off from the speaker, say 2 meters or more. Even though the spectral component of speech is severely degraded, the characteristics of the quasi-periodic excitation are well preserved in the direct speech picked up by a distant microphone. It is true that reflected and delayed speech signal is also added to the direct speech signal. If the strength of the excitation is low, then, even voiced speech will sound more like whispered speech and hence cannot be perceived at long (say 2 m or more) distances from the speaker. Note that in the case of reverberation, scaled versions of the speech signal are added to the direct signal at random instants. These random instants are different for microphones placed at different locations. Thus if the delay between the direct speech signal components at two microphones is compensated, then the strengths of the instants due to direct speech are reinforced, and simultaneously the strengths due to reflected speech components are distributed in time. It is also important to note that, since no two speakers can be at the same location simultaneously, the delays due to the direct speech components between a pair of microphones from any two speakers are different. These properties of speech production and propagation of sound in rooms form the basis for the proposed method for speech enhancement from multispeaker speech.

Most of the multispeaker enhancement methods in the literature involve modification of spectral features representing the vocal tract system [69–74,108]. They use the knowledge of pitch to separate the individual speakers in multispeaker case. Hence the performance of these methods depends on the accuracy of the estimated pitch. Estimation of pitch from degraded speech is a difficult task in itself. Moreover, pitch is only one feature of the excitation of the vocal tract system. In this work, we propose a method of speaker separation from speech collected over multiple microphones, using other important characteristics of excitation of voiced speech. In particular, we exploit the characteristics of the strength of excitation at the GC events, and the robustness of the relative spacing of the GC events in the speech signals collected at different microphones. We use Hilbert envelope of the LP residual as a representation for the sequence of impulses corresponding to the instants of significant excitation of the vocal tract system. 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. Using the knowledge of the enhanced speaker characteristics in the coherently-added sequence of impulses, a weight function is derived, which in turn 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 signal. Enhancement in the resulting speech is primarily due to enhancement of the excitation characteristics, which are important perceptually.

5.4.1 Usefulness of GC Events for Speaker Separation

The large amplitude peaks of Hilbert envelope of LP residual occur mostly around the GC events. Even if some of them do not occur exactly at the GC events, it is not critical, as long as the intervals between the peaks due to successive GC events remain the same. There could be some spurious peaks in the Hilbert envelope of the residual. These are mostly due to the effects of noise and reverberation. The Hilbert envelope due to an impulse has a point property, in the sense that the peak of the Hilbert envelope is due to the residual samples at that instant and at the instants immediately adjacent to it. Therefore, peaks in Hilbert envelope of the LP residual occur at the same relative instants due to the direct speech at all the microphones, whereas the peaks in the Hilbert envelope due to noise and reverberation occur at different instants at different microphones.

Their is a unique time-delay between the signals from two microphones for each speaker. This delay is estimated from the microphone signals using the Hilbert envelopes of the LP residuals. The time-delays, estimated for the speech signals collected over two microphones using a frame size of 50 ms and a frame shift of 10 ms, are shown in Figure 5.21. In the recorded data, there is speech from two speakers (*spkr-1* and *spkr-2*). It can be seen that there are two prominent delay values (*delay-1* and *delay-2*) corresponding to each speaker. The random delay values in the plot generally correspond to nonspeech regions. The two delays are obtained by summing the number of evidences for each delay and then considering the two delays with highest number of evidences.

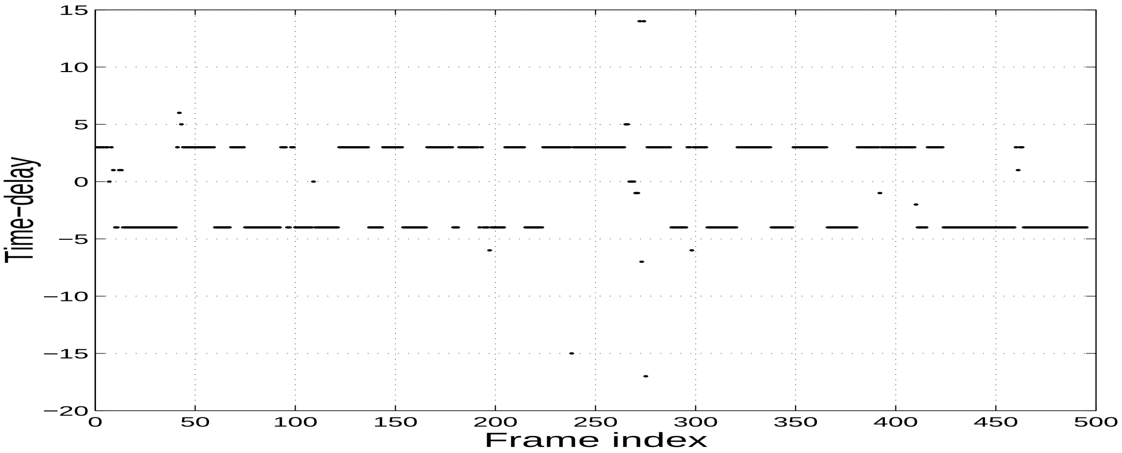

Figure 5.21: Time-delays computed using the Hilbert envelopes of two microphones for every frame of 50 ms with a shift of 10 ms. The two main delay values correspond to the two speakers.

The Hilbert envelopes of the LP residuals of speech signals from the two microphones are added after compensating for the delays and the results are shown in Figure 5.22. When the Hilbert envelopes are added after compensating for *delay-1*, the excitation source information of the corresponding speaker is reinforced in the coherently-added signal. Similarly the excitation source information of the second speaker is reinforced in the coherently-added signal using *delay-2*. This property of reinforcement of the excitation information is exploited for separation of speech of individual speakers.

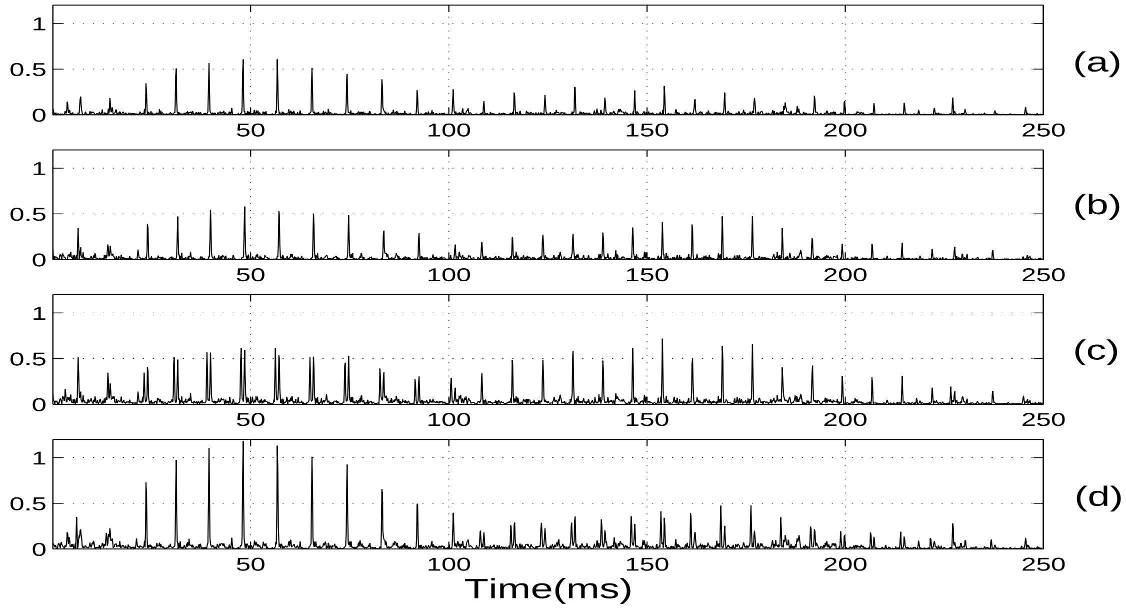

Figure 5.22: Hilbert envelopes derived from two microphone signals. Hilbert envelope of the LP residual of (a) *mic-1* signal, (b) *mic-2* signal, (c) Coherently-added Hilbert envelope using *delay-1* and (d) Coherently-added Hilbert envelope using *delay-2*.

Note that the reinforcement of the excitation information of a given speaker in the coherently-added Hilbert envelopes of the LP residuals can be best seen at the peaks around the GC events. While it may be difficult to give an objective measure, one can easily see from the plots of the coherently-added Hilbert envelopes that the peaks for the desired speaker reinforce and the peaks for the other speaker are spread out. One can determine the effect of coherence in the plots of the standard deviation for the four cases shown in Figure 5.23. The standard deviation plot is obtained by computing it for every 2 ms interval around every sample, corresponding to 17 samples. The choice of 2 ms is not critical. Any small interval less than a pitch period but greater than twice the delay is adequate. The standard deviation plots clearly show that the values are significantly reduced in the regions where the addition of the peaks in the Hilbert envelopes is incoherent. For the example shown in Figure 5.23, the coherent and incoherent regions are somewhat nonverlapping. On the other hand Figure 5.24 is an illustration of a segment where there is significant overlap of the speech from the two speakers. Even in this case, one can observe the coherent and incoherent regions, as there is a significant reduction in the values of the standard deviation in the incoherent regions.

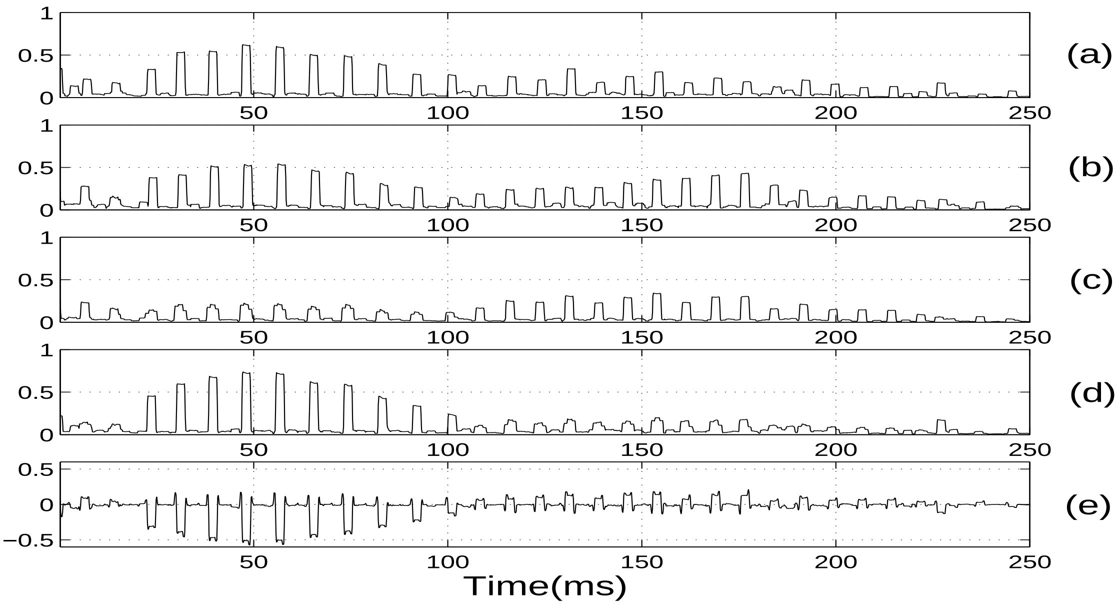

Figure 5.23: Standard deviation plots computed using a frame size of 2 ms with a shift of one sample for the Hilbert envelopes shown in Figure 5.22. Standard deviation of the Hilbert envelope of (a) LP residual of *mic-1* signal, (b) LP residual of *mic-2* signal, (c) coherently-added using *delay-1*, (d) coherently-added using *delay-2* and (e) difference between the two standard deviation plots shown in (c) and (d).

5.4.2 Speech Enhancement from Multispeaker Data

The proposed method of speaker separation exploits the characteristics of the coherently-added Hilbert envelopes of the LP residuals of speech from a pair of microphones. For example, by subtracting the standard deviation values of the coherently-added Hilbert envelopes of both the speakers, we get the waveforms shown in Figures 5.23(e) and 5.24(e) for the two segments. The regions of significant excitation for the desired speaker are indicated by the positive pulses, and for the other speaker, by the negative pulses. To derive the regions of significant excitation of the desired speaker, the positive pulses of width greater than 1 ms and standard deviation difference values

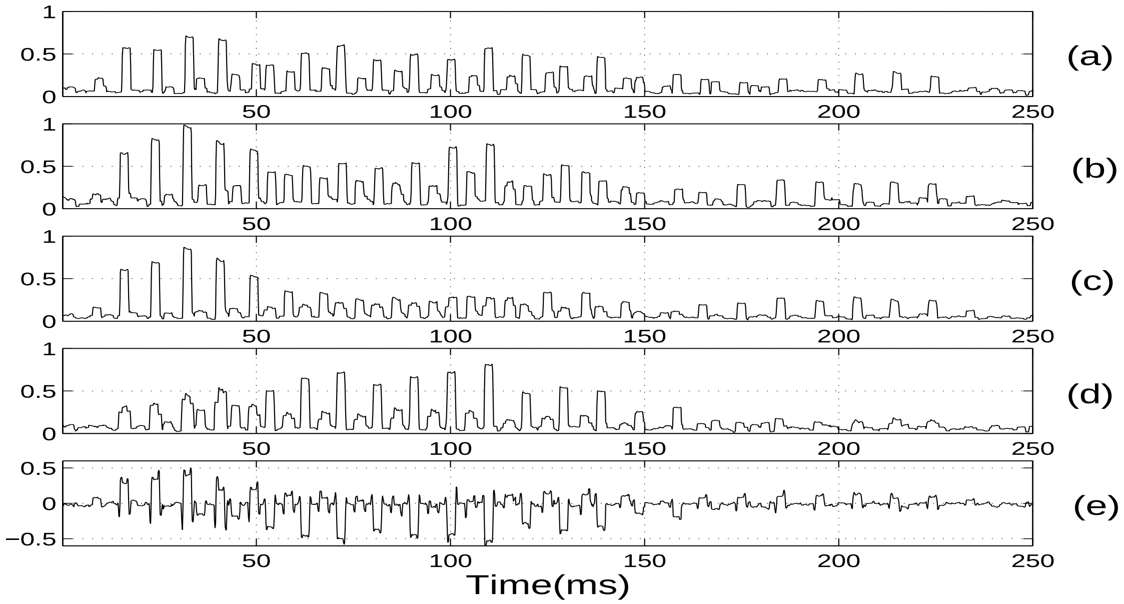

Figure 5.24: Standard deviation plots computed using a frame size of 2 ms with a shift of one sample for the Hilbert envelopes of a overlapping region. Standard deviation of the Hilbert envelope of (a) LP residual of *mic-1* signal, (b) LP residual of *mic-2* signal, (c) coherently-added using *delay-1*, (d) coherently-added using *delay-2*, and (e) difference between the two standard deviation plots shown in (c) and (d).

greater than the mean of the positive values are used. The choice of 1 ms is related to the delay, as it is necessary to ignore the spurious positive regions. The choice of the small positive threshold is to avoid spurious positive regions. To derive the regions of significant excitation of the second speaker, similar logic is used for the plot in Figure 5.23(e) to pick up regions of negative pulses. The positive and the negative pulses for the case in Figure 5.23(e) are shown in Figures 5.25(b) and 5.25(c). The sequences of positive and negative pulses are smoothed using a 5-point mean smoothing to avoid abrupt changes in the weight function. The width of the smoothing window is not critical. In order to suppress the undesired speaker, it is preferable to have more amplitude for the negative pulses. In this case a value of -4 is chosen. This value was chosen after listening to the enhanced speech for different values of the amplitudes for the negative pulses. While the chosen value is not critical, too small a value does not reduce the level of the second speaker and on the other hand, too large a value will produce distortion of the desired speaker. Ideally one should choose the negative weight value adaptively based on the strength of the GC event of the undesired speaker. In this work we have chosen a fixed threshold value of -4. The resulting plots are shown in Figures 5.25(d) and 5.25(e). The final weight function is derived from this plot by taking exponential of the sum of the plots in Figures 5.25(d) and 5.25(e), and normalizing the maximum value to one. Figure 5.25(f) gives the desired weight function.

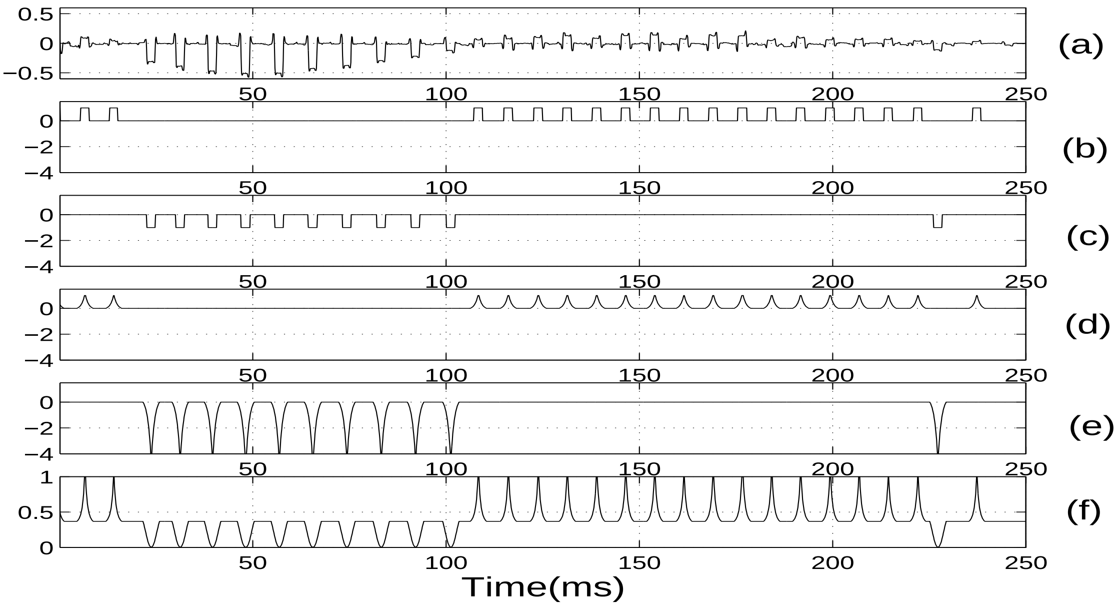

Figure 5.25: Illustration of the steps involved in deriving the weight function. (a) The plot of difference of standard deviations shown in Figure 5.23(e) is reproduced here. (b) Positive pulses indicate regions of significant excitation of the desired speaker. (c) Negative pulses indicate regions of significant excitation of the undesired speaker. (d) Smoothed version of (b). (e) Smoothed version of (c). (f) Normalized weight function derived using the positive and negative pulses shown in (d) and (e).

This weight function is used to multiply the LP residual of the speech from *mic-1*. The resulting modified residual is used to excite the time varying all-pole filter obtained from the LP analysis. Thus we get the processed speech from *mic-1*, where the desired speaker is enhanced. Similarly the speech from *mic-2* is processed to enhance the same speaker. Both these enhanced speech signals are coherently added to obtain the enhanced signal of the desired speaker. This coherent addition at the signal level helps to reduce the effect of background noise in the degraded signals. For the enhancement of speech of the desired speaker no special attempt is made for the unvoiced regions. Perceptually it may not be necessary to separate the speech of each speaker in these regions. Since the weight values for the unvoiced region is less than the weight values for the undesired speaker, the unvoiced segments are also clearly perceived, even though the speakers are not separated in those regions. The steps in processing the two microphone signals for enhancement are summarized in Table 5.4.

