SIGNIFICANCE OF EXCITATION SOURCE INFORMATION FOR SPEECH ANALYSIS

A THESIS

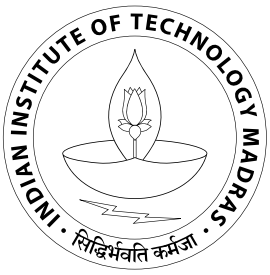
submitted by

SRI RAMA MURTY KODUKULA

for the award of the degree

of

DOCTOR OF PHILOSOPHY



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

MARCH 2009

To My Parents

Lakshmana Rao *and* **Sarada**

and My Guide

Prof. B. Yegnanarayana

THESIS CERTIFICATE

This is to certify that the thesis titled **Significance of Excitation Source Information for Speech Analysis**, submitted by **Sri Rama Murty Kodukula**, to the Indian Institute of Technology Madras, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under our 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.

Prof. B. Yegnanarayana
Dept. of Computer Science and Engg.

Dr. C. Chandra Sekhar
Dept. of Computer Science and Engg.

Date :

Chennai, 600 036

Acknowledgements

I would like to express my deepest respect and sincere gratitude to my guide Prof. B. Yegnanarayana, for his constant guidance and encouragement at all stages of my work. I still remember the day, 30 July 2002, when I received a telegram from him informing me that I was considered for MS programme under his guidance. By sheer coincidence, on 30 July 2004, he accepted to convert my registration to PhD programme. I am grateful to him for twice accepting me as his student. I am fortunate to have had numerous technical discussions with him, from which I have benefited enormously. I sincerely thank him for all his personal interest in me, and all those valuable long hours he has spent with me in observing speech signals and analyzing results. I thank him for patiently correcting several drafts of my papers. His 3-D principle of dedication, determination and discipline, has been a constant source of inspiration for me to continue in the field of research. I thank him for the excellent research environment he has created for all of us to learn.

I am extremely grateful to Dr. C. Chandra Sekhar, my co-guide, for his constant encouragement and support all through my research work. I am thankful to him for all the invaluable advice on both technical and nontechnical matters. The discussions with him have been a source of motivation and energy for me to persist in the field of research. I thank him for patiently correcting the final draft of my thesis within a short duration.

I thank Dr. Hema A. Murthy for all the thought provoking discussions and constructive criticism. I thank Prof. S. Raman and Prof. Timothy A. Gonsalves, chairpersons of the department during the course of my research work, for providing constant support and excellent facilities to carryout my research. I thank the members of my doctoral committee, Prof. D. Janaki Ram, Dr. V. Kamakoti, Prof. C. Sujatha and Dr. K. Sridharan, for

sparing their valuable time to evaluate the progress of my research work.

Thanks are due to Dr. G. F. Mayer for readily providing access to the Keele pitch evaluation database used in this work. Thanks are also due to the many people who generously created, edited and distributed the software codes that were used for comparative studies. I offer my sincere apologies in the event that my choice of parameters did not do the methods justice. Thanks to all the anonymous reviewers of our publications for their in-depth criticism which has helped me a lot in improving the presentation of this work.

Sincere thanks to Prof. Rajiv Sanghal, Director, IIIT Hyderabad, for allowing me to stay and pursue my research work at IIIT for the past two years, and generously sponsoring me a travel grant to Belgium for presenting my paper at Interspeech-2007. I am extremely grateful to the DRDO laboratories, DRDL-Hyderabad, NSTL-Vizag, and CAIR-Bangalore, for financially supporting me throughout the course of my research work.

I thank Dr. S. Rajendran for maintaining an excellent computing facility and for the valuable suggestions he gave me at different stages of my work. My special thanks to Dr. S. R. M. Prasanna for motivating me to pursue research. I am thankful to Dr. Suryakanth for his immense support and help all through my research work, especially during my stay at IIIT Hyderabad.

I will forever remember the wonderful time I have had with my dearest friends Anand, Anil, Dhanu and Guru. I thank Anand, the computer wizard, for all his instant support on softwares and data recording. I am thankful to Anil, the coolest guy, for all the moral support and encouragement he provided during the periods of distress. I am indebted to Dhanu, the synonym of generosity, for his immense help throughout my research work, especially during the period I fell sick. My sincere thanks to Guru, the moving encyclopedia of cricket and movies, for helping me in writing my papers and patiently explaining me the meanings of Hindi songs.

Special thanks to Dileep for the immense help and support he extended during my visits to Chennai. I thank my MS batchmates Chaitu, Satish, Suresh and Panuku for the wonderful moments we shared together. I thank all the past and present lab-mates for the friendly and conductive atmosphere in SVL.

Needless to mention the love and moral support of my parents, sisters and brother-in-laws. This work would not have been possible but for their support. Words are not enough to express my special appreciation to my fiancee Sloka, for her constant encouragement, patience and understanding of my mood-swings.

Finally, I would like to dedicate this thesis to my parents, Lakshmana Rao and Sarada, and to my guide, Prof. B. Yegnanarayana.

Sri Rama Murty

Abstract

The primary mode of excitation of the vocal-tract system during speech production is due to the vibration of the vocal folds. For voiced speech, the most significant excitation takes place around the instant of glottal closure, called the *epoch*. The objective of this work is to extract the epoch locations and estimate their excitation strengths from the speech signal. Conventional methods for extracting the excitation source features rely on modeling the response of the vocal-tract as parameters of an all-pole filter, and then inverse filtering the speech signal to estimate the source information. Accuracy of these methods depends critically on our ability to model the time-varying response of the vocal-tract system. In this work, we propose methods to extract the features of excitation source using the impulse-like nature of excitation. The proposed methods do not depend on modeling the response of the vocal-tract system.

The effect of an impulse is spread uniformly across the frequency domain including at zero-frequency. Around the zero-frequency, the response of the vocal-tract system is significantly low compared to the response of the impulse-like excitation. In this work, the impulse-like nature of excitation is exploited by filtering the speech signal at zero-frequency to extract the epoch locations and their strengths of excitation. Using the epoch locations as anchor points within each glottal cycle, a method to estimate the instantaneous fundamental frequency of voiced speech is proposed. The strengths of excitation at the epochs are used to detect the regions of vocal fold vibration, which is referred to as glottal activity. Using the robustness of relative spacing between the epochs in speech signals collected over a pair of microphones, methods for pitch extraction in reverberant environment and multispeaker environment are proposed. The proposed method of extracting the glottal activity together with linear prediction analysis is used to study the

role of excitation source in the analysis of manner of articulation of stop consonants. Robustness of the proposed epoch extraction and fundamental frequency estimation methods has been studied and compared with the state-of-the-art methods.

Keywords: *Epoch extraction, glottal closure instant, instantaneous frequency, pitch, strength of excitation, multimicrophone processing, manner of articulation, stop consonants, zero-frequency resonator.*

Contents

List of Tables

List of Figures

Abbreviations

BSS	- Blind source separation
DET	- Detection error trade-off
DYPSA	- Dynamic programming projected phase-slope algorithm
EER	- Equal error rate
EGG	- Electroglottograph
FAR	- False acceptance rate
FRR	- False rejection rate
GAD	- Glottal inverse filtering
GCC	- Generalized crosscorrelation
GCI	- Glottal closure instant
IDFT	- Inverse discrete Fourier transform
IFT	- Inverse Fourier transform
IIR	- Infinite impulse response
LP	- Linear prediction
MLED	- Maximum likelihood epoch detection
SHS	- Subharmonic summation
SNR	- Signal-to-noise ratio
SVD	- Singular value decomposition
TDOA	- Time difference of arrival
VOT	- Voice onset time

Chapter 1

Introduction

Speech signal can be considered as the output of a linear system for which neither the excitation nor the system response is known. In particular, voiced speech is the output of a quasistationary vocal-tract system excited with quasiperiodic puffs of air produced due to vibration of vocal folds. Although the vibration of vocal folds produces a sequence of glottal pulses, the significant excitation to the vocal-tract system within each glottal cycle can be considered to occur around the instant of glottal closure, called *epoch*. Epoch location marks the start of the closed glottis region during which there is little or no airflow through the glottis. Accurate identification of closed glottis region allows the blind deconvolution of the vocal-tract and excitation source. Characterization of the excitation source features has great potential for use in speech analysis, synthesis, coding, speaker recognition, and diagnosis of voice disorders.

During speech production, the vocal-tract responses at successive glottal pulses overlap forming a composite signal. Extracting the excitation information from speech signals is a challenging task, as it is difficult to suppress the response of the time-varying vocal-tract system in the speech signal. The existing methods for extracting the excitation information from the speech signal are based on glottal inverse filtering [1, 2, 3]. These methods assume that the speech signal is produced as the response of a linear time-invariant system to an excitation signal having a flat spectrum. Glottal inverse filtering involves the estimation of the characteristics of vocal-tract system in terms of the parameters of a

linear filter. For instance, linear prediction (LP) analysis [4] is the standard method used to estimate the parameters of the filter under the assumption that the vocal-tract can be modeled as an all-pole filter. The effectiveness of the LP analysis in characterizing the excitation information depends on the accuracy of the all-pole model, and the nature and quality of the speech signal. Moreover, accurate estimation of the vocal-tract response and excitation source are interdependent problems, as the accurate estimation of one depends on the accurate estimation of the other. Reliable glottal inverse filtering requires accurate estimation of the parameters of the all-pole filter representing the vocal-tract system, which in turn depends on the accurate identification of the closed glottis region. It is desirable to characterize the excitation information from speech signals independent of the influence of the vocal-tract system.