Table 5.4: Steps in the proposed method for speech enhancement in multispeaker environment

Sl. No.	Description
1	Collect the speech signals (sampling frequency 8 kHz) from two speakers over two spatially separated microphones in a live room.
2	Derive the (10^{th} order) LP residuals from the speech signals.
3	Compute the Hilbert envelopes of the LP residuals.
4	Estimate the time-delays for each speaker using the cross-correlation of the Hilbert envelopes.
5	Add the Hilbert envelopes using the estimated time-delays to produce the coherently-added Hilbert envelope for each speaker.
6	Derive the weight function using the standard deviation plots from the coherently-added Hilbert envelopes.
7	Derive the modified LP residual signal from each microphone signal.
8	Synthesize the enhanced speech for each microphone signal.
9	Coherently add the speech signals of the desired speaker derived from both the microphone signals.

5.4.3 Experimental Results

To compare the results of the proposed method for enhancement with other methods, we consider an example (Case 1B) from Independent Component Analysis (ICA) database prepared for evaluating new methods for multispeaker speech enhancement in multichannel case [109]. The example consists of speech of two male speakers, collected for 10 seconds over two microphones in a room of dimension 3.4m × 3.8m × 2m. The room was not completely anechoic, as it did have a short reverberation. Information regarding distance between the speakers and the microphones was not given in the description of the database. The outputs of the microphones and the processed speech signals of each speaker are given in the database, and are used as reference for comparison in this study. The two microphone signals are processed by the proposed method to enhance the speech of each speaker. The signals from the individual microphones, the enhanced speech signals of the speakers available in the database, and the enhanced speech signals obtained by the proposed method are shown in Figure 5.26.

The segments of the speech signals shown in Figures 5.26(d) and (f) are due to the new method. The narrowband spectrograms for all the speech signals are also given in Figure 5.26. From the spectrograms it is difficult to see the result of enhancement. But the waveforms in Figures 5.26(d) and (f) clearly show the separation of the speakers. Perceptually, the proposed method gives a significantly better suppression of the undesired speaker and also better quality than the enhancement results given in the database. Note that even though the spectrum is not manipulated, the enhancement of the speaker is achieved purely by emphasizing the regions of excitation around the GC events of the desired speaker and at the same time deemphasizing the regions around the GC events of the undesired speaker.

Speech data from two speakers speaking simultaneously was collected from two spatially separated (0.6 m) microphones in a laboratory. The microphones are at distances of over 1 m from each of the speakers. The speech data for 20 seconds collected from two male speakers is processed by the proposed method, and the results are shown in Figure 5.27. The separation of the speakers is evident in the waveform plots. Speech from two female speakers, and one male and one female speakers (each of 20 seconds) were also collected and processed by the proposed method. In all the cases the speech of the desired speaker is found to be significantly enhanced compared to the respective degraded signals. These results indicate that it is possible to separate multispeaker data collected at multiple microphones using the proposed method. The wave files for these cases are available for listening at the website *http://speech.cs.iitm.ernet.in/Main/result/multispkr.html.*

5.4.4 Subjective Evaluation

In this section, results of the listening tests are given. Mean Opinion Score (MOS) rating method was used to asses the quality of the degraded and processed speech signals [110]. In this method, subjects were asked to rate the speech under test on a five-point scale given in Table 5.5 [110].

The subjects for this evaluation consisted of 25 graduate students who volunteered for the task. All the subjects are familiar with speech processing as they have taken a full semester course on speech technology. The evaluation was conducted by play-

Table 5.5: Mean opinion score five-Point Scale.

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory	Very annoying and objectionable

ing the speech signals through a loudspeaker in the laboratory environment. Initially the subjects were presented with degraded and the processed speech signals, different from those used for evaluation, to give familiarity about the rating on the five-point scaling. In the formal evaluation, the degraded speech signals from the two microphones are presented followed by the enhanced speech signals. The subjects were asked to give their rating for each case. The summary of their ratings is shown in the form of histograms in Figure 5.28, for the example considered from the ICA database (Figure 5.26). The rating is lowest for the degraded speech and highest for the speech enhanced by the proposed method. The enhanced speech signals obtained by the proposed method shows better rating compared to the enhanced speech signals given in the database. The mean opinion scores are 3.03, 2.03 and 1.40 for the enhanced speech signals (*spkr-1* and *spkr-2*) by the proposed method, enhanced speech signals (*spkr-1* and *spkr-2*) given in the database and degraded speech signals (*mic-1* and *mic-2*), respectively. Thus, this study show that the proposed method is effective for enhancement of multispeaker speech, and the knowledge of the excitation source is indeed useful for processing speech for enhancement. The summary of the ratings for the data collected in the laboratory environment (three examples, namely, two male, two female, one male and one female) is given in the histograms in Figure 5.29. The mean opinion scores for the enhanced speech is 3.58 and it is 1.70 for the degraded speech. This example illustrates the effectiveness of the proposed method for processing speech collected in different environmental conditions.

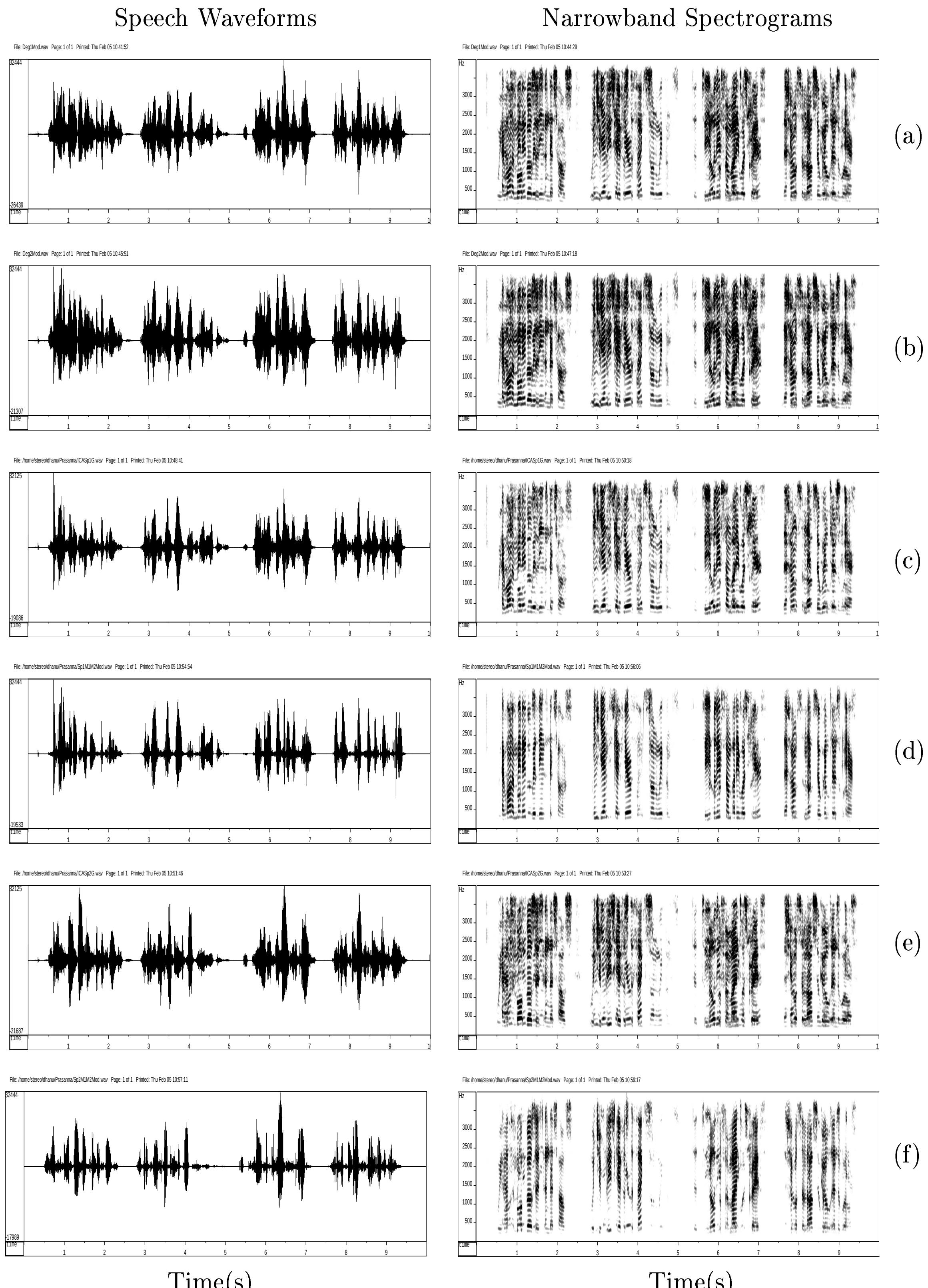

Figure 5.26: Speech waveforms and narrowband spectrograms for (a) *mic-1* signal, (b) *mic-2* signal, (c) enhanced *spkr-1* signal (given), (d) enhanced *spkr-1* signal (proposed), (e) enhanced *spkr-2* signal (given) and (f) enhanced *spkr-2* signal (proposed) (The speech data was collected from ICA99 database).

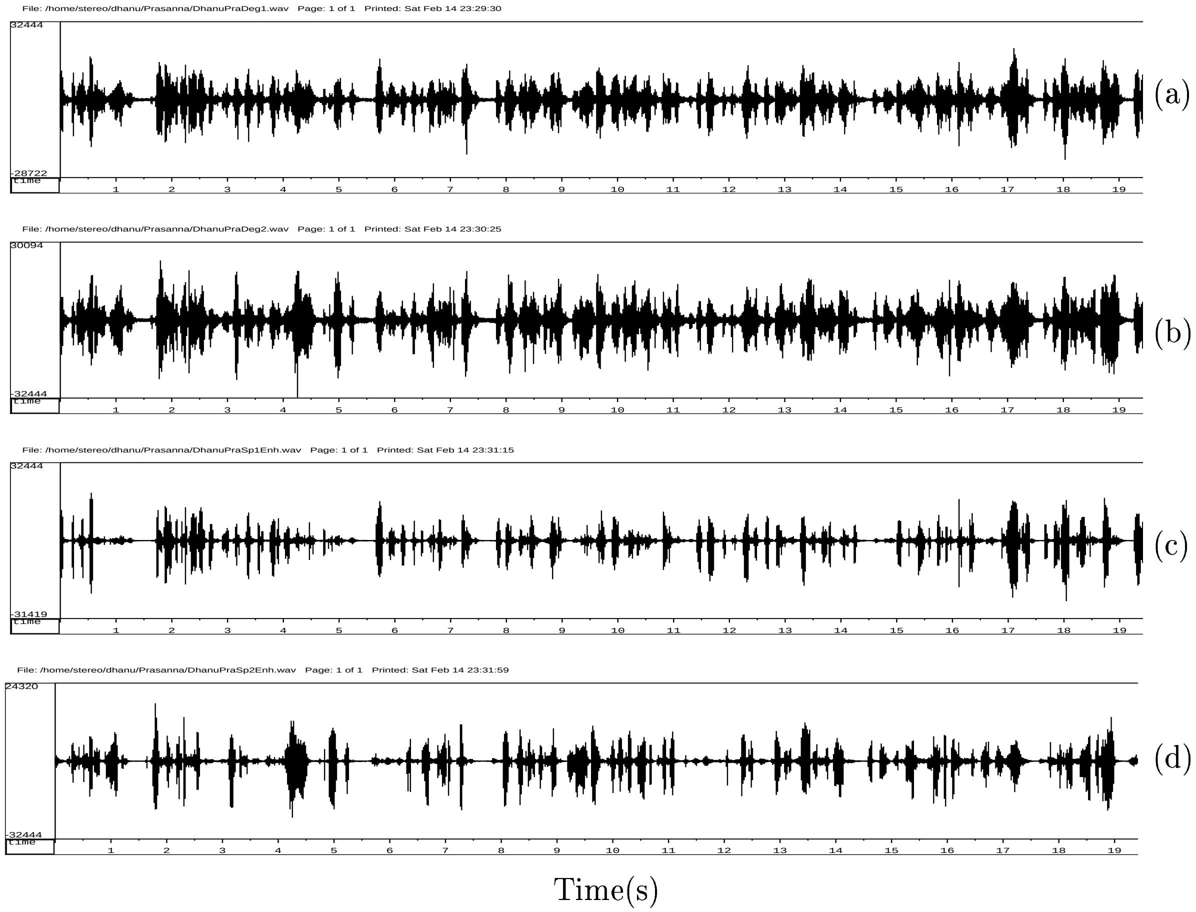

Figure 5.27: Speech waveforms of two male speakers collected in the laboratory. (a) *mic-1* signal, (b) *mic-2* signal, (c) enhanced *spkr-1* signal and (d) enhanced *spkr-2* signal.

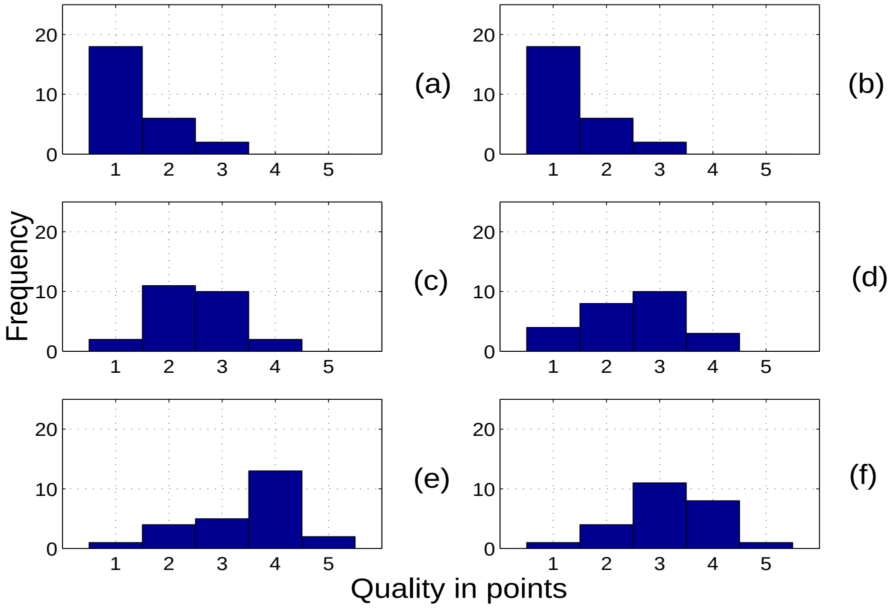

Figure 5.28: Frequency histogram showing the frequency distribution of the scores given to the quality of speech on a five-point scale for (a) degraded speech from *mic-1*, (b) degraded speech from *mic-2*, (c) enhanced speech of *spkr-1* (given), (d) enhanced speech of *spkr-2* (given), (e) enhanced speech of *spkr-1* (proposed) and (f) enhanced speech of *spkr-2* (proposed) for the signals from ICA99 database.

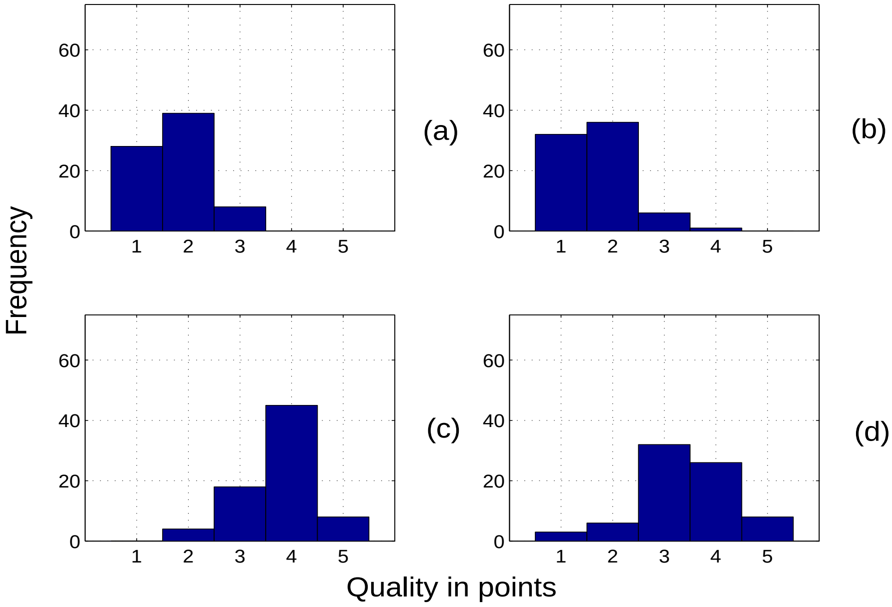

Figure 5.29: Frequency histogram showing the frequency distribution of the scores given to the quality of speech on a five-point scale for (a) degraded speech from *mic-1*, (b) degraded speech from *mic-2*, (c) enhanced speech of *spkr-1*, and (d) enhanced speech of *spkr-2* for the signals collected in the laboratory environment.

5.5 Summary

In this chapter a method is proposed for estimating the time-delays from speech signals collected over spatially distributed microphones. The method uses the knowledge of the excitation source, unlike the commonly used spectrum-based methods. Since time-delays can be estimated accurately even from short segments of speech, it is also possible to develop algorithms to track a moving speaker. A method for enhancement of speech in multichannel case is proposed. The Hilbert envelope of the LP residual signal is used for developing the method for enhancement. The resulting coherently-added Hilbert envelope exhibits some important properties. Using these properties a weight function is derived, which when multiplied with the LP residual of the coherently-added signal, enhances the high SNR regions. The enhanced residual is used to excite the time varying all-pole filter, which is derived from the coherently-added signal, to synthesize speech signal. A method based on the knowledge of the excitation source of speech production is also proposed for enhancing speech of the desired speaker in a multispeaker environment. The enhancement is achieved by deriving a modified excitation sequence for each speaker and synthesizing the speech signal using this sequence. It was found that in the synthesized signal, the speech of the desired speaker is enhanced significantly compared to that of the other speaker. A summary of the various issues discussed in this chapter is given in Table 5.6.

In Chapter 3, methods for the detection of GC events were discussed. In the previous and this chapter, usefulness of GC events was illustrated in some applications. In the following two chapters we discuss the detection of the VOP events and usefulness of the VOP events in the detection of end-points.

Table 5.6: Summary of the issues discussed with respect to the applications of GC events for Multichannel case.

Time-Delay Estimation

- Issues in Time-Delay Estimation
  - Existing methods use vocal tract system features, which are easily affected by degradation.
  - Long segments, typically 500-1000 ms are used for delay estimation to minimize the effects of degradation. But considering long segments is not advisable for applications like tracking moving speakers.
- Proposed method for Time-Delay Estimation
  - Uses excitation source information of speech production, which is robust to degradation, that is, the relative spacing between the GC events will not change due to degradation.
  - Hilbert envelope of the LP residual is used as the excitation source information.
  - Cross-correlations of segments of size 50-100 ms are sufficient for estimating the delay.

Speech Enhancement in Multichannel Case

- Issues in Speech Enhancement in Multichannel Case
  - Estimation of time-delay using spectral features.
  - Estimating characteristics of degradations and subtracting the same from the degraded speech.
- Proposed method for the enhancement in multichannel case
  - Time-delay estimation using excitation source information.
  - Exploiting excitation source information derived from the GC events.
  - Identify and enhance excitations of speech component.

Speech Enhancement in Multispeaker Environment

- Issues involved in speech enhancement in multispeaker environment
  - Existing methods use vocal tract system features.
  - Existing methods use the knowledge of pitch and estimating pitch in degraded conditions is a difficult task in itself.
- Proposed method for the enhancement in multichannel environment
  - Uses excitation source information based on the knowledge of GC events.
  - Pitch information is not used, and only the fact that the relative spacing between the GC events will not change, is exploited.

Chapter 6

VOWEL ONSET POINT EVENT FOR SPEECH ANALYSIS

In the previous three chapters, the proposed event-based approach was demonstrated using GC events. Acoustic cues were proposed for the detection of GC events. A method for automatic detection of GC events using group delay functions was discussed. The group delay based method detects GC events accurately. The Hilbert envelope of the LP residual gives approximate information of the GC events. The usefulness of GC events was demonstrated in applications like extraction of pitch, enhancement of speech, time-delay estimation and enhancement of speech in multi-speaker environment. The significantly improved performance in these applications illustrates the potential of the proposed event-based approach. In this and the next chapter, the event-based approach is demonstrated using VOP events. In particular, this chapter discusses the issues involved in the detection of VOP events.

The organization of this chapter is as follows: Issues involved the detection of the VOP events are discussed in Section 6.1. The acoustic descriptions of VOP events for different categories of CV units are given in Section 6.2. Section 6.3 proposes a set of acoustic cues for the detection of VOP events. Preparation of the reference database using the proposed acoustic cues is discussed in Section 6.4. Section 6.5 gives an algorithm for automatic detection of VOP events for isolated utterances of CV units. An algorithm for the detection of VOP events in continuous speech is proposed in

Section 6.5.2. In Section 6.6, a summary of the issues discussed in this chapter is given.

6.1 Issues in the Detection of VOP events

Important information for the analysis of speech lies around the VOP event, and hence a reliable algorithm for the automatic detection of the VOP event is essential. The proposed algorithm needs to be evaluated against a reference database, containing speech data with manually marked VOP events.

Careful observation of the characteristics of different CV units shows that there are some categories of CV units which are difficult even for manual marking of VOP event. The cues normally used for manual marking are the amplitude of the signal, voicing level and pitch periodicity of the vowels. For difficult cases of CV units such as voiced aspirated, nasal and semivowel CV units, it is necessary to use other cues for manual marking. Even to select a set of acoustic cues, it is essential to know the characteristics of the signal around the VOP event for each category of sound unit. Hence a study of the characteristics of the signal around the VOP event is made to obtain the acoustic description of the VOP event for each category of CV units.

In most of the methods described in the literature for the detection of VOP events, the features of the vocal tract system are used. But there is significant information in the excitation source features too, which may be exploited to determine the VOP event. There are some sound units, especially in Indian languages, like unvoiced aspirated and voiced aspirated sounds, in which the onset of vowel may be determined better using the excitation source features, as there are very few changes in the characteristics of the vocal tract system at the VOP event. Even though the onset of vowel is an instant property, most of the existing methods treat the VOP event as a region property, and hence the results are poor in resolution. The aim of this work is to develop a method for accurate detection of VOP events, and also explore the usefulness of excitation source features in the detection of VOP events.

6.2 Acoustic Description of the VOP Events

Even though not exhaustive, the most frequently used 145 CV units of the Indian language Hindi are chosen for this study. The 145 CV units may be broadly classified into Stop Consonant Vowel (SCV) units, nasal, fricative, affricate and semivowel CV units. The SCV units may be further classified depending on the type of vowel, Place of Articulation (POA) and Manner of Articulation (MOA). In this study the MOA criterion is used for classification. According to this criterion the SCV units can be further divided into Unvoiced Unaspirated (UVUA), Unvoiced Aspirated (UVA), Voiced Unaspirated (VUA) and Voiced Aspirated (VA) CV units. The different categories of the 145 CV units are shown in Table 6.1. Note that the CV units are shown only for the vowel ending /a/. Similar CV units exist for the vowel endings /i/, /u/, /e/ and /o/. In this section the acoustic descriptions of the VOP event in each category of the CV units are given. This description helps in the identification of suitable acoustic cues for the detection of the VOP event.

Table 6.1: Categories of CV units in Hindi.

Category of CV units	Sub-Category	Sound units
SCV	UVUA	/ka/, /Ta/, /ta/, /pa/
	UVA	/kha/, /Tha/, /tha/, /pha/
	VUA	/ga/, /Da/, /da/, /ba/
	VA	/gha/, /Dha/, /dha/, /bha/
Nasal		/na/, /ma/
Semivowel		/ya/, /ra/, /la/, /va/
Fricative		/sa/, /sha/, /ha/
Affricate		/cha/, /Cha/, /ja /, /Jha/

For the UVUA CV units, the VOP event is characterized by the changes in the source as well as the system characteristics. The change in the source is from the burst release to the glottal vibration (quasi-periodic laryngeal source). The change in the system characteristics is from total closure to wide opening. The consonant region is characterized mainly by the burst region, and the vowel region is characterized by quasi-periodic signal and regular formant contours.

The VOP event in the case of UVA CV units is characterized mainly by the change in the source characteristics. This is because, the change in the vocal tract shape occurs at the onset of aspiration, which is ahead of the VOP event, and hence there will be few changes in the system characteristics. The change in the source is from the noise due to turbulent airflow at the glottis to the quasi-periodic glottal vibration. The consonant region is characterized by burst and aspiration. The vowel region is characterized by a quasi-periodic signal and regular formant contours.

For the VUA CV units, changes in both source and system characteristics occur at the VOP event. The change in the system characteristics is from total closure to wide opening. Even though the glottal vibration remains the same during the production of consonant as well as vowel, its characteristics change due to loading of the system on the glottis during the production of the consonant. The consonant region is characterized by low voiced region and weak frequency formant structure. The vowel region is characterized by high voiced region and regular formant contours.

The VOP event in the VA CV units is characterized by the changes mainly in the source characteristics only. Major changes in the system characteristics occur at the beginning of the aspiration region. In the initial part of the consonant, there is only glottal vibration, superimposed by the additional noise source at the glottis in the aspiration region. This is followed by change only due to the glottal vibration at the VOP event. The consonant region is characterized by burst and aspiration. The vowel region is characterized by high voiced region and regular formant contours.

In case of nasal CV units, the VOP event is characterized by changes in the source and system characteristics. The change in the source characteristics is due to the loading effect of system on the glottis. The change in the system characteristics is from total closure to wide opening of the oral cavity accompanied by coupling and decoupling of nasal cavity, respectively. The consonant region is characterized by a comparatively low voiced region and low frequency formant structure. The vowel region is characterized by high voiced region and regular formant contours.

The VOP events in fricative and affricate CV units may be characterized by the changes in the source and system characteristics. The change in the source is from noise due to turbulent airflow at a narrow constriction, to glottal vibration. The change in the system characteristics is from narrow constriction to wide opening. The consonant region is characterized by a noise-like signal and high frequency formant structure due to frication, followed by a quasi-periodic signal and regular formant contours in the vowel region.

The VOP event in semivowel CV units is also characterized by changes in the source and the system characteristics. The change in the source characteristics is due to the loading effect of the system on the glottal vibration during the production of the consonants. The change in the system characteristics is from partial opening to wide opening. The amount of loading is less compared to other sound units like nasal CV units, due to partial opening of the vocal tract system during the production of the consonant. Semivowels are characterized by the formant contours both in the consonant as well as in the vowel region with transition at the VOP event.

To summarize, in all the categories of the CV units, there are changes in the excitation source characteristics at the VOP event. Thus the knowledge of the excitation source features may also be explored for detecting the VOP event. In VA, nasal and semivowel CV units, the similarity in the characteristics of the signal in the consonant and the vowel regions makes it difficult to detect the VOP event directly from the signal. Hence there is a need for exploring a new set of acoustic cues which shows significant change at the VOP event in these categories.

6.3 Acoustic Cues for the detection of the VOP Events

In this study GC events (instants of significant excitation) are used as the pitch markers, and acoustic cues based on the GC events are explored [42, 43]. These acoustic cues will have good temporal resolution, consistency in representation and robustness. The different acoustic cues considered are: (1) Formant transition ($F_{tr}(t)$), (2) epoch intervals ($E_i(t)$), (3) strength of instants ($S_i(t)$), (4) symmetric Itakura distance ($I_d(t)$) and (5) ratio of signal energy to residual energy ($S_r(t)$). Among these, the cues $F_{tr}(t)$, $I_d(t)$ and $S_r(t)$ indicate changes in the vocal tract system characteristics and the cues $E_i(t)$ and $S_i(t)$ indicate changes in the excitation source characteristics.

6.3.1 Formant Transition

The acoustic cue formant transition ($F_{tr}(t)$) can be obtained from the plot of formant contours ($F_i(t)$) plot. In this study the method based on the instants of significant excitation proposed in [2] is used for deriving the formant frequencies. Knowledge of the instants of significant excitation enables us to choose the position and size of the analysis frame within a pitch period in such a way that consistent results can be obtained [2]. Frames of 3 ms duration immediately after the instants of significant excitation are used for LP analysis.

The formant frequencies are derived from the roots of the prediction polynomial

$$A(z) = a_0 + a_1 z^{-1} + ... + a_p z^{-p} \tag{6.1}$$

where a_k, $k = 0, 1, 2, ...p$ are the Linear Prediction Coefficients (LPCs) estimated using the covariance method [46]. The covariance method gives better resolution compared to the autocorrelation method when the frame size is small. Roots with a magnitude above a certain threshold, say 0.8, and with absolute frequency above a certain threshold, say corresponding to a frequency of 200 Hz, are considered as resonances corresponding to formants [111] and other roots are ignored.

The $F_{tr}(t)$ is significant during transition from consonant to vowel, that is, at the VOP event. This is due to the change in the system characteristics occurring at this point. Hence this can be used as an acoustic cue for detecting the VOP event. Figure 6.1 shows the UVUA velar SCV /ka/ and its VOP event marked using $F_{tr}(t)$ cue. The VOP event is identified as the instant at which transition to the following vowel begins.

6.3.2 Epoch Intervals

The distance between two successive instants of significant excitation is the epoch interval ($E_i(t)$). After extracting the instants, the $E_i(t)$ plot is obtained by plotting the successive time intervals between the epochs. The $E_i(t)$ value varies randomly in the case of unvoiced sounds, but remains nearly constant in the case of voiced sounds. Hence this plot will have uniform contour for the voiced sounds. The beginning of such a contour indicates the VOP event in case of CV units with unvoiced consonants.

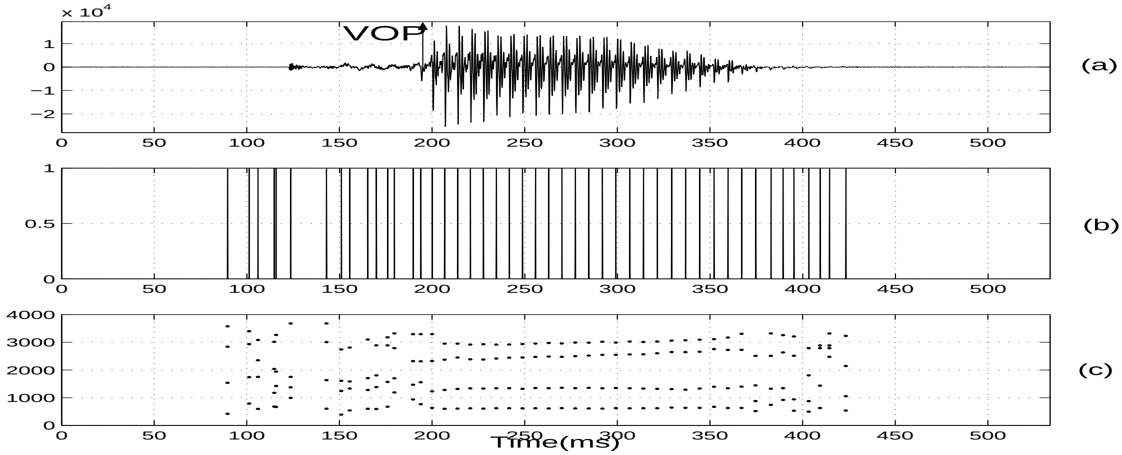

Figure 6.1: UVUA velar SCV /ka/. (a) Waveform, (b) instants of significant excitation and (c) formant contours.

Hence $E_i(t)$ can be used as an acoustic cue for detecting the VOP events. Figure 6.2 shows the VOP event marking for the fricative CV /sha/ using the $E_i(t)$ plot.

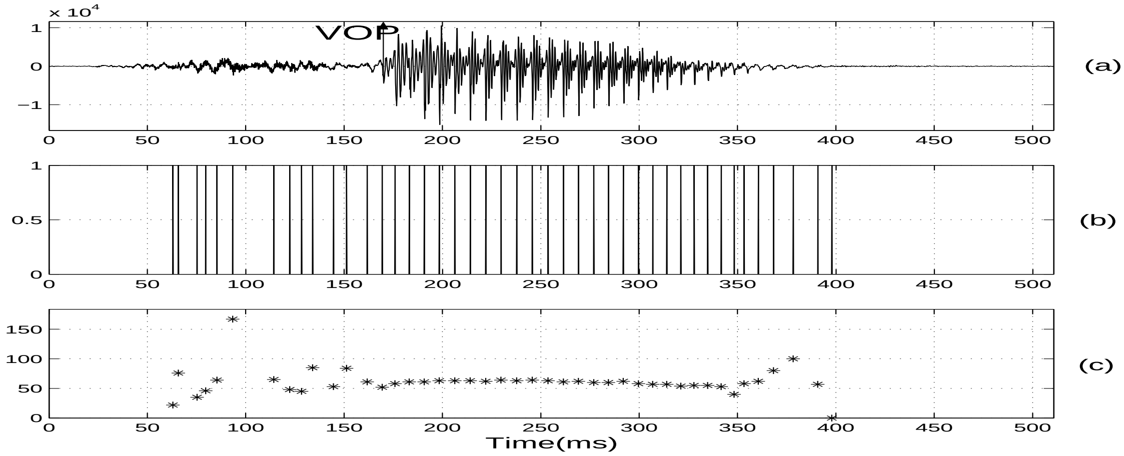

Figure 6.2: Fricative CV /sha/. (a) Waveform, (b) instants of significant excitation and (c) epoch intervals.

6.3.3 Strength of Instants

The strength of instants ($S_i(t)$), which indicates the strength of excitation, mainly depends on the amount of loading of the vocal tract system on the source. The strength of the instants for voiced sounds is generally higher compared to the strength of the random instants present in the unvoiced sound. In particular, the strength of instants for vowels is higher compared to the strength of voiced consonants. Also $S_i(t)$ shows a significant change at the transition from consonant to vowel for most of the CV units. Hence $S_i(t)$ can be used as an acoustic cue for detecting the VOP event. Figure 6.3 shows the nasal CV /mi/, the LP residual, Hilbert envelope of the LP residual, the instants and the strength of the instants. The strengths at the instants are obtained by picking the amplitude of the Hilbert envelope at the locations of the instants. The figure also shows the manually marked VOP event using $S_i(t)$ as the acoustic cue.

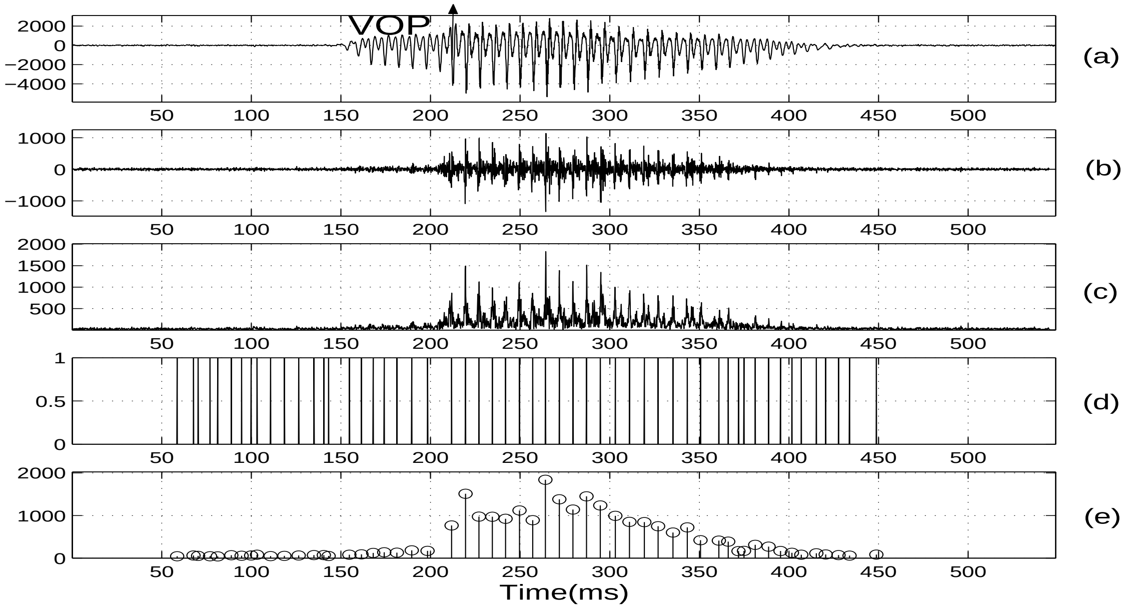

Figure 6.3: Nasal CV /mi/. (a) Waveform, (b) residual, (c) Hilbert envelope of residual, (d) instants of significant excitation and (e) strengths of instants.

6.3.4 Symmetric Itakura Distance

The change in system characteristics during the production of sound units will be manifested as spectral change. The amount of spectral change or distortion can be measured using the Itakura distance ($I_d(t)$) [112]. The spectral change is significant at the VOP event, and hence the $I_d(t)$ can be used as an acoustic cue for detecting the VOP event.

Given two frames, the symmetric Itakura distance is computed using the following relations [113]:

$$d_{12}(t) = \frac{\mathbf{a_2'} R_1 \mathbf{a_2}}{\mathbf{a_1'} R_1 \mathbf{a_1}} \tag{6.2}$$

$$d_{21}(t) = \frac{\mathbf{a_1'} R_2 \mathbf{a_1}}{\mathbf{a_2'} R_2 \mathbf{a_2}} \tag{6.3}$$

$$I_d(t) = \frac{d_{12}(t) + d_{21}(t)}{2} \tag{6.4}$$

where $\mathbf{a_1}$ and $\mathbf{a_2}$ are LPCs of *frame1* and *frame2*, respectively, R_1 and R_2 are the signal autocorrelation matrices corresponding to $\mathbf{a_1}$ and $\mathbf{a_2}$, $d_{12}(t)$ and $d_{21}(t)$ are the asymmetric Itakura distances and $I_d(t)$ is the symmetric Itakura distance.

In this study $I_d(t)$ values are computed at each instant by considering speech frames of 3 ms at the given instant and at the next instant. Figure 6.4 shows the manually marked VOP event for the UVUA velar CV /ka/ using $I_d(t)$ as the acoustic cue.