1.1 Objective and scope of the work

The objective of this work is to extract important features of the source of excitation from the speech signal independent of the influence of the vocal-tract. The features of the excitation source considered in this work include the locations of the epochs and their strengths of excitation. The locations of the epochs along with their strengths of excitation can be used in several speech analysis situations. Fundamental frequency of the voiced speech can be accurately estimated using epoch as the anchor point in each glottal cycle. The strengths of excitation of the epochs can be used to detect the regions of glottal activity. Locations of the epochs along with their strengths can be used to analyze the manner of articulation of the stop consonants.

This work is based on the assumption that the excitation to the vocal-tract system can be approximated by a sequence of impulses of varying strengths. Hence, the methods proposed in this work are not likely to work well when the degradations produce additional impulse-like sequences in the collected speech signal as in the case of reverberation. The methods may also not work well when there is interference of speech from other speakers.

1.2 Organization of the thesis

The evolution of ideas presented in this thesis is listed in Table 1.1. The contents of the thesis are organized as follows:

In **Chapter 2**, we highlight the significance of epochs in speech analysis, and review the existing methods for epoch extraction. In this chapter, we also review methods for characterizing the strengths of excitation of epochs, detecting glottal activity and estimating fundamental frequency of voiced speech.

In **Chapter 3**, we propose a novel approach to epoch extraction from speech signals, by confining the analysis to a narrow-band of frequencies around zero-frequency. The performance of the proposed approach is evaluated and the results are compared with the state-of-the-art methods for epoch extraction.

In **Chapter 4**, a method for determining the strength of excitation of the epoch from speech signals is proposed. The proposed measure of strength of excitation has a close linear relationship with the negative peak amplitude of the differentiated glottal flow. Since the differentiated glottal flow is almost zero in unvoiced regions, we present a method for glottal activity (voicing) detection based on the estimated strengths of excitation. The strengths of excitation and the regions of glottal activity derived from the speech signal are compared with simultaneous recordings of EGG signals.

In **Chapter 5**, we highlight the need for determining the instantaneous fundamental frequency as compared to the "average pitch" obtained by the conventional block processing methods. Then, we propose a method for estimating the instantaneous fundamental frequency of voiced speech segments using the epoch as anchor point in each glottal pulse. The performance of the proposed method is evaluated, and compared with existing techniques under different noisy environments at varying levels of degradation.

In **Chapter 6**, we propose methods based on epoch extraction to process multimicrophone data in order to overcome the issues involved in pitch estimation in reverberant environment and multispeaker environment. A method for estimating time-delay of arrival between a pair of spatially separated microphones using the excitation source information

Table 1.1: Evolution of ideas presented in the thesis

- In voiced speech, the primary acoustic excitation normally occurs at the instant of glottal closure (epoch), and is impulse-like.
- The effect due to the impulse-like excitation is reflected uniformly across all the frequencies, irrespective of the state of the vocal-tract system.
- The impulse-like excitations reflect as discontinuities in the time domain, which can be highlighted by computing the instantaneous frequency of the speech signal filtered through a narrow-band filter. The center frequency of the narrow-band filter depends critically on the vocal-tract response.
- The contribution of the vocal-tract response at zero-frequency is significantly less compared to the contribution of the response of the impulse-like excitation. A method based on zero-frequency resonator is proposed for the extraction of the epochs and their strengths of excitation.
- The detected epoch is used as anchor point in each glottal cycle to estimate the instantaneous fundamental frequency of speech signals. Since this method does not depend on correlation of the speech segments in adjacent pitch cycles, the method is robust even for diplophony and creaky voices.
- Using the robustness of relative spacing between the epochs in speech signals collected over a pair of microphones, methods for pitch extraction in reverberant environment and multispeaker environment are proposed.
- The excitation information derived from proposed method along with the LP analysis is used to study the nature of excitation in the stop consonants.

is discussed.

In **Chapter 7**, we apply the excitation features derived using the epoch based method proposed in this thesis, along with the LP analysis for studying the production characteristics of stop consonants, specifically the manner of articulation.

In **Chapter 8**, we summarize the contributions of the present work, and discuss some issues which prompt further investigation for extracting excitation features from speech signals collected in practical environments.

Chapter 2

Extraction of Excitation Information - A Review

This chapter reviews some of the existing methods for extracting and processing excitation source information and highlights the issues involved. In particular, we review specific methods for extraction of the epochs, their strengths and fundamental frequency. In Section 2.1, we explain the significance of epochs in speech analysis. In Section 2.2, we illustrate the important features of glottal flow as measured by the electroglottograph signal. Section 2.3 reviews the existing methods for epoch extraction from speech signals. In Section 2.4, we highlight the significance of negative peak amplitude of differentiated glottal flow, which we refer to as strength of excitation, in voice source analysis and review approaches to estimate it. In Section 2.5, we review the existing approaches for pitch estimation from speech signals. In Section 2.6, we review methods to process multimicrophone speech data for time-delay estimation and multispeaker speech processing. Section 2.7 reviews existing methods for analysis of manner of articulation of stop consonants and estimating the voicing onset times. Finally Section 2.8 summarizes the review and highlights the important issues addressed in this thesis.

2.1 Significance of epochs in speech analysis

Voiced speech analysis consists of determining the frequency response of the vocal-tract system and the glottal pulses representing the excitation source. Although the source of excitation for voiced speech is a sequence of glottal pulses, the significant excitation of the vocal-tract system within a glottal pulse, can be considered to occur at the instant of glottal closure (GCI), called the *epoch*. Many speech analysis situations depend on accurate estimation of the location of the epoch within a glottal pulse. For example, knowledge of the epoch locations is useful for accurate estimation of the fundamental frequency (f_0). Often the glottal airflow is zero soon after the glottal closure. As a result, the supralaryngeal vocal-tract is acoustically decoupled from the trachea. Hence the speech signal in the closed glottis region represents free resonances of the supralaryngeal vocal-tract system. Analysis of speech signals in the closed glottis regions provides an accurate estimate of the frequency response of the supralaryngeal vocal-tract system [5, 6]. With the knowledge of the epochs, it may be possible to determine the characteristics of the voice source by a careful analysis of the signal within a glottal pulse. The epochs can be used as pitch markers for prosody manipulation, which is useful in applications like text-to-speech synthesis, voice conversion and speech rate conversion [7, 8]. Knowledge of the epoch locations may be used for estimating the time-delay between speech signals collected over a pair of spatially separated microphones [9]. The segmental signal-to-noise ratio (SNR) of the speech signal is high in the regions around the epochs, and hence it is possible to enhance the speech by exploiting the characteristics of speech signals around the epochs [10]. It has been shown that the excitation features derived from the regions around the epoch locations provide complementary speaker-specific information to the existing spectral features [11, 12, 13].

As a result of significant excitation at the epochs, the regions in the speech signal that immediately follow them are relatively more robust to (external) degradations than other regions. The instants of significant excitation play an important role in human perception also. It is because of the epochs in speech that human beings seem to be able to perceive speech even at a distance (e.g. 10 feet or more) from the source, even though the spectral components of the direct signal suffer an attenuation of over 40 dB. For example, we

may not be able to get the message in whispered speech by listening to it at a distance of 10 feet or more due to absence of regular epochs. The neural mechanism of human beings seems to have the ability of processing selectively the robust regions around the epochs for extracting the acoustic cues even under degraded conditions. It is the ability of human beings to focus on these microlevel events that may be responsible for perceiving speech information even under severe degradation such as noise, reverberation, presence of other speakers and channel variations.

2.2 Extraction of excitation source information from electroglottography

Electroglottography is a noninvasive method of measuring the vocal fold contact during voicing without affecting speech production. The electroglottograph (EGG) measures the variation in impedance to a very small electrical current between a pair of electrodes placed across the neck as the area of contact of the vocal folds changes during voicing. The demodulated impedance signal is referred to as EGG signal. During voiced speech, the EGG signal exhibits quasiperiodicity according to the frequency of vocal fold vibration.

Fig. 2.1 shows a few stylized glottal cycles of the EGG signal for a voiced speech segment. The glottal cycle of the EGG signal can be divided into four distinct phases: closing phase, closed phase corresponding to the region of maximum contact, opening phase and open phase. This relation between the EGG signal and the area of contact of the vocal folds has been verified using high-speed larynx photography and X-ray flashing imaging [14]. As long as the glottis is open, the impedance measure across the larynx is maximum and almost flat (region - 4 in Fig. 2.1). When the glottis closes, the laryngeal impedance decreases, and the EGG signal shows a steep downward slope (region - 1 in Fig. 2.1). The opening of the glottis, on the other hand, happens much more gradually (region - 3 in Fig. 2.1). Note that some authors invert the EGG signal from that shown in Fig. 2.1.