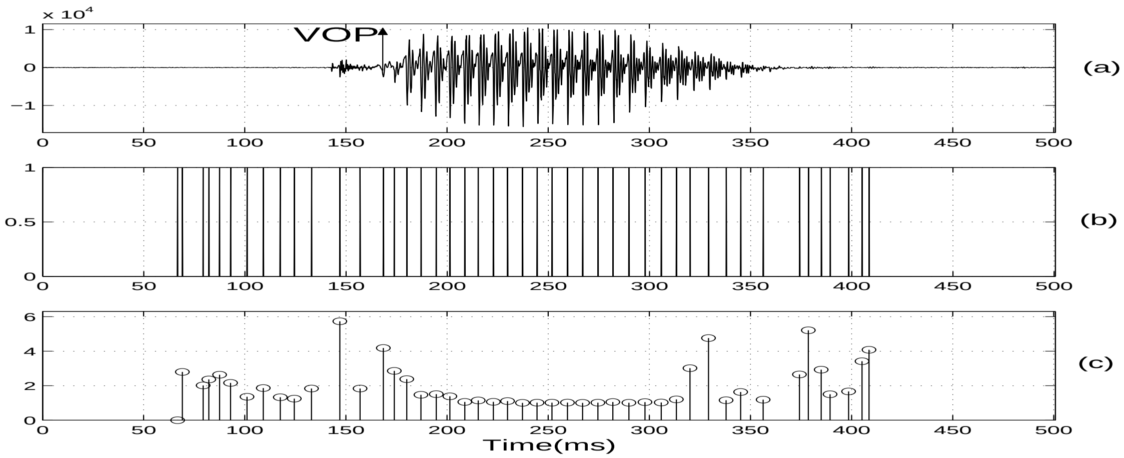

Figure 6.4: UVUA CV /ka/. (a) Waveform, (b) instants of significant excitation and (c) symmetric Itakura distances.

6.3.5 Ratio of Signal Energy to Residual Energy

Energy of the speech signal is generally higher in voiced regions as compared to the energy in unvoiced regions. In the case of residual signal, the energy level may be higher even in the unvoiced region. Therefore the ratio of the signal energy to the residual energy ($S_r(t)$) is low in the consonant region and high in the vowel region, thus indicating a significant change at the VOP event for some of the sound units. For extracting this acoustic cue, short-time energies for 3 ms of the speech signal and residual signal around each instant are computed. The ratio of these two short-time energies is the required acoustic cue. The manually marked VOP event using this cue is shown in Figure 6.5 for the fricative CV /ha/.

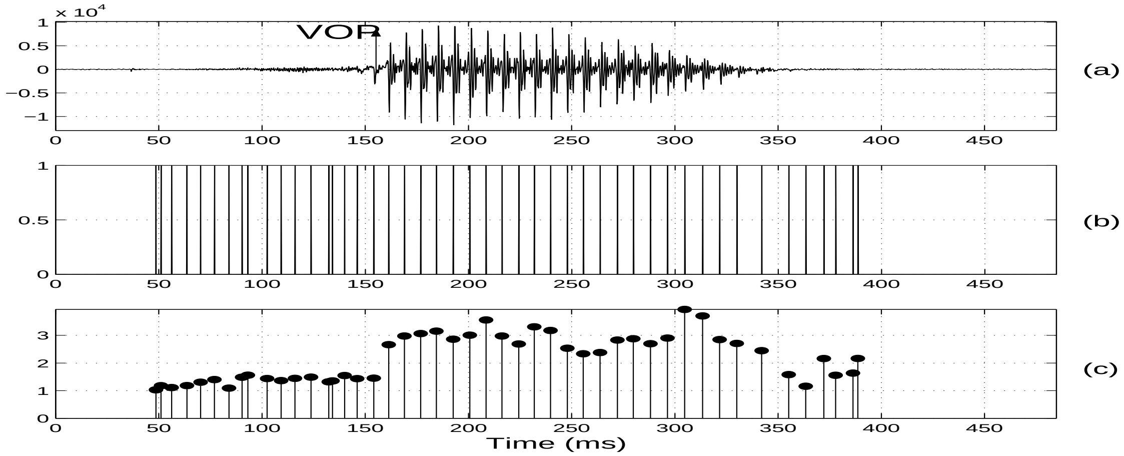

Figure 6.5: Fricative CV /ha/. (a) Waveform, (b) instants of significant excitation and (c) ratio of the signal energy to residual energy.

6.3.6 Summary of the Acoustic cues

With $F_{tr}(t)$ as the cue, the hypothesis is that the VOP is marked by the beginning of formant contours. The categories of CV units for which VOP events can be marked using this cue are UVUA, UVA, VUA, nasal, fricative, affricate and semivowel CV units. This acoustic cue may not be suitable for the case of VA CV units, as the formant transition begins in the consonant region itself. The beginning of uniform contour in the $E_i(t)$ is useful for marking VOP event. This cue is suitable for UVUA, UVA, VUA, fricative and affricate CV units. This cue may not be suitable for CV units with high voiced consonants like VA, nasal and semivowel CV units, as the uniform contour begins in the consonant region. The amount of change in the $S_i(t)$ plot occurring at the VOP event is significant for UVUA, UVA, VUA, nasal, fricative and affricate CV units. The amount of change may not be significant in some cases of VA and semivowel CV units and hence marking VOP events is difficult. In $I_d(t)$, the

VOP event is associated with the instant having large spectral change at the beginning of the vowel. Hence the $I_d(t)$ cue is suitable for UVUA, UVA, VUA, nasal, fricative, and affricate CV units. The spectral change may not be large at the VOP event in some cases of VA and semivowel CV units. The $S_r(t)$ cue shows significant change at the VOP event in the case of CV units having unvoiced consonants. The possible categories for which this cue is suitable are UVUA, UVA, fricative and affricate CV units. The cue may not show significant change at the VOP event in some cases of CV units with high voiced consonants.

The above description is tabulated in Table 6.2 for quick reference. The conclusions that can be made from the above discussion are: The change occurring in an acoustic cue at the VOP event depends on the type of CV unit. Any one acoustic cue is not directly suitable for locating the VOP event for all the 145 CV units. All these acoustic cues may be used together to mark the VOP event in a given CV unit.

Table 6.2: Summary of the acoustic cues for manual marking of VOP event.

Sl. No.	Acoustic cue	Hypothesis for marking the VOP event	Category of CV units for which this cue is suitable
1	$F_{tr}(t)$	Beginning of formant transition	UVUA, VUA, VUA, nasal, fricative, affricate, semivowel
2	$E_i(t)$	Beginning of uniform contour	UVUA, UVA, VUA, fricative, affricate,
3	$S_i(t)$	Significant change in $S_i(t)$ contour	UVUA, UVA, VUA, nasal, fricative, affricate
4	$I_d(t)$	Instant with large distortion	UVUA, UVA, VUA, nasal, fricative, affricate
5	$S_r(t)$	Significant change in $S_r(t)$ contour	UVUA, UVA, VUA, affricate, fricative

6.4 Preparation of Reference Database

To prepare the reference database, speech data was collected at a sampling frequency of 8 kHz for all the 145 CV units in a laboratory environment from three male speakers.

For each CV unit, 12 utterances were collected per speaker. Thus, there are 1740 utterances per speaker and a total of 5220 utterances in the database. For each CV unit, three utterances per speaker (a total of 1305 utterances) were randomly chosen for evaluating the consistency in marking the VOP events using the proposed acoustic cues. The consistency of marking the VOP events using multiple cues is evaluated on the selected 1305 utterances, with the help of 11 subjects (including the author). The subjects were instructed to mark the instant at which evidences due to two or more acoustic cues coincide, as the VOP event. The results of the study is tabulated in Table 6.3. The entires in the table are obtained by considering the deviation of each subject's marking with the author's markings. The relatively poor performance in the case of VA, nasal and semivowel CV units indicates the ambiguity in locating the VOP events in such CV units.

Table 6.3: Results of marking the VOP events using all the acoustic cues by 11 speakers, computed by considering the deviation of each subject's markings with the author's markings, given in percentage. In the table, abbreviations NAS, FRI, AFF, SVOW and AVG refer to nasals, fricatives, affricates, semivowels and average efficiency, respectively.

Sl.No.	DEV	CV category (%)								AVG
	(ms)	UVUA	UVA	VUA	VA	NAS	FRI	AFF	SVOW	(%)
1	10	96.7	92.6	92.0	83.2	82.9	94.1	81.9	72.5	86.9
2	20	99.3	97.8	98.7	94.1	93.8	96.4	91.2	81.1	94.1
3	30	99.8	99.2	99.6	95.1	95.8	97.8	95.8	84.6	95.9
4	40	99.9	99.8	99.8	96.6	96.2	98.7	98.3	88.0	97.2
5	50	100	99.9	99.9	97.7	98.7	99.1	99.1	91.3	98.2
6	> 50	0	0.1	0.1	2.3	1.3	0.9	0.9	8.7	1.8

This study shows that using the proposed acoustic cues, the subjects are consistent in marking the VOP events in most of the cases. It was found that the accuracy of marking is high for sound units like UVUA, UVA, VUA, fricative and affricate CV units and accuracy is low for other sound units like VA, nasals and semivowels. The high accuracy can be attributed to the fact that the acoustic cues are able to locate the start of the voicing better for the case of CV units with unvoiced consonants. The manual marking of the VOP event is illustrated for the difficult cases like aspirated, nasal, semivowel CV units in Figures 6.6 to 6.9 using the proposed cues.

Finally, using the proposed acoustic cues, markings of the VOP events for all the 5220 utterances was performed by the author, for the reference database.

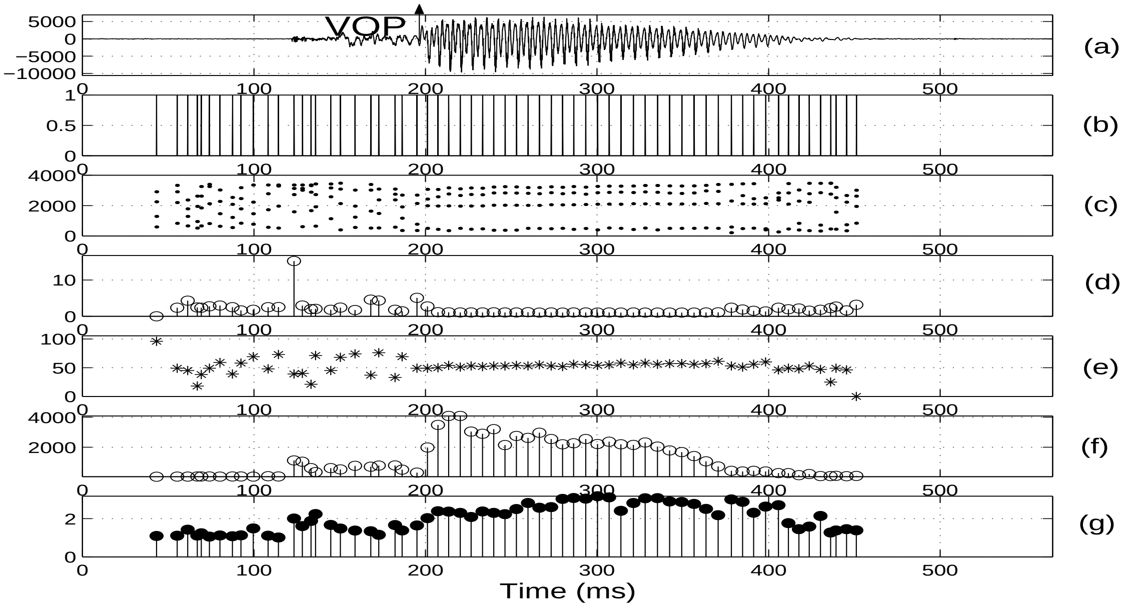

Figure 6.6: UVA CV /khi/. (a) Waveform, (b) instants, (c) formant contours, (d) symmetric Itakura Distance, (e) epoch intervals, (f) strength of instants and (g) ratio of the signal energy to residual energy.

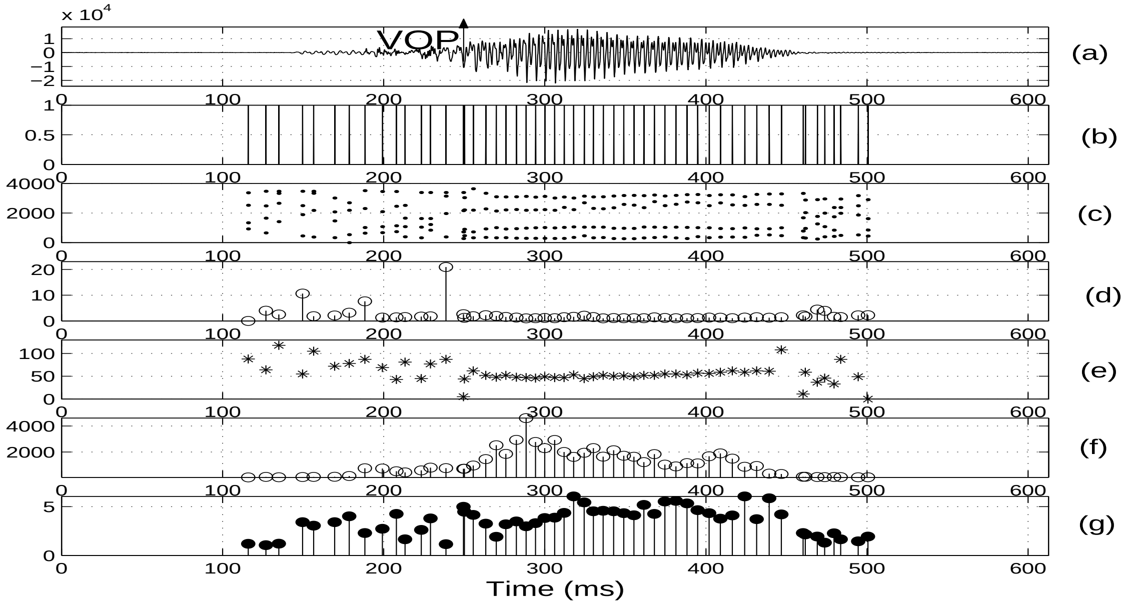

Figure 6.7: VA CV /ghu/. (a) Waveform, (b) instants, (c) formant contours, (d) symmetric Itakura distance, (e) epoch intervals, (f) strength of instants and (g) ratio of the signal energy to residual energy.

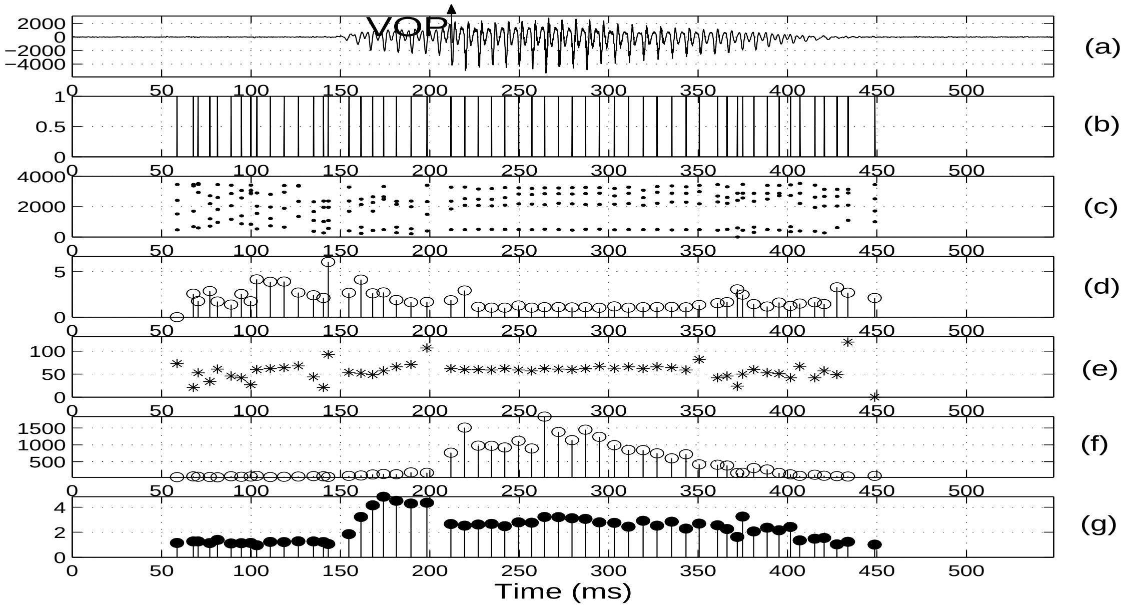

Figure 6.8: Nasal CV /mi/. (a) Waveform, (b) instants, (c) formant contours, (d) symmetric Itakura distance, (e) epoch intervals, (f) strength of instants and (g) ratio of the signal energy to residual energy.

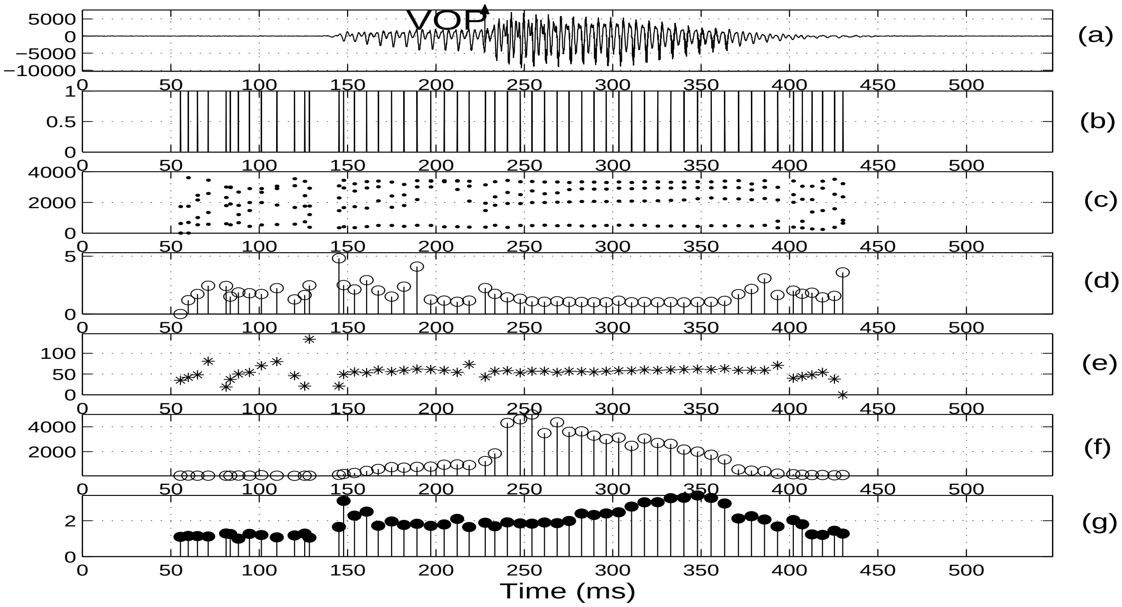

Figure 6.9: Semivowel CV /li/. (a) Waveform, (b) instants, (c) formant contours, (d) symmetric Itakura distance, (e) epoch intervals, (f) strength of instants and (g) ratio of the signal energy to residual energy.

6.5 Automatic Detection of VOP Events

6.5.1 VOP events in Isolated Utterances of CV units

In the previous section, we discussed the preparation of the reference database of CV units with manually marked VOP events using the proposed acoustic cues. While preparing the reference database, significant changes occurring in the acoustic cues are visually observed for a given utterance to mark the VOP event. But to use the knowledge of the VOP event for any application, automatic detection of VOP event is needed. In this section development of an algorithm for automatic detection of the VOP event for isolated utterances of CV units is discussed. From the acoustic description, it is known that the VOP event may also be characterized by changes in the source characteristics. Hence, in this study, only $S_i(t)$ is chosen as the acoustic cue for developing an algorithm for automatic detection of the VOP event.

The speech signal is preemphasized and low pass filtered to 2.5 kHz (5 kHz sampling frequency) to select only the high SNR regions. The LP residual is computed for every frame of 20 ms with a shift of 10 ms using an LP order of 8. The instants of significant excitation are computed from the LP residual. Also, Hilbert envelope of the LP residual is computed. The $S_i(t)$ values are obtained from the Hilbert envelope using the knowledge of the instants.

At the next level, the instant at which there is a significant change in the strength of excitation is to be detected. For this, a Gabor filter (modulated Gaussian pulse) with parameters spatial spread of the Gabor filter $\sigma = 100$, angular frequency of the sinusoidal component $\omega = 0.0114$ and a filter length $n = 800$ is used [114]. The parameters of the Gabor filter are chosen in such way that the negative part of the window is larger than the positive part [1]. This is to ensure a peak only at the VOP event. The Gabor filter is shown in Figure 6.10. The filter parameters are not crucial, except that the general shape as in Figure 6.10 is to be maintained.

$S_i(t)$ is multiplied with the Gabor filter at each sampling instant, and the sum of the product is noted as evidence for the VOP event at that instant. The plot of the evidence is termed as the *VOP Evidence Plot.* In the VOP evidence plot the relative maxima occurs at the instants where the strength of the instants rises sharply and this maximum is detected as the VOP event. The algorithm is given in Table 6.4. The $S_i(t)$ plot and the corresponding VOP evidence plot for UVA velar CV /khi/ are shown in Figure 6.11.

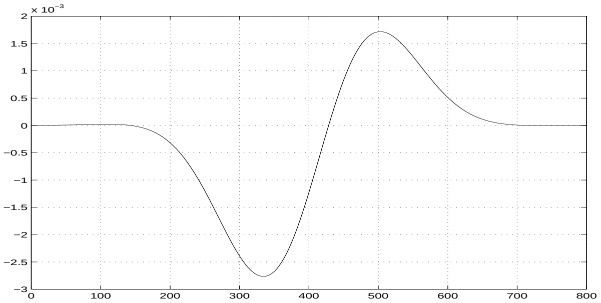

Figure 6.10: Gabor window for $\sigma = 100$, $\omega = 0.0114$ and $n = 800$.

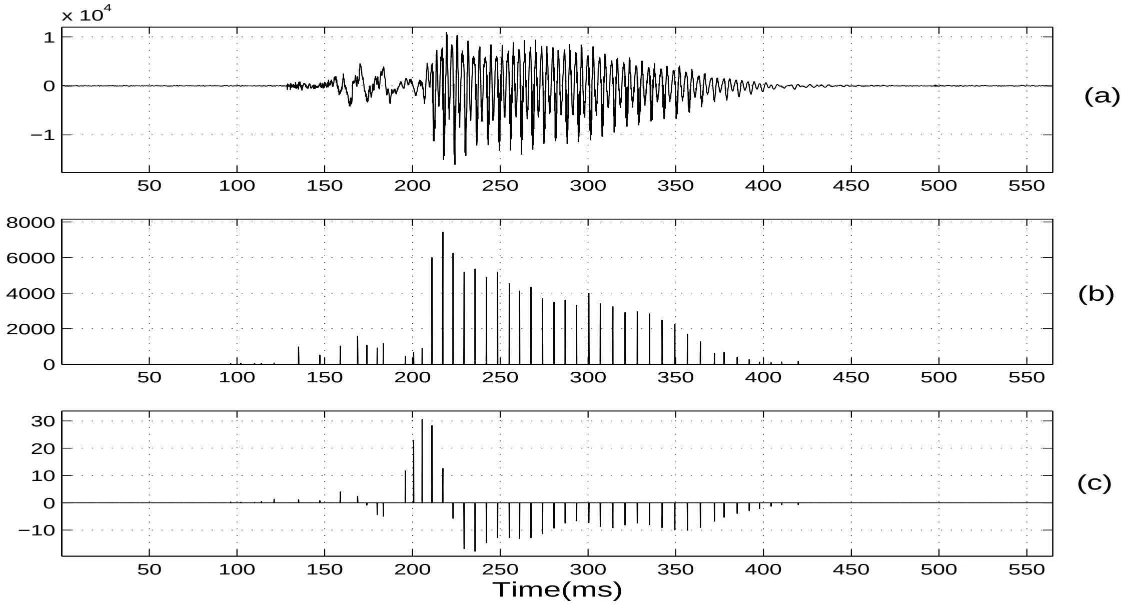

Figure 6.11: UVA CV /khi/ (a) Waveform, (b) strength of instants and (c) VOP evidences.

To evaluate the performance of the proposed algorithm, the reference database is used. The efficiency is found out by computing the deviation between the hypothesized VOP event and the manually marked VOP event. Performance of the algorithm for each category of CV units for different deviations with respect to the manually marked VOP event is given in Table 6.5. The performance is good for the CV categories in which the consonant part is of the unvoiced type like UVUA, UVA and fricatives. The performance seems to degrade for the CV categories the voiced consonants like VUA,

Table 6.4: Algorithm for automatic detection of VOP events in isolated utterances of the CV units.

1. Preemphasize input speech.
2. Select only high SNR portions of input speech (up to 2.5 kHz) by low pass filtering.
3. Compute LP residual with 8^{th} order LP analysis, frame size of 20 ms and shift of 10 ms.
4. Find the instants of significant excitation for every frame of 10 ms with one sample shift.
5. Compute Hilbert envelope of the LP residual.
6. Find the strength of instants.
7. Obtain the VOP event evidence plot from strength of instants using Gabor filter.
8. Find the location of global maximum in the VOP evidence plot, which is hypothesized as the VOP event.

VA, nasals, affricates and semivowels. The performance may be improved using the evidence from the other acoustic cues.

For comparison, algorithms based on system features like energy derivative based method and neural network based method are briefly discussed here [33, 83]. In the case of energy derivative based method the hypothesis is that the VOP event is the point at which there is a significant increase in the energy of a CV utterance. This point may be detected by computing the derivative of the short-time energy of the speech signal, and locating the point at which the positive derivative is maximum. The performance of the energy derivative based method for the CV units in the reference database is shown in Table 6.6. The hypothesis for the neural network based method is that the characteristics of the acoustic cues (signal energy, LP residual energy and spectral flatness) are significantly different in the regions immediately before and after the VOP event. A multilayer perceptron network is trained to detect the VOP event by using the trends in these parameters at the VOP event. The performance of the neural network based method for the same reference database is given in Table 6.6.

The performance of the proposed algorithm is comparable (even slightly better) to the energy derivative and neural network methods [33, 83]. For a deviation of ± 30 msec, the proposed algorithm determines 88 % of the total 5220 VOP events correctly, whereas the energy derivative method detects only 75.0 % and neural network method detects 86.7 % of the VOP events.

Table 6.5: Performance of the proposed algorithm based on excitation source information for the detection of VOP event in isolated utterances of CV units. In the table abbreviations DEV, NAS, FRI, AFF, SVOW and AVG refer to deviation, nasals, fricatives, affricates, semivowels and average efficiency, respectively.

Sl.No.	DEV	CV category (%)								AVG
	(ms)	UVUA	UVA	VUA	VA	NAS	FRI	AFF	SVOW	(%)
1	10	91.1	75.8	69.8	41.5	82.5	84.4	51.8	59.1	69.5
2	20	98.3	90.0	82.7	56.2	89.8	94.2	73.6	70.1	81.8
3	30	99.3	95.2	92.0	66.8	90.6	95.2	87.6	76.9	88.0
4	40	100	96.6	95.2	75.7	92.0	97.3	92.7	82.6	91.5
5	50	100	97.0	96.8	82.5	92.6	98.3	95.5	90	94.1
6	> 50	0.0	3.0	3.2	17.5	7.6	1.7	4.5	10	5.9

Table 6.6: Average performance of energy derivative and neural network based methods for the detection of the VOP events in isolated utterances of 145 CV classes.

Sl.No.	Deviation (ms)	Energy Derivative based Method (%)	Neural Network based Method (%)
1	10	51.5	68.5
2	20	65.4	78.7
3	30	75.0	86.7

6.5.2 VOP Events in Continuous Speech

To hypothesize the VOP events in continuous speech, the algorithm proposed for isolated utterances of CV units may be used with slight modification. For continuous speech finding $S_i(t)$ is computation intensive, because computation of instants of significant excitation from the residual is obtained by the group delay processing for every sample shift. Alternatively, since Hilbert envelope of the LP residual represents approximately the strength of instants, this can be used instead of $S_i(t)$ for hypothesizing the VOP event. The VOP event evidence is obtained from the Hilbert envelope of the LP residual by multiplying it with the Gabor filter ($\sigma = 100$, $\omega = 0.0114$, and $n = 800$), and taking the sum of the product for every sample shift. In the VOP evidence plot the peaks are located using a peak picking algorithm. Spurious peaks are eliminated using characteristics of the shape of the VOP evidence plot, namely, between two true VOP events, there exists a negative region of sufficient strength due to the vowel region. The algorithm for the detection of VOP events in continuous speech is given in Table 6.7.

Table 6.7: Algorithm for detection of VOP events in continuous speech using excitation source features.

1. Preemphasize input speech.
2. Select only high SNR portions of input speech (upto 2.5 kHz) by low pass filtering.
3. Compute LP residual with 8^{th} order LP analysis, with a frame size of 20 ms and shift of 10 ms.
4. Compute Hilbert envelope of the LP residual
5. Obtain the VOP event evidence plot from Hilbert envelope for every sample shift using Gabor filter.
6. Identify the peaks in the VOP event evidence plot using peak picking algorithm.
7. For each peak, if there is no negative region with reference to next peak, then eliminate such a peak as it is spurious.
8. Hypothesize remaining peaks as the VOP events.

The above procedure is illustrated for a Hindi sentence /antarAshtriyA bassevA pichale mahIne shuru huyithi/. In this sentence there are 16 VOP events, as marked in Figure 6.12(a) using the proposed acoustic cues. The Hilbert envelope and the VOP evidence plots are shown in Figures 6.12(b) and (c), respectively. The output of the peak picking algorithm is given in Figure 6.12(d), and the hypothesized VOP events after eliminating the spurious ones are shown in Figure 6.12(e). Comparing the manually marked VOP events and the hypothesized VOP events, it can be seen that the algorithm has hypothesized 14 VOP events correctly within a deviation of ± 20 ms, 2 VOP events are not detected and 1 is a spurious VOP event.

To study the effectiveness of the proposed algorithm, 25 sentences from five Hindi news bulletins from five different speakers (2 male and 3 female) are chosen. For each of these sentences, the VOP events are manually marked using the proposed acoustic cues. There are 236 VOP events in the selected data. Out of the total 236 VOP events, 209 (88.5 %) are detected within a resolution of ± 20 ms, 22 (9.32 %) are missing and 29 (12.3 %) are wrongly hypothesized.

To compare the performance of the proposed algorithm with that of the algorithm based on system features, frame energies of band-pass speech (500-2500 Hz) for blocks of 10 ms with every sample shift are computed. The band energy in the range 500-2500 Hz typically represents the energy of the first two formants and it is assumed to represent the system features. The VOP event evidence plot is obtained from the energies computed using the Gabor filter for every sample shift. The VOP events are hypothesized from the VOP event evidence plot as explained earlier. The algorithm is given in Table 6.8. The algorithm is illustrated for the utterance /antarAshtriyA bassevA pichale mahIne shuru huyithi/. From Figures 6.12 and 6.13, we can see that both the methods hypothesize the VOP events approximately at the same places.

Table 6.8: Algorithm for automatic detection of VOP events in continuous speech using vocal tract system features.

1. Preemphasize input speech.
2. Select only high SNR portions (from 500 - 2500 Hz) by band pass filtering.
3. Compute the frame energies for blocks of 10 ms for every sample shift.
4. Obtain the VOP event evidence plot from the energies for every sample shift using Gabor filter.
5. Identify the peaks in the VOP event evidence plot using peak picking algorithm.
6. For each peak, if there is no negative region with reference to next peak, then eliminate such a peak.
7. Hypothesize remaining peaks as the VOP events.

The performance of both the algorithms, that is, algorithm based on source as well as spectral features is tabulated in Table 6.9. As given in the table, the performance of both the algorithms are nearly the same.

Table 6.9: Performance of the proposed algorithm based on source features and the algorithm based on system features for detection of the VOP events in continuous speech. In the table the abbreviations HYPO, SPU and MISS represent hypothesized, spurious and missing, respectively.

Method	Total VOPs	HYPO VOPs	SPU VOPs	MISS VOPs	VOPs in ± 10 ms	VOPs in ± 20 ms	VOPs in ± 30 ms	VOPs in ± 40 ms
Source feature	236	243	29 (12.3%)	22 (9.3%)	182 (77.1%)	209 (88.5%)	213 (90.2%)	214 (90.6%)
System feature	236	237	24 (10.1%)	23 (9.7%)	186 (78.8%)	210 (89.0%)	211 (89.4%)	213 (90.2%)

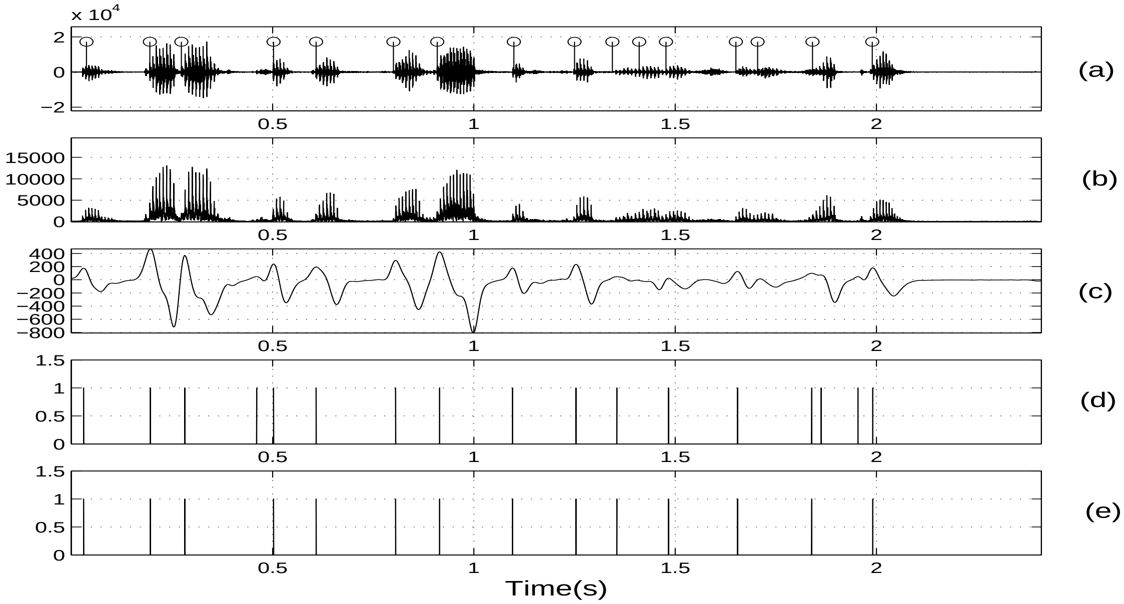

Figure 6.12: Hindi sentence /antarAshtriyA bassevA pichale mahIne shuru huyithi/. (a) Waveform with manual marked VOP events, (b) Hilbert envelope of the LP residual, (c) VOP evidences, (d) output of peak picking algorithm and (e) hypothesized VOP events after eliminating some spurious peaks.

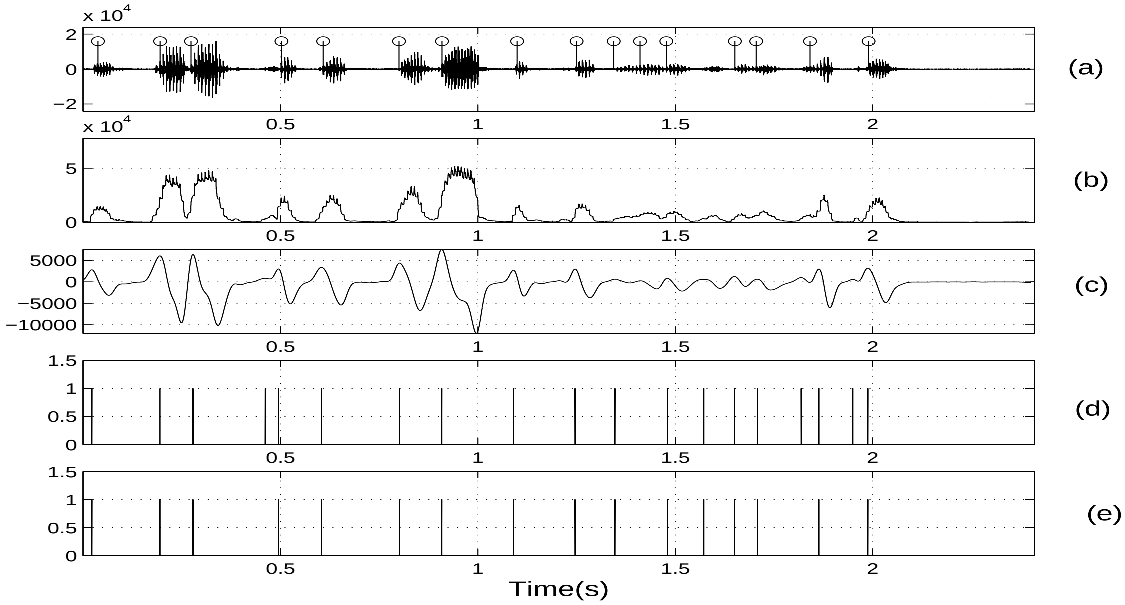

Figure 6.13: Hindi sentence /antarAshtriyA bassevA pichale mahIne shuru huyithi/. (a) Waveform with manual marked VOP events, (b) frame energies, (c) VOP evidences, (d) output of peak picking algorithm and (e) hypothesized VOP events using the energy of speech signal.

6.6 Summary

The acoustic description of the VOP event in different categories of CV units was given. With the knowledge of this acoustic description, a set of acoustic cues based on the GC events for detecting the VOP events was proposed. A reference database of CV units with manually marked VOP events was prepared using the proposed acoustic cues. An algorithm for automatic detection of the VOP event in isolated utterances of CV units was proposed using the strength of instants as an acoustic cue. An algorithm for the detection of VOP events in continuous speech is proposed using Hilbert envelope of the LP residual. Summary of the various issues discussed in this chapter is given in Table 6.10.

In the next chapter, usefulness of VOP events is demonstrated in the detection of end-points of a speech utterance.

Table 6.10: Summary of the issues with respect to detection of the VOP events.

VOP Event for Speech Analysis

- Issues in the detection of VOP events
  - Difficult to detect the VOP events in VA, nasal and semivowel CV units.
  - VOP event is an instant property, but the existing methods employ block processing for the detection of VOP events.
  - Existing methods mainly use vocal tract system features for the detection of VOP events.
- Acoustic description of VOP events
  - Excitation source of speech production also contains significant information about the VOP event.
- Acoustic cues for the detection of VOP events
  - Formant transition, Itakura distance and the ratio of signal energy to residual energy indicate the changes in vocal tract characteristics.
  - Epoch intervals and strength of instants indicate the changes in excitation source characteristics.
- Preparation of the reference database
  - Instant where two or more acoustic cues show a significant change is marked as the VOP event.
- Automatic detection of VOP events
  - Strength of instants is used as a cue for developing the algorithm.
  - VOP evidence plot is computed from the strength of instants using the Gabor filter
  - Peaks in the VOP evidence plot (after eliminating spurious) are hypothesized as the VOP events.