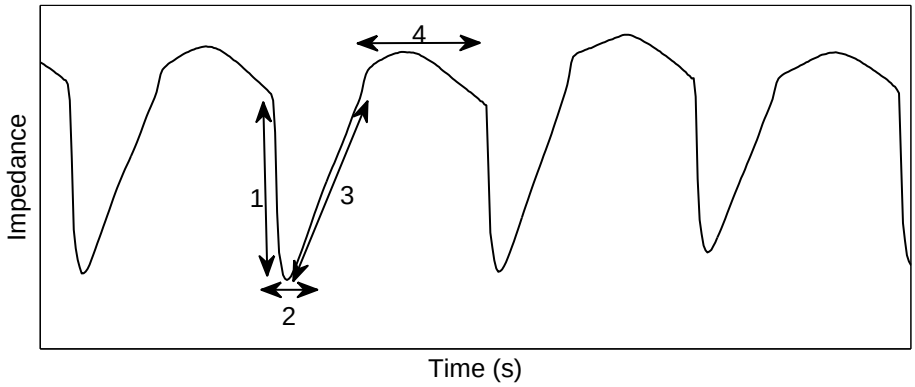


Fig. 2.1: EGG signal for a segment of voiced speech taken from a continuous utterance. Four distinct phases of a glottal cycle in the EGG signal can be identified as (1) closing phase, (2) closed phase with maximum contact, (3) opening phase and (4) open phase.

According to the theory of voice excitation [15][16], the instant of glottal closure is the point of maximum excitation to the vocal-tract system, and it is justified to define it to be the starting point of a pitch period. Although the instant of glottal closure is the most abrupt event, it nevertheless needs a finite amount of time. The definition of the starting point of the period, however, requires identification of a unique point in time, that is less subjected to errors. Though identifying a unique point directly from the speech waveform is not possible, such a feature is well manifested in the EGG signal [17]. Moreover, since the EGG signal measures directly the laryngeal impedance, it is not affected by the ambient noise. The point of inflection during the steep fall of the EGG signal, i.e., the instant of maximum change of the laryngeal impedance is typically selected to represent the instant of glottal closure [18]. Hess and Indefrey defined an epoch to occur at the maximum of the time-derivative of the smoothed EGG signal during a glottal cycle [19]. Huckvale developed an algorithm that identifies epoch locations as the positive-going zero-crossings in the smoothed time-derivative of the EGG signal [20].

Fig. 2.2 shows a segment of voiced speech, its EGG signal and the differenced EGG signal. The locations of sharp negative peaks in the differenced EGG signal denote the instants of glottal closure. The negative peak amplitude of the differenced EGG signal denotes the maximum flow declination rate, which can be hypothesized to be the strength of excitation around the epoch. Notice that, in contrast to the speech signal, the EGG signal is hardly affected by the time-varying vocal-tract system, and the changes in shape

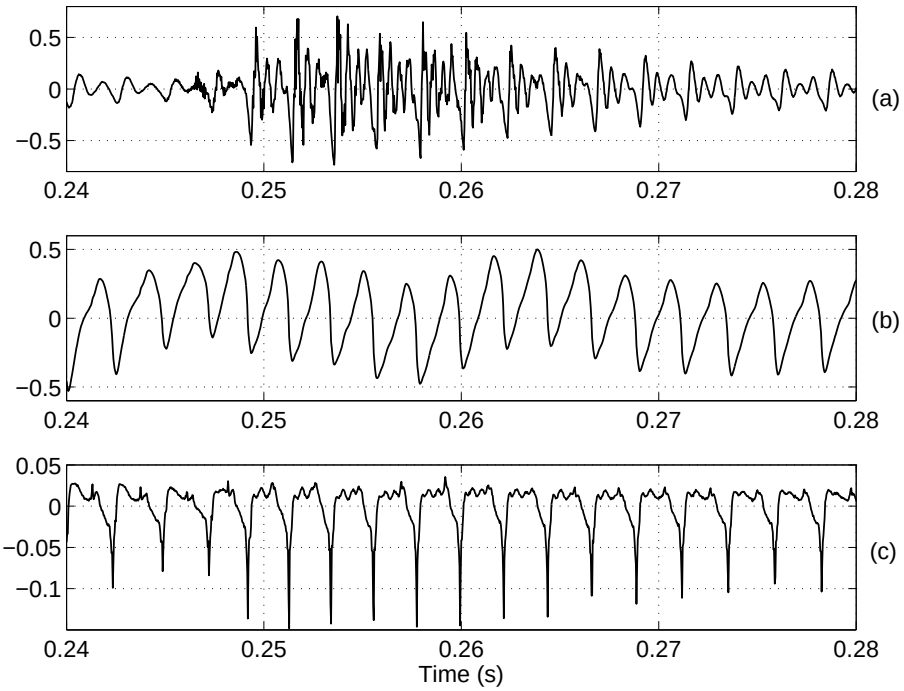


Fig. 2.2: Extraction of epoch locations from differenced EGG signal. (a) A segment of voiced speech taken from a continuous utterance, (b) EGG signal and (c) differenced EGG signal. Locations of the negative peaks in the differenced EGG signal correspond to the instants of glottal closure.

and amplitude are relatively small. Hence, the epoch locations and their strengths can be accurately determined from the EGG signal even in the dynamic regions where the vocal-tract system is not stationary.

Since every glottal cycle is represented by a single pulse, the EGG signal can be used for accurate determination of instantaneous fundamental frequency of the voiced speech segments. In addition, the EGG signal provides the basis for a good voiced-unvoiced discrimination, since the differenced EGG signal is almost zero during unvoiced segments where the glottis is always open. Fig. 2.3(a) and Fig. 2.3(b) show a segment of speech signal and simultaneously recorded EGG signal, respectively. Notice that the differenced EGG signal shown in Fig. 2.3(c) is almost zero in the unvoiced regions. Hence the voiced regions, i.e., the regions of glottal activity, can easily be detected from the differenced EGG signals. The regions of glottal activity for the speech signal shown in Fig. 2.3(a) are marked with dashed lines in Fig. 2.3(b) using the differenced EGG signal. Finally, the instantaneous fundamental frequency computed from time intervals between the negative peaks of differenced EGG signal is shown in Fig. 2.3(d). The finer cycle-to-cycle varia-

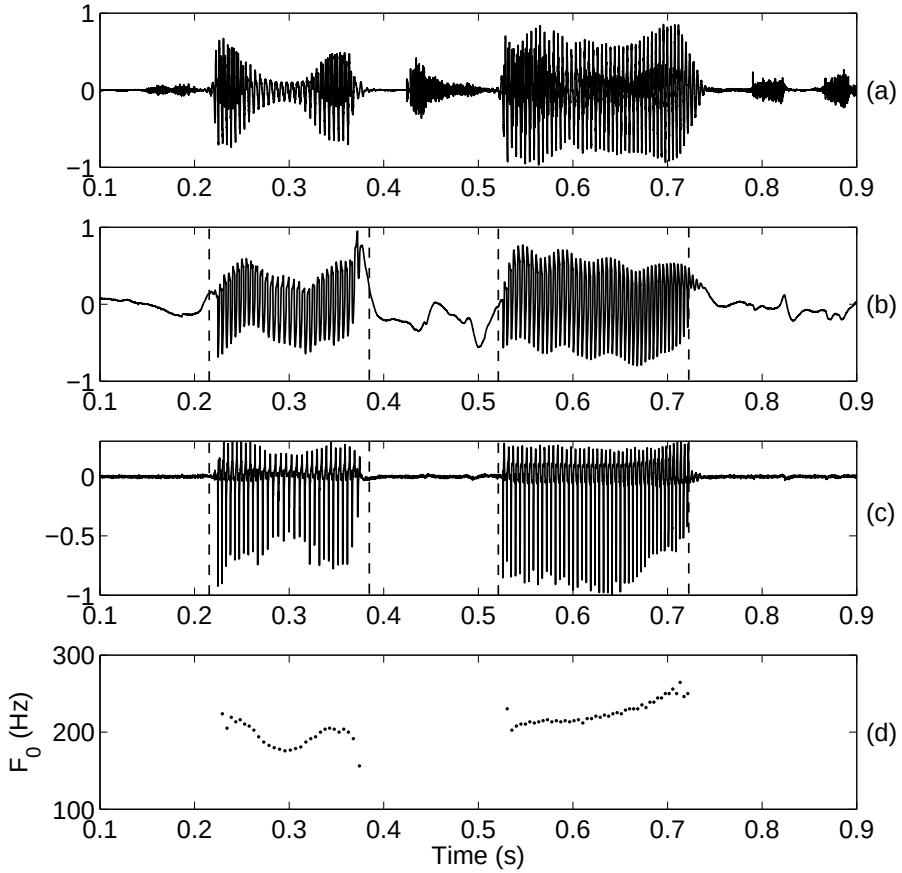


Fig. 2.3: Glottal activity detection and pitch estimation from EGG signal. (a) A segment of speech signal taken from a continuous utterance, (b) EGG signal, (c) differenced EGG signal, and (d) pitch contour obtained by taking the reciprocal of time intervals between locations of successive negative peaks in the differenced EGG signal. The regions of glottal activity are marked with dashed lines.

tions reflected in the pitch contour (Fig. 2.3(d)) are crucial for incorporating naturalness in the synthesized speech signals.

The EGG signal can be effectively used for characterizing the important features of excitation source of the speech signal. The EGG signal can be used for (a) accurate identification of epoch locations, (b) reliable estimation of the strengths of excitation of epochs, (c) glottal activity detection and (d) estimation of instantaneous fundamental frequency of the voiced segments that provides finer cycle-to-cycle variations in pitch. Since the EGG signal is not normally available in practice, there exists strong motivation to develop techniques for extracting these features from the speech signal alone. Several such techniques have been presented in the literature to address these issues independently. The following sections review the techniques for extracting the above mentioned excitation features

directly from the speech signal.

2.3 Overview of epoch extraction methods

Several methods have been proposed for estimating the instants of the glottal closure from a speech signal without the use of the EGG signal. For convenience, we categorize these methods as follows: (a) Methods based on short-time energy of the speech signal; (b) Methods based on the predictability of an all-pole linear predictor; and (c) Methods based on the properties of group-delay, i.e., the negative going zero-crossings of a group delay measure derived from the speech signal. Notice that the methods placed in one category could also belong to another, given another interpretation of the method.

2.3.1 Epoch extraction from linear prediction

Many methods for epoch extraction rely on the discontinuities in a linear model of speech production. An early approach used a predictability measure to detect epochs by finding the maximum of the determinant of the autocovariance matrix of the speech signal [21, 22]. Consider a sequence of observation vectors consisting of segments of the speech signal obtained by advancing a rectangular window of length $p + 1$ samples, one sample further successively. The following data matrix can be formed from the samples $s[n]$ of the speech signal:

$$\mathbf{S} = \begin{pmatrix} s[1] & s[2] & \cdots & s[p] & s[p+1] \\ s[2] & s[3] & \cdots & s[p+1] & s[p+2] \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ s[m] & s[m+1] & \cdots & s[p+m-1] & s[p+m] \end{pmatrix}. \tag{2.1}$$

Let $\mathbf{s}_i$ denote the i^{th} column vector of matrix $\mathbf{S}$. In the absence of excitation, the linear filter model of order p imposes a linear dependence between the vectors $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{p+1}$.

Consequently, the determinant of the matrix $\mathbf{S}^T\mathbf{S}$ as a function of time increases sharply when the speech segment covered by the data matrix $\mathbf{S}$ contains an excitation, and it decreases when the speech segment is excitation free. Therefore the determinant value can be used to detect the location of epochs in the speech signal. This is, in essence, Strube's method for detection of the epochs [22], which is equivalent to computing the product of all squared singular values of the matrix $\mathbf{S}$. This method, however, does not work well for some vowel sounds, particularly when many pulses occur in the determinant computed around the instant of closure. Furthermore, it is computationally expensive. The Cholesky factorization of $\mathbf{S}^T\mathbf{S}$ provides, however, an efficient recursive scheme to perform this computation [22].

The error signal obtained in the LP analysis, referred to as the LP residual, is known to contain information pertaining to epochs. A large value of the LP residual within a pitch period is supposed to indicate the epoch location [23]. However, epoch identification directly from the LP residual is not recommended [22], because the LP residual contains peaks of random polarity around the epochs. Further, since the digital inverse filter does not compensate the phase response of the vocal-tract system exactly, there is an uncertainty in the estimated epoch locations. This makes unambiguous identification of the epochs from the LP residual difficult. Fig. 2.4(b) shows the LP residual derived through a 10^{th} order LP analysis of the speech segment shown in Fig. 2.4(a). The epoch locations can not be unambiguously identified from the LP residual shown in Fig. 2.4(b) because of the occurrence of multiple peaks of either polarity around (0.59 s to 0.6 s) the reference epoch locations shown by the differenced EGG signal in Fig. 2.4(d). A detailed study was made on the determination of the epochs from the LP residual [2], by considering the effect of following factors: (a) the shape of the glottal pulses, (b) inaccurate estimation of formants and bandwidths, (c) phase response of resonances of the vocal-tract system at the instants of significant excitation, and (d) zeros in the vocal-tract system. By taking these factors into account, a method for unambiguous identification of epochs from the LP residual was proposed in [2]. In this work, the amplitude envelope of the analytic signal of the LP residual, referred to as the Hilbert envelope of the LP residual, is used for epoch extraction. Computation of the Hilbert envelope overcomes the effect due to inaccurate phase compensation during inverse filtering. Fig. 2.4(c) shows the Hilbert en-

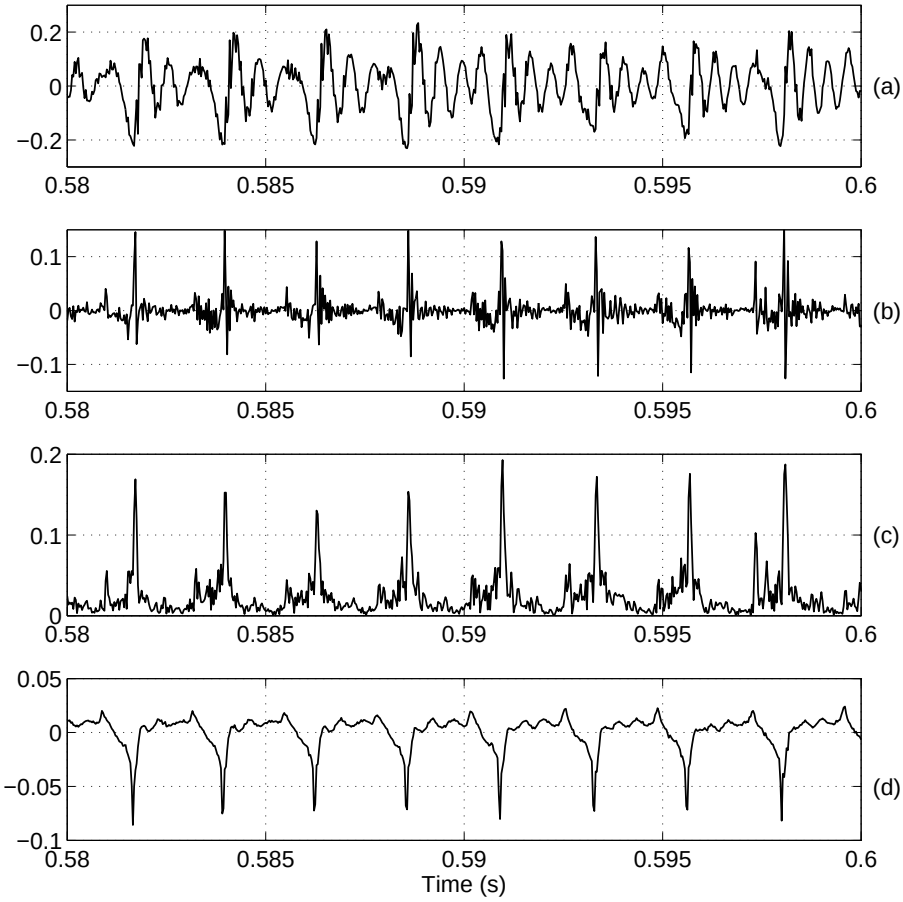


Fig. 2.4: Hilbert envelope of LP residual of the speech signal for epoch extraction. (a) A segment of voiced speech signal taken from a continuous utterance, its (b) LP residual, (c) Hilbert envelope of the LP residual. (d) Differenced EGG signal for observing the reference epoch locations

velope of the LP residual signal shown in Fig. 2.4(b). Unlike the LP residual, the Hilbert envelope shows sharp unambiguous peaks which are in close agreement with the reference epoch locations as shown by the differenced EGG signal in Fig. 2.4(d). Though this method works well on clean signals, the performance of the method degrades under noisy conditions due to the sensitivity of LP analysis to noise in the signal.

Wong, et al., used covariance analysis in the least squares approach for accurately performing the glottal inverse filtering from the acoustic speech waveform [1]. In this work, epochs were detected based on a measure derived from the total energy of the LP residual derived over a sliding window. To ensure that the results are not a function of the absolute system gain (such as recording or voice level), the normalized error which is the ratio of the energy of the LP residual to the energy of the speech signal was used as a measure

of goodness. The glottal closure instant is identified in each glottal cycle as the beginning of the period over which the normalized error stays small. This method was further enhanced by Plumpe, et al., [12] using the observation that the formant modulations are slower in the closed phase region than in the open phase region [24].