Chapter 7

APPLICATION OF VOP EVENTS

The issues in the detection of VOP events were discussed in the previous chapter. Acoustic cues for detection of VOP events were developed using the knowledge of GC events. Methods for detection of VOP events using the excitation source information derived from the GC events were proposed. This chapter discusses the significance of VOP events in speech analysis.

7.1 Introduction

A method for detection of end-points of a speech utterance using the knowledge of VOP events is proposed. An algorithm for the detection of VOP event for text-dependent continuous speech is discussed. The VOP event helps in overcoming the difficulties present in coming up with multiple thresholds followed in most of the existing end-points detection algorithms. The VOP event of the first vowel is used as an anchor point for further analysis to detect the beginning of the speech utterance. Similarly, the VOP event of the last vowel is used as an anchor point for detecting the end point. The performance of the proposed end-points detection algorithm is compared with that of the existing energy-based approach by conducting text-dependent speaker verification experiments. The speaker verification system using the knowledge of VOP events for the detection of end-points shows a significant improvement in the performance.

The chapter is organized as follows: Section 7.2 gives a description of the speaker verification system using energy-based end-points detection algorithm. The VOP-based end-points detection method is explained in Section 7.3. Section 7.4 concludes with a summary of the issues discussed in this chapter.

7.2 Speaker Verification using Energy-based End-points Detection

7.2.1 Speech Database

The speech database for this study was collected from 30 cooperative speakers (21 male and 9 female) over microphone as well as telephone channels. A typical telephone channel has a passband of 300-3300 Hz. In addition to bandwidth limitation, telephone channels may introduce noise and distortion to the spectral characteristics of speech signals. The speech data is collected for 10 sentences of Hindi (an Indian language). The number of words in these sentences vary from 5 to 7, and the durations of the sentences from 2 to 3 seconds. Each of the 10 sentences was uttered 18 times by each speaker. The data was collected in a laboratory environment in different sessions for microphone and telephone cases. However, one set of all the 18 utterances for each sentence by a speaker was collected in a single session. Thus, with this data, it is not possible to obtain inter-session variability for the same channel. However, the effect of inter-session variation can be studied along with the effect of inter-channel variation by matching the microphone data with the reference templates for the telephone data or vice versa. The speech data was sampled at 8 kHz and stored as 8 bit samples.

7.2.2 Speaker Verification System

The speaker verification system consists of four stages: Preprocessing, feature extraction, pattern classification and decision making. Preprocessing involves mainly the detection of the end-points of a speech utterance. Correct detection of the end-points increases the accuracy of aligning the reference and test utterances [115] [92]. An algorithm based on the energy of the speech signal is used for detection of the end-points in the baseline system [90]. The speech signal is blocked into frames of 20 ms with a frame shift of 5 ms. The energy of each frame is determined and mean and standard deviation of these energies are computed. Ten percent of the sum of the mean and standard deviation is taken as the threshold for a frame to be considered as a speech frame. Starting from the first frame, the frame at which a block of at least 20 consecutive speech frames begins is marked as the begin point. Similarly starting from the last frame and moving backwards, the frame at which a block of at least 20 consecutive speech frames begins is marked as the end point.

Spectral information is extracted for each differenced and Hamming windowed frame of the speech signal using LP analysis [96]. The spectral information is represented using Weighted Linear Prediction Cepstral Coefficients (WLPCC) and the corresponding delta cepstral coefficients [96] [89]. A 10^{th} order LP analysis is used to derive the 20 weighted linear prediction cepstral coefficients for each frame of 20 ms. The delta cepstral coefficients are obtained by deriving the average slope of the contour for each of the WLPCC from 7 successive frames [89]. Only the first 5 delta cepstral coefficients are considered, as it was experimentally found that the other delta coefficients did not contribute much to the performance of the speaker verification system [34]. Thus the feature vector for each frame consists of 25 components (20 WLPCCs and 5 delta cepstral coefficients). We use this 25 dimension vector to represent segmental features of speech.

Both the reference and test utterances are represented by a sequence of 25 dimension feature vectors. The reference and test utterances are matched using Dynamic Time Warping (DTW) algorithm [116, 117]. The matching score is the minimum distance, which is obtained along the optimal warping path of the DTW algorithm.

For each speaker, out of the 18 utterances for each sentence, 3 utterances are used for creating reference templates. The remaining 15 utterances are used for conducting the genuine speaker tests. Thus there are 45 genuine trial scores (15 × 3) for each speaker. Hence the total number of genuine speaker tests per sentence for 30 speakers is 1350 (30 × 45). Imposter tests for each speaker are conducted by using the utterances of the remaining 29 speakers in the database. For each speaker, three utterances of the same sentence are taken for testing. Thus, there are 261 impostor trial scores (87 × 3) for each speaker for each sentence. Hence, the total number of imposter speaker tests per sentence for 30 speakers is 2610 (30 × 87). Since there is data for ten sentences, the total number of genuine speaker trials are 13500 (1350×10) and the total number of impostor trials are 26100 (2610×10).

The performance of the speaker verification system is evaluated as follows: For each speaker for each sentence, the genuine and the impostor scores are normalized to the range from -1 to 1. The threshold is linearly varied from the -1 to 1 and at each threshold the fraction of the False Acceptance (FA) and the fraction of the False Rejection (FR) are noted. The point at which the FA and FR curves as a function of the threshold meet, is the EER for that speaker. The average value of EER for all the speakers and for all the sentences is given in Table 7.1. Analysis of the results shows that most of the failure cases are due to errors in the detection of the end-points. The baseline system which uses an energy-based approach for the end-points detection fails mostly in the cases where the speech data is noisy, which happens especially in the case of telephone speech. To minimize the errors in the end-points detection a method based on the knowledge of VOP events is described in the next section.

Table 7.1: Performance of the text-dependent speaker verification system which uses energy-based end-points detection.

End-points detection	Reference Patterns	Test Patterns	Equal Error Rate
Energy	Microphone	Microphone	5.80
	Telephone	Telephone	6.82
	Telephone	Microphone	11.41

7.3 VOP-based End-points Detection

Detection of the end-points can be improved by using the knowledge of VOP events [33,83,118]. The VOP events are obtained using Hilbert envelope of LP residual [119]. A segment of continuous speech, its LP residual and Hilbert envelope of LP residual are shown in Figure 7.1. Also, the places of significant change in the strength of excitation are probable candidates for the VOP events.

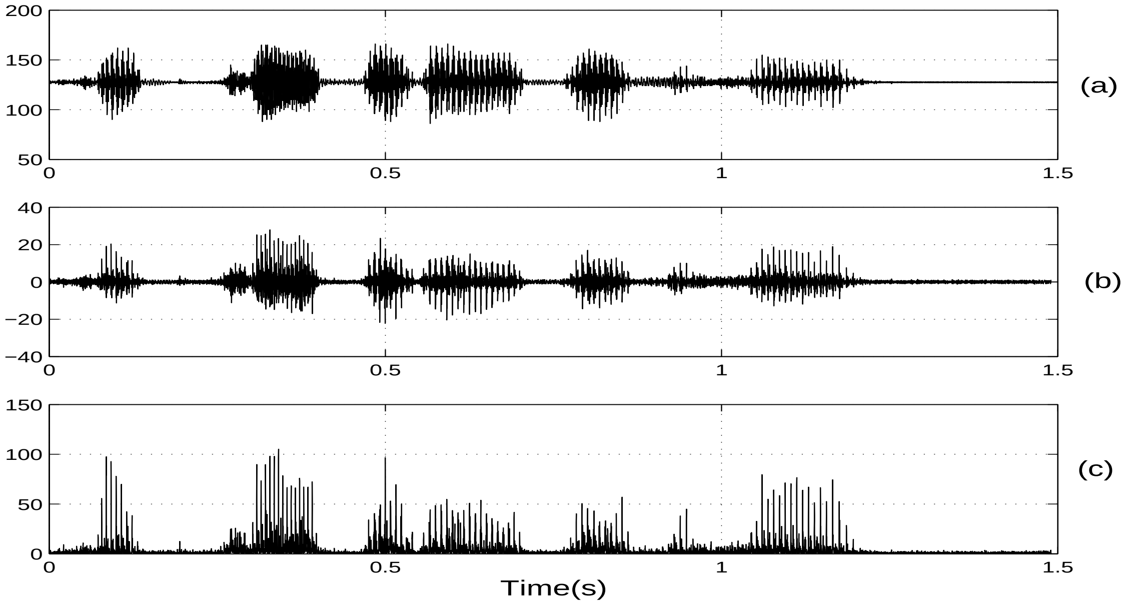

Figure 7.1: (a) Speech segment, (b) LP residual and (b) Hilbert envelope of the LP residual.

Table 7.2 gives the algorithm, a modified version of the algorithm given in the previous chapter, for detecting the VOP events. A speech utterance from the database and its detected VOP events are shown in Figure 7.2. As shown in the figure, the speech utterance has 8 VOP events. The proposed algorithm hypothesized all the 8 VOP events correctly. Additionally, one spurious VOP is also hypothesized. However the spurious VOP event is in between the first and last VOP events and will not degrade the performance of end-points detection. The heuristics used in the algorithm ensures that no spurious VOPs are hypothesized either at the begin or at the end. This is because the heuristics uses speech knowledge to eliminate spurious ones. It was observed that the spurious ones that may occur either at the beginning or at the end are mainly due to transients like clicks.

To evaluate the performance of the VOP detection algorithm, VOP events for 60 randomly chosen utterances from the database are manually marked using the knowledge of Hilbert envelope of the LP residual. There are totally 480 VOP events. For the same 60 utterances, the VOP events are detected automatically using the proposed algorithm. It was found that 95.6% of the VOP events are correctly detected within a deviation of ±30 ms [119]. In the chosen 60 utterances, among the total 480 VOP events, 459 are correctly hypothesized, 21 are missing and 26 are spurious.

Table 7.2: Algorithm for automatic detection of VOP events in text-dependent type continuous speech

1. Preemphasize the input speech.
2. Low pass (cut off freq 2.5 kHz) filter the speech signal.
3. Compute the LP residual using 10^{th} order LP analysis, using a frame size of 20 ms and a frame shift of 5 ms.
4. Compute Hilbert envelope of the LP residual.
5. Obtain the *VOP evidence plot* from the Hilbert envelope by passing the signal through a filter given by
   $g(n) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{n^2}{2\sigma^2}} cos(\omega n)$
   where σ is the spatial spread and ω is the modulating frequency.
   ($\sigma = 100$, $\omega = 0.0114$ and analysis window size = 800)
6. Find the maximum in the *VOP evidence plot* and identify the peaks in the *VOP evidence plot* greater than 5% of the maximum as the candidates for the VOP events.
7. Eliminate the spurious peaks by checking for the presence of the vowel region between two peaks, which is indicated by a negative region in the *VOP evidence plot*.
8. In continuous speech two vowels cannot occur in less than 50 ms duration. Hence eliminate peaks which are at a distance less than 50 ms with respect to their next peak.
9. Also in a text-dependent continuous speech case, the VOP events cannot be at a distance more than 500 ms and eliminate such peaks on either side of the neighborhood peaks.
10. Hypothesize the remaining peaks as the VOP events.

Among the missing VOP events, very few of them correspond to either the first or the last VOP event of the utterance. These missing VOP events are the cases when the strengths of the first vowel and the last vowel are comparable to that of the noise level. These failures can be attributed to the VOP event detection algorithm, which presently uses only the strength of the LP residual.

The first and the last VOP events are used to locate the end-points. The point 300 ms before the first VOP event is marked as the begin point of the speech utterance. Similarly, the point 300 ms after the last VOP event is marked as the end point of the utterance. Table 7.3 shows the performance of the text-dependent speaker verification system using the VOP-based end-points detection method. It can be seen that the performance of the system has improved significantly when the knowledge of the VOP events is used for the detection of the end-points.

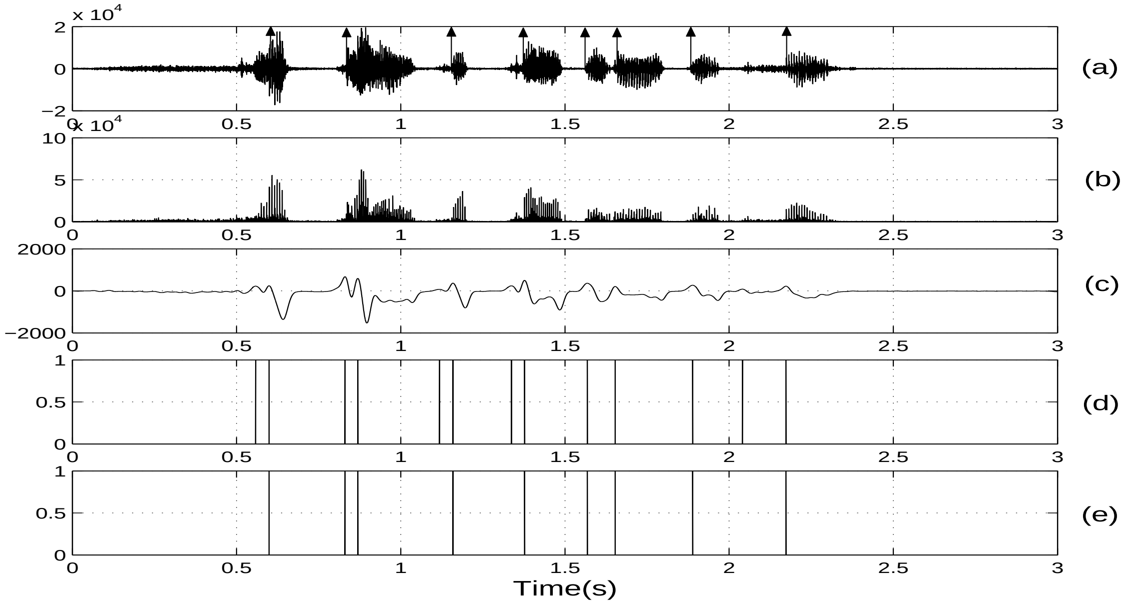

Figure 7.2: Steps in the detection of VOP events. (a) Speech signal with manually marked VOP events, (b) Hilbert envelope of LP residual, (c) *VOP evidence plot*, (d) peaks as candidates for VOP events and (e) hypothesized VOP events.

Table 7.3: Performance of the text-dependent speaker verification system which uses VOP-based end-points detection.

End-points detection	Reference Patterns	Test Patterns	Equal Error Rate
VOP	Microphone	Microphone	2.54
	Telephone	Telephone	2.77
	Telephone	Microphone	3.73

7.4 Summary

A robust method for detection of the end-points based on the knowledge of VOP events, is proposed in this chapter, which is crucial for tasks like text-dependent speaker verification system based on template matching. A summary of the various issues discussed in this chapter is given in Table 7.4.

Table 7.4: Summary of the discussion related to the application of VOP events.

Application of the VOP Events

- Issues in End-points Detection
  - Existing methods use energy as the feature and multiple thresholds are employed in coming up with a decision on the end-points. The performance of these methods will be poor for degraded speech.
- VOP-based End-points Detection
  - Knowledge of VOP events is used for detecting the end-points.
  - As the events are robust to degradations, the proposed method detects the end-points accurately even in degraded speech.

Chapter 8

SUMMARY AND CONCLUSIONS

8.1 Summary of the Work

In this thesis, an event-based approach for analysis and processing of speech is proposed. The proposed approach is based on the nature of speech production. The events are used as anchor points for processing speech and hence the proposed approach is termed as *Event-based Analysis of Speech*. To discuss the various issues involved in the proposed approach, two events, based on GC and VOP events are chosen for the study. Two methods, namely, group delay approach and Hilbert envelope of the LP residual, are used for the detection of GC events. Hilbert envelope of LP residual gives approximate information about GC events. The group delay based approach provides more accurate results. Throughout this work, whenever approximate information about the GC events is sufficient, Hilbert envelope of LP residual was used and for accurate analysis group delay based approach was used.

One accurate method for detection of the pitch is to locate the GC events and obtain their successive time differences. In this work, Hilbert envelope of the LP residual is used as a representation of the GC events. Since detection of peaks at the GC events in the Hilbert envelope of the LP residual is a difficult task, especially in adverse conditions, autocorrelation analysis is performed to find the values of pitch. In case of speech collected over a severely degraded channel the samples in the Hilbert envelope of the LP residual corresponding to the speech regions, in particular, regions around the GC events show high correlation, compared to the samples in the nonspeech regions. This property is exploited to develop a method for enhancement of speech in single channel case.

A method based on GC event information is proposed for time-delay estimation. The cross-correlation of segments of Hilbert envelopes of the LP residuals from two microphone signals show a prominent peak in the cross-correlation sequence. In multichannel case since there will be more than one signal, coherently adding the signals after compensating for time-delay will provide enhancement mainly against the background noise. To provide additional enhancement against reverberation, Hilbert envelopes of the LP residuals from the multiple microphones are coherently-added. In the coherently-added Hilbert envelope, the samples around the GC events show high deviation compared to other places. This property of the coherently-added Hilbert envelope is used for developing a method for enhancement of speech in multichannel case. One important point to be noted in the multispeaker environment is that, as the speakers are spatially distributed, unique time-delay will be associated with respect to each speaker. It is observed that when Hilbert envelopes of the LP residuals from the microphones are coherently added with respect to a particular delay (say, *delay-1*), then, some excitation instants are enhanced in the coherently-added Hilbert envelope. Similarly, with respect to the other delay (say, *delay-2*), some other excitation instants in the coherently-added Hilbert envelope are found to be enhanced. This behavior of the samples in the coherently-added Hilbert envelopes is used for coming up with a method for enhancement of speech in multispeaker environment.

In the studies related to VOP events, one of the important observations made is that their is always a change associated with the characteristics of the excitation source at the VOP event. Hence, methods for detection of VOP events using the excitation source information are proposed. In this work a method for detection of end-point based on the knowledge of VOP events is also proposed. After detecting the VOP events, the first and last VOP events are used as anchor points for coming up with the decision on the end-points.

8.2 Major Contributions of the Work

The important contribution of the research work reported in this thesis is an event-based approach for analysis and processing of speech. Events are used as anchor points and careful analysis of the characteristics of the signal around the events is carried out to develop methods for processing speech to achieve the desired objective in different applications. The major contribution of the thesis is in developing methods for the following:

- Detection of GC events
- Detection of VOP events using excitation source information
- Extraction of pitch in adverse conditions
- Enhancement of speech in single channel case
- Estimation of time-delay using excitation source information
- Speech enhancement in multichannel case
- Enhancement of speech in multispeaker environment
- End-points detection

8.3 Scope for Future Work

- Hilbert envelope of LP residual gives approximate location of GC events. A more accurate and computationally efficient method for the detection of GC events may be developed by first identifying the GC events approximately using Hilbert envelope of the LP residual and performing the group delay analysis on a small region of the LP residual around the approximate locations.
- The proposed method for the detection of VOP events uses only information about the excitation source. The performance of this method may be improved by incorporating additional information from the vocal tract system features.