One of the difficulties in using the prediction error for epoch extraction is that it often contains effects due to resonances of the vocal-tract system, as the derived inverse filter does not completely suppress the formant frequencies. As a result, the excitation peaks become less prominent in the residual signal, and hence unambiguous detection of the epoch locations becomes harder. In an attempt to overcome this limitation, Cheng, et al., proposed a method based on maximum likelihood theory for epoch determination [25]. In this method, the speech signal was processed to get the maximum-likelihood epoch detection (MLED) signal. The strongest positive pulse in the MLED signal indicates the epoch location within a pitch period. The MLED signal creates not only a strong and sharp epoch pulse, but also a set of weaker pulses that represent the suboptimal epoch candidates within a pitch period. Hence a selection function was derived using the speech signal and its Hilbert transform, that emphasized the contrast between the epoch and the suboptimal pulses. Using the MLED signal and the selection signal with an appropriate threshold, the epochs were detected. The limitation of this method is the choice of the window for deriving the selection function, and also the use of threshold for deciding the epochs.

Kalman filtering has been applied to detect the closed phase regions in voiced speech [26]. The boundary of the closed phase, i.e., the instant of glottal closure and the instant of glottal opening are detected using the logarithm of the determinant of the error covariance matrix of the Kalman filter. This measure assesses the predictability of the speech signal, and is able to detect the glottal closure instants, but the timing accuracy is poor.

2.3.2 Epoch extraction from short-time energy of speech signal

Glottal closure instants can be detected from the energy peaks in waveforms derived directly from the speech signal [27, 28] or from the features in its time-frequency repre-

sentation [29, 30]. In [31], a method based on the composite signal decomposition was proposed for epoch extraction of voiced speech. A superposition of nearly identical waveforms was referred to as a composite signal. The epoch filter proposed in this work, computes the Hilbert envelope of the highpass filtered composite signal to locate the epoch instants. It was shown that the instants of excitation of the vocal-tract could be identified precisely even for continuous speech. However, this method is suitable for analyzing only clean speech.

The Frobenius norm offers a short-term energy estimate of the speech signal. The Frobenius norm computed using a sliding window gives an estimate of energy value at every speech sample. The locations of peaks in the energy signal indicate glottal closure instants. A Frobenius norm approach for detecting the epochs was proposed in [27]. In this work, a new approach based on singular value decomposition (SVD) was proposed. The SVD method calculates the Frobenius norms of signal matrices, and is therefore, computationally efficient. The method was shown to work only for vowel segments. No attempt was made to detect epochs in difficult cases like nasals, voiced consonants and semivowels.

The energy peaks can also be detected in a time-frequency representation of the speech signal. Wavelet transform has been used to represent the speech and to detect the glottal closure instants [30]. Lines of amplitude maxima in the time-frequency plane were identified, and the epochs were determined to correspond to the line carrying the maximum accumulated amplitude within each pitch period. Alternatively, a Cohen's class time-frequency representation of speech was constructed and used to detect the epochs [29]. The epochs were detected as peaks in a spectral density correlator derived from the time-frequency representation.

2.3.3 Epoch extraction from group-delay measures

A method for detecting the epochs in a speech signal using the properties of minimum phase signals and group-delay function was proposed in [3]. The method is based on the fact that the average value of the group-delay function of a signal within an analysis frame

corresponds to the location of the significant excitation. An improved method based on the computation of the group-delay function directly from the speech signal was proposed in [32]. Robustness of the group-delay based method against additive noise and channel distortions was studied in [33]. Four measures of group-delay (average group-delay, zero frequency group-delay, energy weighted group-delay and energy weighted phase) and their use for epoch detection were investigated in [34]. The effect of the length of the analysis window, the tradeoff between the detection rate and the timing error, and the computational cost of evaluating the measures were also examined in detail. It was shown that the energy weighted measures performed better than the other two measures. A dynamic programming projected phase-slope algorithm (DYPSA) for automatic estimation of glottal closure instants in voiced speech was presented in [35, 36]. The candidates for GCI were obtained from the zero-crossings of the phase-slope function derived from the energy weighted group-delay, and were refined by employing a dynamic programming algorithm. It was shown that DYPSA performed better than the existing methods.

Epoch is an instant property. However, in most of the methods discussed above (except the group-delay based methods), the epochs are detected by employing block processing approaches, which result in ambiguity about the precise location of the epochs. Most of the existing methods rely on the LP residual signal derived by inverse filtering the speech signal. Though these methods work well in most cases, they need to deal with the following issues: (a) Selection of parameters (order of LP analysis, length of the window) for deriving the error signal; (b) Dependence of these methods on the energy of the error signal, which in turn depends on the energy of the signal; (c) The accuracy with which the epochs can be resolved decreases as a result of block processing; (d) Setting a threshold value to take a decision on the presence of an epoch; (e) The excitation impulses need not be periodic, though some of these methods exploit periodicity for accurate estimation of epoch locations. In general, it is difficult to detect the epochs in the case of low voiced consonants, nasals and semivowels, breathy voices and female speakers.

2.4 Estimation of strength of excitation of the epoch

We refer to the amplitude of the significant excitation to the vocal-tract system as strength of excitation of the epoch. During the production of voiced speech, the excitation to the vocal-tract system can be considered to be the differentiated glottal flow (also called the effective driving function) [1]. A negative impulse-like peak dominates the waveform of the differentiated glottal flow, at least for normal and loud phonations [37]. This peak occurs at the instant of glottal closure, and serves as significant excitation to the vocal-tract system.

The amplitude of negative peak of the differentiated glottal flow is one of the most important parameters of the excitation source. Several excitation source analysis situations require estimation of the negative peak amplitude of the differentiated glottal flow. The negative peak amplitude of the differentiated glottal flow is closely related to the vocal intensity. Gauffin and Sundberg observed that there is a strong linear correlation between the negative peak amplitude of the differentiated glottal flow and sound pressure level [38]. Alku, et al., defined a parameter called amplitude domain quotient as ratio of the maximum amplitude of the glottal flow and the negative peak amplitude of the differentiated glottal flow [39]. This parameter was used to discriminate between different phonation types [40]. Normalized amplitude quotient, defined as amplitude quotient normalized by the period of vibration [41], was observed to decrease with an increase in vocal intensity [42]. It was shown that the normalized amplitude quotient is more accurate, consistent and robust measure, for parameterization of glottal flow, compared to the closing quotient which indicates the portion of a period where the glottis is closing [43].

In the above mentioned methods, the glottal flow was estimated using inverse filtering of the speech signal [5, 44, 45, 38], where the vocal-tract system is modeled as an all-pole filter [4, 46]. The negative peak amplitude of the differentiated glottal flow was computed from the first derivative of the glottal flow. The mean of negative peak amplitudes of the differentiated glottal flow over a few consecutive glottal cycles was used in the studies [43, 37]. But, as mentioned in earlier sections, the glottal inverse filtering requires modeling of time-varying supralaryngeal vocal-tract system. Errors may occur

whenever the mathematical model assumed for the supralaryngeal vocal-tract system does not accurately reflect the actual acoustic characteristics. Hence, it is desirable to extract the negative peak amplitude of the differentiated glottal flow without characterizing the vocal-tract system.

2.5 Overview of pitch estimation methods

Accurate estimation of the fundamental frequency of voiced speech plays an important role in speech analysis and processing applications. The variation in the fundamental frequency with time contributes to the speech prosody. Estimation of accurate prosody is useful in various applications such as in speaker recognition [47, 48], language identification [49], and even speech recognition [50, 51]. Prosody also reflects the emotion characteristics of a speaker [52]. Prosody is essential for producing high quality speech synthesis, and also for voice conversion. Prosody features were exploited for hypothesizing sentence boundaries [53], for speech segmentation, and for story parsing [54].

There are several algorithms proposed in the literature for estimating the fundamental frequency from speech signals [55, 56, 57]. Depending on the type of processing involved, the algorithms may be classified into three broad categories: (a) algorithms using time domain properties; (b) algorithms using frequency domain properties; and (c) algorithms using statistical methods to aid in the decision making.

2.5.1 Time domain methods

Algorithms based on the properties in the time domain operate directly on the speech signal to estimate the fundamental frequency. Depending on the size of the segment used for processing, the time domain methods can be further categorized into *block-based* methods and *event-based* methods. In the block-based methods, an estimate of the fundamental frequency is obtained for each segment of speech, where it is assumed that the pitch is constant over the segment consisting of several pitch periods. In this case, variation of the fundamental frequency within the segment is not captured. Event-based pitch detectors

locate unique anchor points in each glottal cycle of the speech waveform and the time interval between two successive anchor points is hypothesized as the fundamental period. For event-based pitch detectors, the measurements often made are: peak and valley measurements, zero-crossing measurements, epoch locations or pitch markings.

Among the time domain block-based methods, the autocorrelation approaches are popular for their simplicity. A correlation function is a measure of the degree of similarity between two signals [58]. The autocorrelation measures how well the input signal matches with a time-shifted version of itself. The autocorrelation sequence $r_{ss}[\tau]$ of a speech segment $s[n]$ is given by

$$r_{ss}[\tau] = \sum_{n=0}^{N-1-|\tau|} s[n]s[n+\tau], \qquad \tau = 0, \pm 1, \pm 2, \ldots, \pm N-1, \tag{2.2}$$

where τ is the time shift. For a periodic signal, its autocorrelation function is also periodic. Due to periodic nature of the voiced speech, the first peak (also called the pitch peak) after the center peak in the autocorrelation function indicates the fundamental period (T_0) of the signal. The reciprocal $F_0 = \frac{1}{T_0}$ is the fundamental frequency. There are several reasons for the success of the autocorrelation methods for pitch detection [59]. The autocorrelation computation is made directly on the speech signal, and involves a straightforward computation. Although high processing rate is required, the autocorrelation computation is amenable to digital hardware implementation, generally requiring only a single multiplier and an accumulator as computational elements. Finally, the autocorrelation computation is largely phase insensitive. Thus, it is a good method to use to detect the pitch of speech which has been transmitted over a telephone channel, or has suffered some degree of phase distortion during transmission.

Although the autocorrelation-based pitch detector has some advantages, there are several problems associated with the use of this method. Although the autocorrelation function of a segment of voiced speech generally displays a fairly prominent peak at the pitch period, peaks due to formant structure of the signal are also often present. Thus, one problem is to decide which of the several peaks in the autocorrelation function corresponds to the pitch period. Another problem with the autocorrelation computation is the

required use of a window for computing the short-time autocorrelation function. The use of a window for analysis leads to some difficulties. First, there is the problem of choosing an appropriate window. Second, there is the problem that, no matter which window is selected, the effect of the window is to taper the autocorrelation function as the autocorrelation index increases. This effect tends to compound the difficulties mentioned above in which the formant peaks in the autocorrelation function (which occur at lower indices than the pitch period peak) tend to be of greater magnitude than the peak due to the pitch period. A final difficulty with the autocorrelation computation is the problem of choosing an appropriate analysis frame (window) size. The ideal analysis frame should contain at least 2 to 3 complete pitch periods. Thus, for high pitch speakers the analysis frame should be short (5-20 ms), whereas for low pitched speakers it should be long (20-50 ms).

A wide variety of preprocessing techniques have been proposed in the literature to address the above mentioned issues. To partially eliminate the effects of the higher formant structure on the autocorrelation function, most methods use a sharp cutoff low-pass filter with cutoff around 1000 Hz. This will, in general, preserve a sufficient number of pitch harmonics for accurate pitch detection, but eliminates the second and higher formants. In addition to linear filtering to remove the formant structure, a wide variety of methods have been proposed for directly or indirectly flattening the short-time spectrum of the speech signal to remove the effects of the first formant [60, 61, 62]. Included among these techniques are center clipping and spectral equalization by filter bank methods [61], inverse filtering using linear prediction methods [62], and spectral flattening by a combination of center and peak clipping methods [63]. Rabiner presented an investigation of the properties of a class of nonlinearities applied to the speech signal prior to autocorrelation analysis with the purpose of spectrally flattening the signal [59]. A solution to the problem of choosing an analysis frame size which adapts to the estimated average pitch of the speaker is also presented in [59] .

In the computation of the autocorrelation function, fewer samples are included as the lag increases. This effect can be seen as the roll-off of the autocorrelation values for higher lags. The values of the autocorrelation function at higher lags are important, especially for low-pitched male voices. For a 50 Hz pitch, the lag between successive pitch pulses is

200 samples at a sampling frequency of 10 kHz. To overcome the roll-off caused by the windowing, Boersma suggested dividing the autocorrelation sequence of the windowed signal with the autocorrelation sequence of the window [64]. This correction does not let the resulting correlation sequence taper to zero as the lag increases, which helps in accurate identification of the peak corresponding to the fundamental period.

To overcome this limitation of the autocorrelation function, a crosscorrelation function which operates on two different data windows is also proposed [65]. Direct computation of the crosscorrelation function is influenced by the energy of the speech segments. Rapid changes in energy is common at voicing onsets and voicing endings. In order to make the crosscorrelation function independent of the energy of the speech segments, the crosscorrelation values are compensated based on the energy in the sliding window. The resulting normalized crosscorrelation function is given by

$$c[\tau] = \frac{\sum_{n=0}^{N-1} s[n]s[n+\tau]}{\sqrt{\sum_{n=0}^{N-1} s^2[n] \sum_{n=0}^{N-1} s^2[n+\tau]}} \qquad \tau = 0, \pm 1, \pm 2, \ldots, \pm N-1. \tag{2.3}$$

As the number of samples involved in the computation of $c[\tau]$ is constant, this estimate is unbiased, and has lower variance than that of the autocorrelation. Unlike the autocorrelation method, the window length could be lower than the pitch period, so that the assumption of stationarity is more valid, which results in better time resolution. While the pitch trackers based on the normalized crosscorrelation typically perform better than those based on the autocorrelation, they also require more computation.

One drawback of the correlation-based methods is the need for multiplication, which is relatively expensive for implementation, especially in those processors with limited functionality. To overcome this problem, the average magnitude difference function (AMDF) was proposed [66]. This function is defined by

$$d[\tau] = \frac{1}{N} \sum_{n=0}^{N-1} |s[n] - s[n-\tau]|, \qquad \tau = 0, \pm 1, \pm 2, \ldots \pm N-1. \tag{2.4}$$

For short segments of voiced speech, it is reasonable to expect that $d[\tau]$ is small for

$\tau = 0, \pm T_0, \pm 2T_0, \ldots$, with T_0 being the fundamental period of the signal. Thus by computing the AMDF for the lag range of interest, the fundamental period can be estimated by locating the lag index associated with the minimum magnitude difference. Notice that multiplication operation is not involved in implementation of the AMDF method.

The methods discussed so far can only find integer valued fundamental periods. That is, the resultant fundamental period values are multiples of the sampling period. For a speech signal sampled at 8 kHz, the fundamental period can only be computed with a precision of multiples of 0.125 ms. In many applications, higher resolution is necessary to achieve good performance. In fact, the fundamental period of the original continuous-time (before sampling) signal is a real number. Thus, integer periods are only approximations, and may introduce errors that might have negative impact on the performance of the system. Multirate signal processing techniques can be used to improve the resolution beyond the limits set by the fixed sampling rate. Interpolation, for instance, is a widely used method where the actual sampling rate is increased. Medan, et al., proposed a super-resolution pitch determination algorithm which is based on a linear interpolation technique [67].

2.5.2 Event-based Methods

The basic assumption behind the event-based methods is that, if a quasiperiodic speech signal is suitably processed to minimize the effects of the formant structure and highlight certain anchor point in each glottal cycle, then simple time-domain measurements provide a good estimate of the period. Gold and Rabiner proposed a pitch detection method using parallel processing of events derived from the speech signal [68]. In this approach, the speech signal is first low-pass filtered to a bandwidth of 900 Hz. Then a series of measurements are made on the peaks and valleys of the low-pass filtered signal to give six separate functions. Each of these six pitch functions is processed by an elementary pitch period estimator, giving six separate estimates of the pitch period. The six pitch estimates are then combined by a sophisticated decision algorithm which determines the pitch period. As a byproduct of this algorithm, a voiced-unvoiced decision is obtained

based on the degree of agreement among the six pitch detectors.

Miller proposed a data reduction pitch detector which places pitch markers directly on a low-pass filtered (0-900 Hz) speech signal, and thus is a pitch-synchronous pitch detector [69]. To obtain appropriate pitch markers, the data reduction method first detects excursion cycles in the waveform based on intervals between major zero-crossings. The remainder of the algorithm tries to isolate and identify principal excursion cycles, i.e., those which correspond to true pitch periods. This is accomplished through a series of steps using energy measurements, and logic based on permissible pitch periods and anticipated syllabic rate changes of pitch. Finally, an error correction procedure is used to provide a reasonable measure of continuity in the pitch markers.

Wavelet transforms have been used to determine the pitch period by locating the instants at which glottis closes (called events), and then measuring the time interval between two such events [70, 71, 72, 73, 74]. In [70], wavelet transforms are used for pitch period estimation based on the assumption that the glottal closure causes sharp discontinuities in the derivative of the airflow. The transients in the speech signal caused by the glottal closure result in maxima in the scales of the wavelet transform around the instant of discontinuity. In this method, one needs to detect the correlated maxima across these scales by heuristic algorithms, which is often prone to error especially in the case of noisy signals. To overcome this, an optimization scheme is proposed in the wavelet framework using a multipulse excitation model for the speech signal, and the pitch period is estimated as a result of this optimization [75].

2.5.3 Frequency domain methods

Algorithms based on the properties in the frequency domain assume that if the signal is periodic in the time domain, then the frequency spectrum of the signal contains a sequence of impulses at the fundamental frequency and its harmonics. Then simple measurements can be made on the frequency spectrum of the signal, or on a nonlinearly transformed version of it (as in the cepstral pitch detector [76]) to estimate the fundamental frequency of the signal.