- The proposed method for enhancement of speech in single channel case enhances the speech only at the gross level. Inside the detected speech regions, there are high SNR components like the GC events. In the detected speech regions, further enhancement can be done both at the excitation source level and the vocal tract system level to improve perceptual quality of the processed signal.

- Time-delay estimation is done using only the knowledge of the excitation source features. Approaches may be developed to combine time-delay values estimated by the proposed method with the time-delay values from the spectral-based GCC method to improve the performance of the combined method.

- Speech enhancement in multichannel case is achieved by enhancing the excitations in the LP residual. It may be possible to improve the performance by a better choice of the system parameters, which may be obtained by selecting the high SNR regions immediately after the GC events for parameter estimation.

- Speech enhancement in multispeaker environment is achieved using the knowledge of time-delays and the coherently-added Hilbert envelopes. We have made no attempt to modify the parameters of the time varying filter. It may be possible to derive the parameters of the filter corresponding to each speaker using the knowledge of the instants of significant excitation of the desired speaker. It is also possible to obtain significant improvement in signal separation from a multispeaker environment if the speech data is collected from a number of spatially distributed microphones.

Appendix–A
SIGNIFICANCE OF EVENTS IN SPEECH PERCEPTION

Studies have been conducted to know the activity taking place in the mammalian auditory nerve fibers when speech sounds are presented to the auditory system [1,6–16]. It has been observed that mammalian auditory nerve fibers produce a maximum firing at the onset of tone bursts of constant sound intensity [1]. This can be observed in the Post-Stimulus Time (PST) histograms shown in Figure A.1, which are taken from [1] for illustration. The histograms have been obtained with a large 21 dB intensity increment applied at several time delays. As shown in the figure, the response of the auditory nerve fiber is maximum at the onset of the tone, which indicates that significant information for perception is present around the onset.

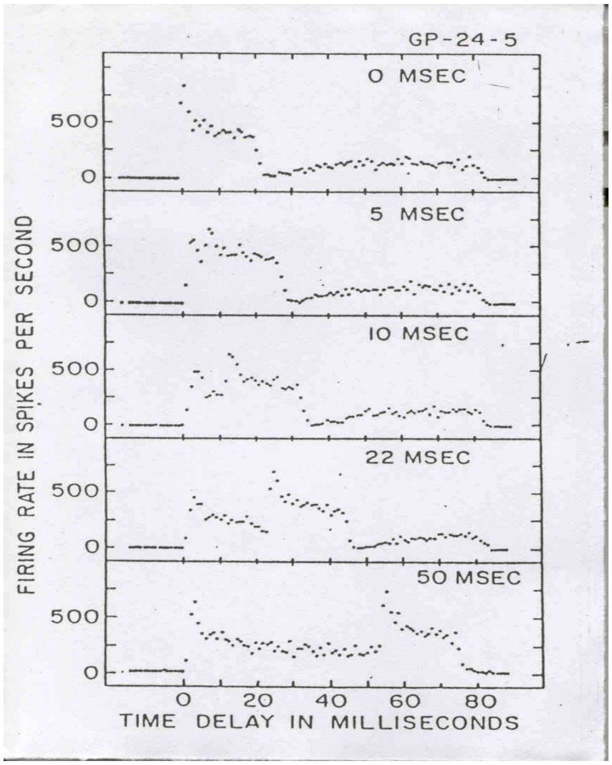

Figure A.1: PST histograms of responses of an auditory-nerve fiber (taken from [1]).

The following are some of the important observations made with respect to the auditory nerve fibers: High spontaneous fibers tended to be more active at the onset of the syllable [13]. The range of auditory nerve fiber is larger at the onset of sound [11].

The spectrum near the onset of the consonant is well preserved in the profiles of average discharge rate versus characteristic frequency [14]. Many fibers track a formant from its onset to its steady-state frequency [13]. The response of fibers to speech-like noise bursts tend to show onsets with appreciable overshoot when the onset is abrupt [12]. A signal undergoes some kind of special processing at an abrupt onset [15,16]. From these observations it can be inferred that the onset of events and some regions around the onset of events are important, which contain discriminatory information for perception.

Analysis of speech sounds using signal processing tools have also been made to find out the acoustic features which contain discriminatory information for further processing [17–21]. These studies show that features extracted from small regions starting from the onset of events, both static and time-varying, contain important information. For instance, static spectrum extracted from the onset of burst event contains information about the place of articulation for stop sounds [20]. Time-varying spectrum extracted from the onset of voicing also contains information about the place of articulation [21].

Several perceptual studies by human subjects have been conducted to identify which part of a given speech sound contains crucial information [22–26]. These studies infer that discriminatory information for perception lies at discrete places and it is concentrated near the onset of events. When regions around the onset of events are deleted from the speech sounds, then it is found that the speech is less intelligible. Thus regions around the events bear the discriminatory information for perception.

Appendix–B
LINEAR PREDICTION ANALYSIS

Linear prediction (LP) analysis is a model based approach for analysis of speech signals [38, 96, 120]. Most of the non-model based methods for analyzing speech start by transforming acoustic data into spectral form by performing a short-time spectrum analysis of the speech wave. Although spectral analysis is a well known technique for studying signals, its application to speech signals suffers from a number of serious limitations arising from the nonstationary as well as the quasiperiodic properties of the speech wave. As a result, methods based on spectral analysis often do not provide a sufficiently accurate description of speech articulation. In contrast, LP analysis is a model based approach in which speech is represented directly in terms of time varying parameters related to the transfer function of the vocal tract and the characteristics of the source function. By modeling speech wave itself, rather than its spectrum, the problems occurring in the frequency domain methods are eliminated. For instance, the traditional Fourier analysis methods require a relatively long speech segment to provide adequate spectral resolution. As a result, rapidly changing speech events cannot be accurately detected. Furthermore, because of the periodic nature of voiced speech, little information about the spectrum between pitch harmonics is available; consequently, the frequency domain techniques do not perform satisfactorily for high-pitched female voices.

The Figure B.1 shows a model of speech production for LP analysis. It consists of a time varying filter $H(z)$ which is excited by either a quasi periodic or a random noise source. The type of excitation determines the nature of speech sound that is, voiced or unvoiced and the parameters of the filter determines the identity of the speech sound.

In LP analysis, the output signal $s(n)$ is assumed to be a linear function of past outputs, and present and past inputs, and hence the name linear prediction. Therefore output $s(n)$ is given by

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + G\sum_{l=0}^{q} b_l u(n-l) \qquad b_0 = 1 \tag{B.1}$$

where a_k, $1 \leq k \leq p$, b_l, $1 \leq l \leq q$ and the gain G are the parameters of the filter.

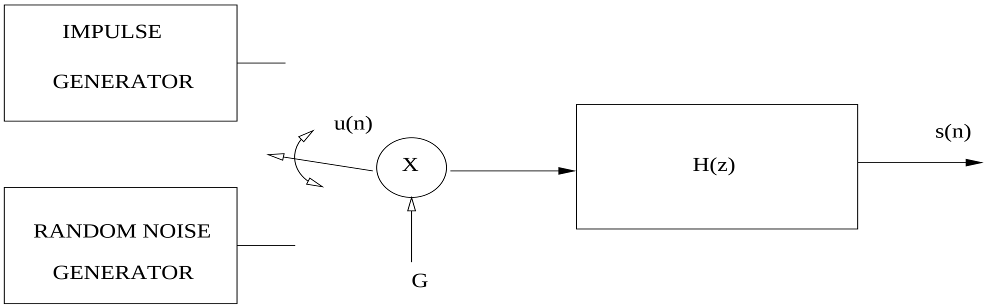

Figure B.1: Model of speech production for LP analysis

For speech analysis an all pole model is considered. For this case, the signal $s(n)$ is assumed as a linear combination of past values and some input $u(n)$. That is

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{B.2}$$

The transfer function of the filter is given by

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{B.3}$$

Given a particular signal $s(n)$, the problem is to determine the Linear Predictor Coefficients (LPC) ($\{a_k\}$) and the gain G in some manner. The LPCs are determined by minimizing the mean squared error over an analysis frame. The coefficients are obtained by solving the set of p normal equations

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), \qquad n = 1, \cdots, p \tag{B.4}$$

where

$$R(k) = \sum_{n=0}^{N-(p-1)} s(n)s(n-k), \qquad k = 1, \cdots, p \tag{B.5}$$

are the autocorrelation coefficients and $\{s(n)\}$ are the speech samples.

If $s(n)$ is the present sample, then it is predicted by the past p samples [96] as,

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{B.6}$$

where $\{a_k\}$ are the LPCs computed by the LP analysis.

The difference between the actual and predicted sample value is termed as prediction error or residual, which is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{B.7}$$

The residual is obtained by passing the speech signal through the inverse filter $A(z)$, which is given by,

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{B.8}$$

As the vocal tract system information is modeled by the linear prediction coefficients ($\{a_k\}$), the LP residual mostly contains information of the excitation source.

As we are operating directly on the speech signal in the time domain, even in high pitched female voices as long as the LP order is less than pitch period, it is possible to estimate the filter parameters. In case of noise/degraded speech and also for sounds not conforming to the all-pole model, the estimation of filter parameters will be poor. This results in large error in the LP residual. However, in the present work as we are further processing the LP residual, accurate estimation of the filter parameters is not very critical.

Appendix–C
HILBERT TRANSFORM RELATIONS

A signal can be either real or complex. All naturally generated signals are real in nature. In some applications, it is desirable to develop a complex signal from the given real signal. For instance, a complex signal helps in obtaining values of amplitude, phase and instantaneous frequency of the signal unambiguously. A complex signal can be generated from a real signal by employing a Hilbert transformer that is characterized by an impulse response $\frac{1}{\pi t}$.

If $s(t)$ is the real signal, then the Hilbert transform of $s(t)$ defined as $\hat{s}(t)$ is obtained by convolving $s(t)$ with $\frac{1}{\pi t}$.

$$\hat{s}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{s(\tau)}{(t-\tau)} d\tau \tag{C.1}$$

The inverse Hilbert transform, by means of which the original signal $s(t)$ is recovered from $\hat{s}(t)$ is defined by

$$s(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\hat{s}(\tau)}{(t-\tau)} d\tau \tag{C.2}$$

The funcations $s(t)$ and $\hat{s}(t)$ are said to constitute a Hilbert transform pair. The Hilbert transform is different from the Fourier transform in that it operates exclusively in the time domain.

From the convolution theorem it is known that the convolution of two functions in the time domain is equal to the multiplication of their Fourier transforms in the frequency domain.

For the time function, $\frac{1}{\pi t}$, we have

$$\frac{1}{\pi t} \rightleftharpoons -jsgn(f) \tag{C.3}$$

where $sgn(f)$ is the signum function, defined in the frequency domain as

$$sgn(f) = \left\{ \begin{array}{cc} 1, & f > 0 \\ 0, & f = 0 \\ -1 & f < 0 \end{array} \right\} \tag{C.4}$$

The Fourier transform $\hat{S}(f)$ of $\hat{s}(t)$ is given by

$$\hat{S}(f) = -jsgn(f)S(f) \tag{C.5}$$

The complex signal of $s(t)$ is defined as

$$s_+(t) = s(t) + j\hat{s}(t) \tag{C.6}$$

where, $\hat{s}(t)$ is the Hilbert transform of $s(t)$. The complex signal is also termed as analytic signal. The Fourier transform of $s_+(t)$ is given by

$$S_+(f) = S(f) + j(-jsgn(f))S(f) \tag{C.7}$$

$$S_+(f) = S(f) + sgn(f)S(f) \tag{C.8}$$

Using the definition of $sgn(f)$, we readily find that

$$S_+(f) = \left\{ \begin{array}{rl} 2S(f), & f > 0 \\ S(0), & f = 0 \\ 0 & f < 0 \end{array} \right\} \tag{C.9}$$

where $S(0)$ is the value of $S(f)$ at frquency $f = 0$. This means that the complex function has no frequency content (that is Fourier transform value) for all negative frequencies.

Hilbert Envelope:

Hilbert envelope of a signal $s(t)$ is defined as

$$s_o(t) = \sqrt{s^2(t) + \hat{s}^2(t)} \tag{C.10}$$

Hilbert Transform Relations for the DFT:

As discussed above one approach for developing the Hilbert transform relations is from the properties of analytic functions. Another approach is to use the properties of even and odd functions and causal sequences. Any sequence can be expressed as the sum of an even sequence and an odd sequence. Specifically, with $s_e(n)$ and $s_o(n)$ denoting the even and odd parts of $s(n)$, then

$$s(n) = s_e(n) + s_o(n) \tag{C.11}$$

where,

$$s_e(n) = \frac{1}{2}[s(n) + s(-n)] \tag{C.12}$$

and

$$s_o(n) = \frac{1}{2}[s(n) + s(-n)] \tag{C.13}$$

Also, $s(n)$ is causal, if $s(n)$ is zero for $n < 0$.

Equations (C.11)-(C.13) apply to an arbitrary sequence whether or not it is causal or whether or not it is real. However, if $s(n)$ is causal, then it is possible to recover $s(n)$ from $s_e(n)$ and to recover $s(n)$ for $n \neq 0$ from $s_o(n)$. Thus, for causal sequences

$$s(n) = \left\{ \begin{array}{rl} 2s_e(n), & n > 0 \\ s_e(n), & n = 0 \\ 0 & n < 0 \end{array} \right\} \tag{C.14}$$

and

$$s(n) = \left\{ \begin{array}{rl} 2s_o(n), & n > 0 \\ 0 & n < 0 \end{array} \right\} \tag{C.15}$$

Equivalently, if we define

$$u_+(n) = \left\{ \begin{array}{ll} 2, & n > 0 \\ 1, & n = 0 \\ 0 & n < 0 \end{array} \right\} \tag{C.16}$$

then

$$s(n) = s_e(n)u_+(n) \tag{C.17}$$

and

$$s(n) = s_o(n)u_+(n) + s(0)\delta(n) \tag{C.18}$$

We note that $s(n)$ can be completely recovered from $s_e(n)$. On the other hand, $s_o(n)$ will always be zero at $n = 0$, and consequently $s(n)$ can be recovered from $s_o(n)$ only for $n \neq 0$.

We can relate the real and imaginary parts of the DFT with a suitable definition of causality. A causal periodic sequence is one for which $s(n) = 0$ for $N/2 < n < N$. That is $s(n)$ is identically zero over the last half of the period. Because of this, it shall be clear that for causal peridic sequence

$$\hat{s}(n) = \left\{ \begin{array}{rl} 2\hat{s}_e(n), & n = 1, 2, ...(N/2) - 1 \\ \hat{s}_e(n), & n = 0, N/2 \\ 0, & n = (N/2) + 1..........(N - 1) \end{array} \right\} \tag{C.19}$$

and

$$\hat{s}(n) = \left\{ \begin{array}{rl} 2\hat{s}_o(n), & n = 1, 2, ...(N/2) - 1 \\ 0 & n = (N/2) + 1, ...(N - 1) \end{array} \right\} \tag{C.20}$$

Equivalently, if we define $\hat{u}_N(n)$ as a periodic sequence

$$\hat{u}_N(n) = \left\{ \begin{array}{ll} 1, & n = 0, N/2 \\ 2, & n = 1, 2, ..., (N/2) - 1 \\ 0, & n = (N/2+1), ..., (N-1) \end{array} \right\} \tag{C.21}$$

Then it follows that for N even we can express $h(n)$ as

$$\hat{s}(n) = \hat{s}_e(n)\hat{u}_N(n) \tag{C.22}$$

and

$$\hat{s}(n) = \hat{s}_o(n)\hat{u}_N(n) + s(0)\delta(n) + s(\frac{N}{2})\delta(n - \frac{N}{2}) \tag{C.23}$$

We note that $\hat{s}(n)$ can be completely recovered from $\hat{s}_e(n)$. On the other hand, $\hat{s}_o(n)$ will always be zero at $n = 0$ and $n = N/2$, and consequently $\hat{s}(n)$ can be recovered from $\hat{s}_o(n)$ only for $n \neq 0$ or $n \neq N/2$.

Bibliography

[1] R. L. Smith and J. J. Zwislocki, “Short-term adaptation and incremental responses of single auditory-nerve fibers,” *Biol. Cybernetics*, vol. 17, pp. 169–182, 1975.

[2] B. Yegnanarayana and N. J. Veldhuis, “Extraction of vocal-tract system characteristics from speech signals,” *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 313–327, July 1998.

[3] K. N. Stevens, *Acoustic Phonetics.* Cambridge, Massachusetts, USA: The MIT Press, 1999.

[4] D. Crystal, *A Dictionary of Linguistics and Phonetics.* Cambridge, Massachusetts, USA: Basil Blackwell Inc., 1985.

[5] M. A. Jack and J. Laver, *Aspects of Speech Technology.* 22 George Square, Edinburgh: Edinburgh University Press, 1988.

[6] R. L. Smith and J. J. Zwislocki, “Responses of some neurons of the cochlear nucleus to tone-intensity increments,” *J. Acoust. Soc. Amer.*, vol. 50, pp. 1520–1525, 1971.

[7] R. L. Smith, “Short-term adaptation in single auditroy-nerve fibers - an additive or a multiplicative effect?,” *J. Acoust. Soc. Amer.*, vol. 55, p. S85(A), 1974.

[8] N. Y. S. Kiang and E. C. Moxon, “Tails of tuning curves of auditory nerve fibers,” *J. Acoust. Soc. Amer.*, vol. 55, pp. 620–630, 1974.

[9] M. B. Sachs and E. D. Young, “Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate,” *J. Acoust. Soc. Amer.*, vol. 66, pp. 470–479, 1979.

[10] E. D. Young and M. B. Sachs, “Representation of steady state vowels in temporal aspects of discharge patterns of populations of auditory nerve fibers,” *J. Acoust. Soc. Amer.*, vol. 66, pp. 1381–1403, 1979.

[11] R. L. Smith and M. Brachman, “Operating range and maximum response of single auditory-nerve fibers,” *Brain Res.*, vol. 184, pp. 499–505, 1980.

[12] B. Delgutte, “Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers,” *J. Acoust. Soc. Amer.*, vol. 68, pp. 843–857, 1980.

[13] D. G. Sinex and C. D. Geisler, “Responses of auditory-nerve fibers to consonant-vowel syllables,” *J. Acoust. Soc. Amer.*, vol. 73, pp. 602–615, 1983.

[14] M. I. Miller and M. B. Sachs, “Representation of stop consonants in the discharge patterns of auditory-nerve fibers,” *J. Acoust. Soc. Amer.*, vol. 74, pp. 502–517, 1983.

[15] B. Leshowitz and E. Cudahy, "Masking patterns for continuous and gated sinusoids," *J. Acoust. Soc. Amer.*, vol. 58, pp. 235–242, 1975.

[16] S. Y. Zhukov, M. G. Zhukova, and L. A. Chistovich, "Some new concepts in the auditory analysis of acoustic flow," *Sov. Phys. Acoust.*, vol. 20, pp. 237–240, 1974.

[17] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 64(5), pp. 1358–1368, 1978.

[18] D. W. Massaro and G. C. Oden, "Evaluation and integration of acoustic features in speech," *J. Acoust. Soc. Amer.*, vol. 67(3), pp. 996–1013, 1980.

[19] K. N. Stevens, "Acoustic correlates of some phonetic categories," *J. Acoust. Soc. Amer.*, vol. 68(3), pp. 836–842, 1980.

[20] S. E. Blumstein and K. N. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," *J. Acoust. Soc. Amer.*, vol. 67(2), pp. 648–662, 1980.