The cepstrum method for extraction of pitch utilizes the frequency domain properties of speech signals [76, 77]. In the short-time spectrum of a given voiced frame, the information about the vocal-tract system appears as a slowly varying component, and the information of the excitation source is in rapidly varying component. These two components may be separated by considering the logarithm of the spectrum, and then applying the inverse Fourier transform to obtain the cepstrum. This operation transforms the information in the frequency domain to the cepstral domain, which has a strong peak at the average fundamental period of the voiced speech segment being analyzed.

Subharmonic summation (SHS), proposed by Hermes [77], performs pitch analysis based on a spectral compression model. Several methods have been proposed for estimating the harmonic frequencies based on the instantaneous frequency of the speech signal [78, 79, 80]. In this approach, the speech signal is decomposed into the harmonic components using a set of bandpass filters, each of whose center frequencies changes with time in such a way that it tracks the instantaneous frequency of its output [78]. As a result, the outputs of the band-pass filters become the harmonic components, and the instantaneous frequencies of the harmonics are accurately estimated. The pitch extraction is accomplished by selecting the correct fundamental frequency out of the harmonic frequencies.

Nakatani and Irino proposed a method for fundamental frequency estimation by selecting the dominant harmonic components of the speech signal [81, 82]. In this work, degree of dominance and dominance spectrum are defined based on instantaneous frequencies. The degree of dominance allows to evaluate the magnitudes of the individual harmonic components of the speech signal relative to background noise. The selection of the dominant harmonic components results in reducing the influence of spectral distortion. The fundamental frequency is more accurately estimated from reliable harmonic components which are easy to select given the dominance spectra.

2.5.4 Statistical methods

The problem of automatic estimation of fundamental frequency can be considered, in some sense, a statistical one. Each input frame is classified into one of a number of groups, representing the fundamental frequency estimator of the signal. Wise, et al., proposed a method for estimating the fundamental period of voiced speech sounds based on a maximum likelihood formulation [83, 84, 85]. In this work, the problem is formulated as that of estimating an unknown periodic signal in white Gaussian noise of unknown intensity. An objective function based on the probability that the signal is periodic with a period of T_0 was derived, and was maximized over T_0 to estimate the fundamental period of the signal. This method is capable of providing finer resolution than one sampling period, and is shown to perform better in the presence of noise than the cepstrum method.

Joseph, et al., proposed a statistical method for pitch tracking, assuming a harmonic model of the speech signal [86]. The harmonic model could be regarded as special case of a sinusoidal speech model, where all sinusoidal components are assumed to be harmonically related, i.e., the frequencies of the sinusoids are at integer multiples of the fundamental frequency. This assumption reduces the number of parameters in the model and achieves more accurate estimates of pitch than the sinusoidal model. Assuming Markovian dynamics, maximum *a-posteriori* probability tracking of the time-varying harmonic signal is performed without prior knowledge of noise variance.

Most of the existing methods for extraction of the fundamental frequency assume periodicity in successive glottal cycles, and thus they work well for clean speech. The performance of these methods is severely affected if the speech signal is degraded due to the noise or due to other distortions. This is because the pitch peak in the autocorrelation function or cepstrum may not be prominent or unambiguous. In fact, during the production of voiced speech, the vocal-tract system is excited by a sequence of impulse-like signals caused by the rapid closure of the glottis in each cycle. There is no guarantee that the physical system, especially due to the time-varying vocal-tract shape, produces similar speech signals for each excitation. Moreover, there is also no guarantee that the impulses occur in the sequence with any strict regularity. In view of this, it is better to

extract the interval between successive impulses, and take the reciprocal of that interval as the instantaneous fundamental frequency.

2.6 Processing multimicrophone data

2.6.1 Time-delay estimation

The problem of time-delay estimation has been handled traditionally by exploiting spectral characteristics of speech signals [87, 88]. Three broad strategies are used in these studies [89]: (a) Steered response power of a beamformer; (b) High resolution spectrum estimation; and (c) 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 [90].

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 signals over the intervals in which noise and speakers are assumed to be stationary, and their estimation parameters are assumed to be fixed [91]. 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 estimation of time differences of arrival (TDOA) are more suitable for time-delay estimation than the previous two approaches [89]. For accurate estimation of time-delays, weighted generalized crosscorrelation (GCC) method is often used [92].

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 is made more robust by deemphasizing the frequency dependent weightings [93]. 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 the 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 [94, 95]. Moreover, this approach is not suitable for estimation of 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.

2.6.2 Multispeaker speech processing

In a multispeaker environment 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 [96, 97, 98, 99, 100]. 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 speaker. In the multichannel case, signals from multiple microphones are processed to enhance speech of the desired speaker. This approach is inspired by the binaural processing of humans. In the multichannel case, speech of two or more speakers may be enhanced using signals from multiple microphones.

Several pitch-based algorithms have been proposed for cochannel speaker separation [97, 98, 99]. 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 pitch harmonics of the desired speaker, most of the energy of the spectral samples would correspond to that speaker. After obtaining the harmonic amplitudes, the time-domain waveform is reproduced using the synthesis algorithm. Harmonic magnitude suppression technique for speech separation was proposed in [101]. 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 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 [102]. 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., have proposed a method for cochannel speaker separation, termed as harmonic enhancement and suppression [99]. The pitch of the stronger speaker was estimated first, and it was then used for recovering his/her harmonics and formants. The weaker speaker information was obtained after suppressing the harmonics and formants information of the stronger speaker 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 [96]. 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). The 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 separation and deconvolution of signals are discussed in [103]. A

method for multichannel signal separation using a dynamical recurrent network is proposed in [104, 105] . Estimation of speech embedded in reverberant environment with multiple sources of noised is proposed in [100, 106], . 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 signal with low energy.

In most of the existing methods, the knowledge of pitch is used for deriving the information related to each speaker. However, reliable estimation of pitch in multispeaker environment is a difficult task.

2.7 Manner of articulation of stop consonants

In Indian languages, there are stop consonants articulated at a given point in the vocal-tract that are minimally distinguished from one another by the nature of the excitation source, referred to as manner of articulation. The manner of articulation of stop consonants is described by the voiced or unvoiced nature of the closure event, and the presence or absence of the aspiration event, leading to four different manners of articulation. The manner of articulation is mainly dictated by the relative timings of onset of vocal fold vibration and instant of closure release. For unvoiced unaspirated stop consonants, the vocal fold vibration begins almost immediately after the closure release. Whereas, for voiced unaspirated stop consonants the vocal fold vibration begins during the closure duration. For unvoiced aspirated stop consonants, the onset of vocal fold vibration is delayed after the instant of closure release to produce aspiration.

Voice onset time (VOT) [107], defined as the interval between the instant of closure release to the onset of vocal fold vibration, is one of the important features used to analyze the manner of articulation of stop consonants. Most of the commonly used methods for measuring the voice onset time are based on the onset of periodicity in the acoustic waveform, possibly supplemented by spectrographic analysis [108] or direct measurements of

airflow [109]. Peterson and Lehiste identified the onset of voicing as the point at which stable striations first become visible in the frequency region of first formant of a wide-band spectrogram [110]. In contrast, Klatt made the measurements of voicing onset at the onset of visible energy in higher formants on the grounds that voicing onset may not always be visible in the first formant region [111]. Liskar and Abramson determined the onset of voicing according to the time of the first vertical striations visible in a wideband spectrogram, presumably irrespective of the frequency (or formant) at which they first appeared [112]. In addition to these spectrographic measures, it is also possible to measure the onset of voicing as the onset of energy visible in the *voice-bar*, i.e., the region of lowest frequency energy in a wide-band spectrogram corresponding to the fundamental frequency, typically found below the first formant [113]. Lieberman and Blumstein measured the voicing onset directly from the acoustic waveform itself, in terms of onset of the first clearly seen periodic pattern in the acoustic signal [114].

One of the issues in spectrographic methods is the choice of acoustic landmark for measuring the VOT. The vertical striations due to voicing onset do not reflect across all the formants at the same instant of time. For instance, the onset of voicing in aspirated sounds appears earlier at the higher formants than at the first formant. The effect of block processing in the spectrogrphic analysis may limit the time-resolution of observation of these features. Since the manner of articulation of stop consonants depends on nature of excitation source, features derived from excitation source may provide a better analysis.

2.8 Summary

In this chapter, we have reviewed some existing methods for extracting and processing the excitation source information in the speech signal. In particular, algorithms for extraction of epochs and estimation of pitch from speech signals are reviewed. Extraction of excitation source information requires suppressing the vocal-tract information from the speech signals. Most of the epoch extraction methods rely on modeling the vocal-tract response using LP analysis, and then inverse filtering the speech signal obtain LP residual. The performance of these methods depends critically on the accuracy of LP analysis in

modeling the vocal-tract response, order of the LP analysis, and nature and quality of the speech signal. Because of the time-varying nature of the vocal-tract, the existing methods for epoch extraction, invariably, employ block processing that introduces effects of windowing. These factors may result in ambiguity about the precise location of the epochs.

The goal of this thesis to demonstrate the significance of impulse-like nature of excitation in extracting the epochs, their strengths of excitation and fundamental frequency of voiced speech. In contrast to the existing approaches, the methods proposed in this work enhance the source information by exploiting the impulse-like nature of excitation rather than attempting to model the vocal-tract response and then suppressing it. Since the proposed methods do not depend on modeling the vocal-tract response, they can be applied on speech data of any length without using block processing.