[21] D. Kewley-Port, "Time-varying features as correlates of place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 73(1), pp. 322–335, 1983.

[22] W. A. Grimm, "Perception of segments of English-spoken consonant-vowel syllables," *J. Acoust. Soc. Amer.*, vol. 40, pp. 1454–1461, 1966.

[23] M. E. Teikeli and W. L. Cullinan, "The perception of temporally segmented vowels and consonant-vowel syllables," *J. Speech Hear. Res.*, vol. 22, pp. 103–121, 1979.

[24] R. N. Ohde and K. N. Stevens, "Effect of burst amplitude on the perception of stop consonant place of articulation," *J. Acoust. Soc. Amer.*, vol. 74(3), pp. 706–714, 1983.

[25] V. C. Tartter, D. Kat, A. G. Samuel, and B. H. Repp, "Perception of intervocalic stop consonants: The contributions of closure duration and formant transitions," *J. Acoust. Soc. Amer.*, vol. 74(3), pp. 715–725, 1983.

[26] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, vol. 80(4), pp. 1016–1025, 1986.

[27] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 309–319, Aug. 1979.

[28] K. S. Rao and B. Yegnanarayana, "Prosodic manipulation using instants of significant excitation," in *Proc. IEEE Int. Conf. Mul., Expo*, vol. I, (Baltimore, MD, USA), pp. 389–392, July 2003.

[29] S. R. M. Prasanna, S. V. Gangashetty, and B. Yegnanarayana, "Significance of vowel onset point for speech analysis," in *Signal Processing and Communications*, (Indian Instiute of Science, Bangalore, India), pp. 81–88, July 2001.

[30] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. I, (Orlando, FL, USA), pp. 541–544, May 2002.

[31] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 267–281, May 2000.

[32] B. Yegnanarayana, S. R. M. Prasanna, and M. Mathew, “Enhancement of speech in multispeaker environment,” in *Proc. European Conf. Speech Processing, Technology*, (Geneva, Switzerland), pp. 581–584, Sept. 2003.

[33] C. C. Sekhar, *Neural network models for recognition of stop consonant-vowel (SCV) segments in continuous speech.* PhD thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, India, 1996.

[34] J. M. Zachariah, *Text-dependent speaker verification using segmental, suprasegmental and source features.* MS thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, India, 2002.

[35] A. N. Sobakin, “Digital computer determination of formant parameters of the vocal tract from a speech signal,” *Soviet Phys.-Acoust.*, vol. 18, pp. 84–90, 1972.

[36] H. W. Strube, “Determination of the instant of glottal closures from the speech wave,” *J. Acoust. Soc. Amer.*, vol. 56, pp. 1625–1629, 1974.

[37] T. V. Ananthapadmanabha and B. Yegnanarayana, “Epoch extraction of voiced speech,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 562–570, Dec. 1975.

[38] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by Linear prediction of the speech wave,” *J. Acoust. Soc. Amer.*, vol. 50(2), pp. 637–655, 1971.

[39] D. Y. Wong, J. D. Markel, and A. H. Gray, “Least squares glottal closure inverse filtering from the acoustic speech waveform,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 350–355, Aug. 1979.

[40] Y. M. Cheng and D. O’Shaughnessy, “Automatic and reliable estimation of glottal closure instant and period,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1805–1815, Dec. 1989.

[41] Y. K. C. Ma and L. F. Willems, “A Frobenius norm approach to glottal closure detection from the speech signal,” *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 258–265, Apr. 1994.

[42] B. Yegnanarayana and R. L. H. M. Smits, “A robust method for determining instants of major excitations in voiced speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, (Detroit, USA), pp. 776–779, May 1995.

[43] R. Smits and B. Yegnanarayana, “Determination of instants of significant excitation in speech using group delay function,” *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 325–333, Aug. 1995.

[44] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, “A comparative performance study of several pitch detection algorithms,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 399–418, Oct. 1976.

[45] A. M. Noll, “Cepstrum pitch determination,” *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, 1967.

[46] J. D. Markel, “The SIFT algorithm for fundamental frequency estimation,” *IEEE Trans. Audio Electrocoust.*, vol. AU-20, pp. 367–377, Dec. 1972.

[47] D. Johnson and D. Dudgeon, *Array Signal Processing- Concepts and Techniques.* New-Jersey: Prentice Hall, 1993.

[48] G. Carter, "Variance bounds for passively locating an acoustic source with a symmetric line array," *J. Acoust. Soc. Amer.*, vol. 62, pp. 922–926, 1977.

[49] J. H. DiBiase, H. Silverman, and M. Brandstein, *Robust Localization in Reverberant Rooms (Ch.7) in Theory and Applications Acoustic Signal Processing for Telecommunications*, pp. 131–154. Boston: Kluwer Academic Publishers, 2000.

[50] W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 608–614, Sept. 1973.

[51] S. Haykin, *Adaptive Filter Theory.* New Jersey: Prentice Hall, second ed., 1991.

[52] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 45–50, Jan. 1997.

[53] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 320–327, Aug. 1976.

[54] A. Stephene and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, (Detroit, MI, USA), pp. 3055–3058, May 1995.

[55] S. Oh and V. Viswanathan, "Hands-free voice communication in an automobile with a microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (San Francisco, California, U.S.A.), pp. 281–284, Mar. 1992.

[56] P. Scalart and A. Benmar, "A system for speech enhancement in the context of hands-free radiotelephony with combined noise reduction and acoustic echo cancellation," *Speech Communication*, vol. 20, pp. 203–214, Dec. 1996.

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

[58] J. L. Flanagan, J. D. Jonston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78(5), pp. 1508–1518, 1985.

[59] H. F. Silverman, "Some analysis of microphone arrays for speech data acquisition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1699–1712, Dec. 1987.

[60] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 145–152, Feb. 1988.

[61] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Processing*, vol. 39, pp. 1943–1954, Sept. 1991.

[62] S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 392–396, Sept. 1996.

[63] P. Satyanarayana, *Short segment analysis of speech for enhancement.* PhD thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, India, 1999.

[64] B. Yegnanarayana, C. Avendaño, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, pp. 25–42, May 1999.

[65] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 731–740, Oct. 2001.

[66] E. Nemer, R. Goubran, and S. Mahmoud, "Speech enhancement using fourth-order cumulants and optimum filters in the subband domain," *Speech Communication*, vol. 36, pp. 219–246, Mar. 2002.

[67] D. V. Compernolle, "Switching adaptive filters for enhancing noisy and reveberant speech from microphone array recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Albuquerque, New Mexico, USA), pp. 833–836, Apr. 1990.

[68] H. Wang and F. Itakura, "An approach of dereverberation using multi-microphone sub-band envelope estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (oronto, Ontario, Canada), pp. 953–956, May 1991.

[69] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, pp. 911–918, 1976.

[70] C. K. Lee and D. G. Childers, "Cochannel speech separation," *J. Acoust. Soc. Amer.*, vol. 83(1), pp. 274–280, 1988.

[71] D. Morgan, E. B. George, L. T. Lee, and S. Kay, "Cochannel speech separation by harmonic enhancement and supression," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 407–424, Sept. 1997.

[72] O. M. M. Mitchell, C. Ross, and G. Yates, "Signal processing for a cocktail party effect," *J. Acoust. Soc. Amer.*, vol. 50, pp. 656–660, 1971.

[73] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 86, pp. 2009–2025, Oct. 1998.

[74] A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavlets," *IEEE Trans. Neural Networks*, vol. 13, pp. 888–893, July 2002.

[75] B. H. Hanson and D. Y. Wong, "The harmonic magnitude supression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (San Diego, CA, USA), pp. 18A.5.1–18A.5.4, 1984.

[76] T. F. Quatieri and R. G. Danisewicz, "An approach to cochannel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 56–69, Jan. 1990.

[77] S.-I. Amari and A. Cichocki, "Adaptive blind signal processing-neural network approaches," *Proc. IEEE*, vol. 86, pp. 2026–2048, Oct. 1998.

[78] S. Choi, H. Hong, H. Glotin, and F. Berthommier, "Multichannel signal separation for cocktail party speech recognition: A dynamic recurrent network," in *Proc. Int. Conf. Spoken Language Processing*, (Beijing, China), pp. 83–87, 2000.

[79] A. K. Barros, F. Itakura, T. Rutkowski, A. Mansour, and N. Ohnishi, "Estimation of speech embedded in a reverberant environment with multiple sources of noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Salt Lake City, Utah, USA), 2001.

[80] D. J. Hermes, "Vowel-onset detection," *J. Acoust. Soc. Amer.*, vol. 87, pp. 866–873, 1990.

[81] J.-F. Wang, C.-H. Wu, Shin-Hung, and J.-Y. Lee, "A hierarchical neural network based on a C/V segmentation algorithm for isolated Mandarin speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 2141–2146, 1991.

[82] J.-F. Wang and S.-H. Chen, "A C/V segmentation algorithm for Mandarin speech signal based on wavelet transforms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 1261–1264, Sept. 1999.

[83] J. Y. S. R. K. Rao, C. C. Sekhar, and B. Yegnanarayana, "Neural network based approach for detection of vowel onset points," in *Int. Conf. Advances in Pattern Recognition and Digital Techniques, (ISI Calcutta, India)*, pp. 316–320, 1999.

[84] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, New Jersy: Prentice-Hall, 1993.

[85] L. R. Rabiner and J. G. Wilpon, "Speaker-independent isolated word recognition for a moderate size (54 word) vocabulary," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 583–587, Dec. 1979.

[86] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 1, pp. 52–59, 1986.

[87] H. Sakoe, "Two-level DP matching–a dynamic programming-based pattern matching algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 589–595, Dec. 1979.

[88] C. S. Myers and L. R. Rabiner, "A level building dynamic time warping algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 575–582, 1978.

[89] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254–272, Apr. 1981.

[90] M. Mathew, *Combining evidences from multiple classifiers for text-dependent speaker recognition.* MS thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, India, 1999.

[91] L. F. Lamel, L. R. Rabiner, and A. E. Rosenberg, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 777–785, 1981.

[92] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, pp. 297–315, Feb. 1975.

[93] M. H. Savoji, "A robust algorithm for accurate endpointing of speech," *Speech Communication*, vol. 8, pp. 45–60, June 1989.

[94] C. Tsao and R. M. Gray, "An endpoint detector for lpc speech using residual look-ahead for vector quantization applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (San Diego, California, USA), pp. 97–100, Mar. 1984.

[95] M. Hamada, Y. Takizawa, and T. Norimatsu, "A noise robust speech recognition system," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, (Kobe, Japan), pp. 893–896, Nov. 1990.

[96] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[97] A. V. Oppenheim and R. W. Schafer, *Digital signal processing.* Englewood Cliffs, New Jersey: Prentice Hall, 1975.

[98] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing.* Upper Saddle River, New Jersey: Prentice Hall, 2000.

[99] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Modeling of Data*, ch. 15, pp. 656–666. Numerical Recipes in C: The Art of Scientific Computing, New Delhi: Foundation Books for Cambridge University Press, second ed., 1992.

[100] M. Brandstein and S. Griebel, *Explicit speech modeling for microphone array applications (Ch.6) in Theory and Applications Acoustic Signal Processing for Telecommunications.* Boston: Kluwer Academic Publishers, 2000.

[101] M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Amer.*, vol. 105, pp. 2914–2919, 1999.

[102] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 609–619, Nov. 1999.

[103] B. Yegnanarayana, S. R. M. Prasanna, and S. V. Gangashetty, "Autoassociative neural network models for speaker recognition," in *proc. Int. Workshop Embedded Systems*, (Hyderabad, India), 2001.

[104] B. Yegnanarayana, C. Avendaño, H. Hermansky, and P. S. Murthy, "Processing linear prediction residual for speech enhancement," in *Proc. European Conf. Speech Processing, Technology*, (Rhodes, Greece), pp. 1399–1402, Sept. 1997.

[105] U. Mittal and N. Phamdo, "Signal/noise klt based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 159–167, Mar. 2000.

[106] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, 1995.

[107] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Seattle, WA, USA), pp. 377–380, May 1998.

[108] R. H. Frazier *et. al.*, "Enhancement of speech by adaptive filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (New York, NY, USA), 1990.

[109] ICA'99, "Int. workshop on independent component analysis and blind signal separation," in *http://www2.ele.tue.nl/ica99/realworld.html*, 1999.

[110] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of speech Signals*. Newyork: Macmillan, 1993.

[111] D. G. Childers and C. K. Lee, "Voice quality factors: Analysis synthesis and perception," *J. Acoust. Soc. Amer.*, vol. 90, pp. 2394–2410, 1991.

[112] P. E. Papamichalis, *Practical approaches to speech coding*. NewJersy: Prentice Hall, Englewood Cliffs, 1987.

[113] P. Satyanarayana, *Short segment analysis of speech for enhancement*. PhD thesis, Indian Institute of Technology Madras, Department of Electrical Engg., Chennai, India, 1999.

[114] D. Gabor, "Theory of communication," *J. IEE (London, UK)*, vol. 93, pp. 429–457, 1946.

[115] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 777–785, Aug. 1981.

[116] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43–49, Feb. 1978.

[117] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 575–582, Dec. 1978.

[118] S. R. M. Prasanna, S. V. Gangashetty, and B. Yegnanaryana, "Significance of vowel onset point for speech analysis," in *Signal Processing and Communications*, (IISc, Bangalore, India), 2001.

[119] S. R. M. Prasanna, J. M. Zachariah, and B.Yegnanarayana, "Begin-end detection using vowel onset points," in *Proc. Workshop on Spoken Language Processing*, (Tata Institute of Fundamental Research, Mumbai, India), pp. 33–40, Jan. 2003.

[120] J. D. Markel and A. H. Gray, *Linear prediction of speech*. Berlin Heidelberg, New York: Springer-Verlag, 1976.

LIST OF PUBLICATIONS

Refereed Journals

- B. Yegnanarayana, S.R.M. Prasanna, R. Duraiswamy, and D. Zotkin,"Processing of reverberant speech for time-delay estimation," accepted for publication in *IEEE Trans. Speech, Audio Processing*, 2004.

- B. Yegnanarayana, S.R.M. Prasanna, J.M. Zachariah and C.S. Gupta,"Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," accepted for publication in *IEEE Trans. Speech, Audio Processing*, 2004.

- V.C. Raykar, B. Yegnanarayana, S.R.M. Prasanna, and R. Duraiswami,"Speaker localization using excitation source information in speech," accepted for publication in *IEEE Trans. Speech, Audio Processing*, 2003.

- S.R.M. Prasanna, C.S. Gupta, and B. Yegnanarayana,"Extraction of speaker-specific information from linear prediction residual of speech," revised and re-submitted to *J. Acoust., Soc., Amer.*, 2003.

- B. Yegnanarayana and S.R.M. Prasanna," Multispeaker speech enhancement in multichannel case using excitation source information," communicated to *J. Acoust., Soc., Amer.*, 2004.

Refereed International Conferences/Workshops:

- S.V. Gangashetty and S.R.M. Prasanna, " Significance of vowel onset point for speech recognition using neural networks," *Proc. Fifth Int. Conf. Cognitive Neural Systems*, Boston University (Boston, MA, USA), May-June 2001.

- S.R.M. Prasanna, S.V. Gangashetty, and B. Yegnanarayana, " Significance of vowel onset point for speech analysis," in *Proc. Signal Proc. Com.*, Indian Institute of Science (Bangalore, India), July 2001.

- B. Yegnanarayana, S.R.M. Prasanna, and S.V. Gangashetty, " Autoassociative neural network models for speaker recognition," in *Proc. Int. Workshop Embedded Systems*, (Hyderabad, India), Dec. 2001.

- B. Yegnanarayana, S.R.M. Prasanna, and K.S. Rao, " Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Orlando, Fl, USA), May 2002.

- S.R.M. Prasanna, and J.M. Zachariah, "Detection of vowel onset point in speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Orlando, Fl, USA), May 2002.

- S.V. Gangashetty, S.R.M. Prasanna, and B. Yegnanarayana, " Linear and nonlinear compression of feature vectors for speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, (Orlando, FL, USA), May 2002.

- C.S. Gupta, S.R.M. Prasanna, and B. Yegnanarayana, " Autoassociative neural network models for online speaker verification using source features from vowels," in *Proc. Int. Joint Conf. Neural Networks*, (Honolulu, Hawaii, USA), May 2002.

- S.V. Gangashetty, A.N. Khan, S.R.M. Prasanna, and B. Yegnanarayana, " Neural network models for preprocessing and discriminating utterances of consonant vowel units," in *Proc. Int. Joint Conf. Neural Networks*, (Honolulu, Hawaii, USA), May 2002.

- S.R.M. Prasanna, C.S. Gupta, and B. Yegnanarayana, " Autoassociative neural network models for speaker verification using source features," in *Proc. Sixth Int. Conf. Cognitive Neural Systems*, Boston University, (Boston, MA, USA), May-June 2002.

- B. Yegnanarayana, S.R.M. Prasanna, and M. Mathew,"Enhancement of speech in multispeaker environment," in *Proc. EUROSPEECH'03*, (Geneva, Switzerland), Sept. 2003.

- V.C. Raykar, R. Duraiswamy, B. Yegnanarayana, and S.R.M. Prasanna, " Tracking a moving speaker using excitation source information," in *Proc. EUROSPEECH'03*, (Geneva, Switzerland), Sept. 2003.

- S.R.M. Prasanna, J.M. Zachariah and B. Yegnanarayana, "Neural network models for combining evidence from spectral and suprasegmental features" *Proc. Int.*

Conf. Intelligent Sensing, Information Processing, (Chennai, India), Jan. 2004.

- S.R.M. Prasanna, and B. Yegnanarayana, "Extraction of pitch in adverse conditions," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Montreal, Canada), May 2004.
- L. Mary, K.S.R. Murty, S.R.M. Prasanna and B. Yegnanarayana, "Features for speaker and language identification", *Proc. ODYSSEY 2004: the speaker and language recognition workshop* (Toledo, Spain), May-June, 2004.
- S.R.M. Prasanna, and B. Yegnanarayana, "Speech enhancement using source features and group delay analysis," to be communicated to *ICASSP*, 2005.

Refereed National Conferences/Workshops:

- S.R.M. Prasanna, J.M. Zachariah and B. Yegnanarayana, " Begin-end detection using vowel onset points," in *Proc. Workshop on Spoken Language Processing*, Tata Institute of Fundamental Research, (Mumbai, India), Jan. 2003.

CURRICULUM VITAE

1. **NAME**: S.R. Mahadeva Prasanna

2. **DATE OF BIRTH**: 08 July 1971

3. **EDUCATIONAL QUALIFICATIONS**:

   - 1993 Bachelor of Engineering (B.E.)
   - 1997 Master of Technology (M.Tech.)
   - 2004 Doctor of Philosophy (Ph.D.)

4. **PERMANENT ADDRESS**:

   s/o S. K. Rajashekharaiah
   Sompura (P), Koratagere (T), Tumkur (D)
   Karnataka (S) - 572121
   Ph: +91-08138-234524

DOCTORAL COMMITTEE

1. **CHAIRPERSON**: Prof. S. Raman

2. **GUIDE**: Prof. B. Yegnanarayana

3. **MEMBERS**:

    - Prof. C. Siva Ram Murthy

    - Prof. D. Janaki Ram

    - Prof. V. V. Rao

    - Prof. Megha Singh