Chapter 3

Epoch Extraction

In this chapter, we present a new method for epoch extraction that is based on the assumption that the major source of excitation of the vocal-tract system is due to a sequence of impulse-like events in the glottal vibration. The impulse excitation to the system results in a discontinuity in the output signal. We propose a novel approach to detect the location of the discontinuity in the output signal by confining the analysis to a narrowband around a single frequency. In Section 3.1, we discuss the basic principle of the proposed method, and illustrate the principle for a few representative cases of synthetic excitation signals. In Section 3.2, we discuss the issues involved in applying the method directly on speech data. In Section 3.3, we propose a method to extract epochs from the speech signal. In Section 3.4, the performance of the proposed method in terms of identification accuracy is given, and the results are compared with three existing methods for epoch extraction. In Section 3.5, the performance of the proposed method is evaluated for different types of degradations, and the results are compared with some existing methods. Finally, in Section 3.6 we summarize the contributions of this chapter, and discuss some limitations of the proposed method which prompt further investigation.

3.1 Basis for the proposed method of epoch extraction

Speech is produced by exciting the time-varying vocal-tract system by one or more of the following three types of excitation: (a) glottal vibration, (b) frication, and (c) burst. The primary mode of excitation is due to glottal vibration. While the excitation is present throughout the production process, it is considered significant (especially during glottal vibration) only when there is large energy in short-time interval, i.e., when it is impulse-like. This impulse-like characteristic is usually exhibited around the instant of glottal closure during each glottal cycle. The presence of impulse-like characteristic suggests that the excitation can be approximated as a sequence of impulses. This assumption on the excitation of the vocal-tract system suggests a new approach for processing the speech signal as discussed in this section.

All physical systems are inertial in nature. The inertial systems respond when excited by an external source. The excitation to an inertial system can be any of the following four types:

(a) *Excitation impulse is not in the observed interval of the signal - Sinusoidal generator:* Output signal is the response of a passive inertial system for an impulse, and the impulses themselves are not present in the observed intervals of the signal.

(b) *Sinusoidal excitation:* Sinusoidal excitation can be viewed as impulse excitation in the frequency domain. Hence, a sinusoidal excitation to an inertial system selects the corresponding frequency component from the transfer function of the system. Though sinusoidal excitation is widely used to analyze synthetic systems, it is not commonly found in physical systems.

(c) *Random excitation:* Random excitation can be interpreted as impulse excitation of arbitrary amplitude at every instant of time. Since impulse excitations are present over all the instants of time, it is difficult to observe them from the output of the system. Random excitation does not possess impulse-like nature either in the time-domain or in the frequency-domain, and hence the impulses cannot be perceived.

(d) *Sequence of impulses as excitation:* In this case, the signals are generated by a

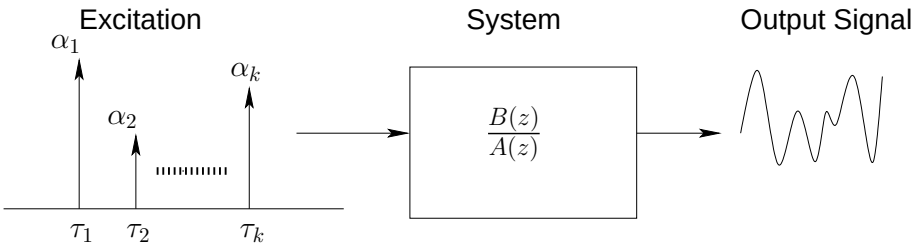


Fig. 3.1: An inertial system excited with a sequence of impulses.

passive inertial system with a fixed sequence of (periodic and/or aperiodic) impulses as excitation. The time instants of impulses may not be observed from the output of the system, but they can be perceived. If the sequence of impulses is periodic in the time-domain, then it corresponds to a periodic sequence of impulses in the frequency-domain as well.

Consider a physical system excited by a sequence of impulses of varying strengths, as shown in Fig. 3.1. One of the challenges in the field of signal processing is to detect the time instants (τ_k) of the impulses and their corresponding strengths (α_k) from the output signal. In a natural scenario like speech production, the characteristics of the system vary with time, and are unknown. Hence the signal processing problem can be viewed as a blind deconvolution, where neither the system response nor the excitation source is known. In this work, we attempt to detect the time instants of excitation (epochs) of the vocal-tract system.

Consider a unit impulse in the time domain. It has all the frequencies equally well represented in the frequency domain. When an inertial system is excited by an impulse-like excitation, the effect of the excitation spreads uniformly in the frequency domain, and is modulated by the time-varying transfer function of the system. The information about the time instants of occurrence of the excitation impulses reflects as discontinuities in the time domain. It may be difficult to observe these discontinuities directly from the signal because of the time-varying response of the system. The effect of the discontinuities can be highlighted by filtering the output signal through a narrowband filter centered around a frequency. The output of the narrowband filter predominantly contains a single frequency component, and as a result, the discontinuities due to the excitation impulses will get manifested as a deviation from the center frequency. The time instants of the

discontinuities can be derived by computing the instantaneous frequency of the filtered output [115]. A tutorial review on the instantaneous frequency and its interpretation is given in [116]. It has been previously observed that isolated narrow spikes in the instantaneous frequency of the bandpass filtered output [117] are attributed to either the valleys in the amplitude envelope or the onset of a new pitch pulse. However, no previous work explored the feasibility of this type of observation for epoch extraction.

3.1.1 Computation of instantaneous frequency

The instantaneous frequency of a real signal $s(t)$ is defined as the time derivative of the unwrapped phase of the complex analytic signal derived from $s(t)$ [115]. The complex analytic signal corresponding to a real signal $s(t)$ is given by

$$s_a(t) = s(t) + js_h(t) \tag{3.1}$$

where $s_h(t)$ is the Hilbert transform of the real signal $s(t)$, and is given by

$$s_h(t) = \mathrm{IFT}(S_h(\omega)), \tag{3.2}$$

where IFT denotes the inverse Fourier transform, and $S_h(\omega)$ is given by

$$S_h(\omega) = \begin{cases} +jS(\omega), & \omega < 0 \\ -jS(\omega), & \omega > 0. \end{cases} \tag{3.3}$$

The analytic signal thus derived contains only positive frequency components. The analytic signal $s_a(t)$ can be rewritten as

$$s_a(t) = |s_a(t)|e^{j\phi(t)}, \tag{3.4}$$

where

$$|s_a(t)| = \sqrt{s^2(t) + s_h^2(t)} \tag{3.5}$$

is called the amplitude envelope, and

$$\phi(t) = \arctan\left(\frac{s_h(t)}{s(t)}\right) \tag{3.6}$$

is called the instantaneous phase. Direct computation of the phase $\phi(t)$ from (3.6) suffers from the problem of phase wrapping, i.e., $\phi(t)$ is constrained to an interval $(-\pi, \pi]$ or $[0, 2\pi)$. Hence, the instantaneous frequency can not be computed by explicit differentiation of phase $\phi(t)$ without first performing the complex task of unwrapping the phase in time. The instantaneous frequency can be computed directly from the signal, without going through the process of phase unwrapping, by exploiting the Fourier transform relations. Taking logarithm on both sides of (3.4), and differentiating with respect to time t, we have

$$\begin{aligned} \log s_a(t) &= \log|s_a(t)| + j\phi(t) \\ \frac{s_a'(t)}{s_a(t)} &= \frac{d}{dt}\log|s_a(t)| + j\phi'(t) \end{aligned} \tag{3.7}$$

where the superscript $'$ denotes the derivative operator, and $\phi'(t)$ is the instantaneous frequency. That is

$$\phi'(t) = -\Im\left(\frac{s_a'(t)}{s_a(t)}\right), \tag{3.8}$$

where $\Im(.)$ denotes the imaginary part. $s_a'(t)$ can be computed by using the Fourier transform relations. The analytic signal $s_a(t)$ can be synthesized from its frequency domain representation through the inverse Fourier transform as follows:

$$s_a(t) = \frac{1}{2\pi}\int_0^\infty S_a(\omega)e^{j\omega t}d\omega, \tag{3.9}$$

where $S_a(\omega)$ is the Fourier transform of the analytic signal $s_a(t)$, and is zero for negative frequencies. Differentiating both sides of (3.9) with respect to time t, we have

$$\begin{aligned} s_a'(t) &= \frac{1}{2\pi}\int_0^\infty S_a(\omega)e^{j\omega t}(j\omega)d\omega \\ &= j\left(\frac{1}{2\pi}\int_0^\infty (\omega S_a(\omega))e^{j\omega t}d\omega\right) \\ &= j\text{IFT}\left(\omega S_a(\omega)\right). \end{aligned} \tag{3.10}$$

The instantaneous frequency $\phi'(t)$ can be obtained from (3.7) and (3.10) as

$$\phi'(t) = \Re\left(\frac{\mathrm{IFT}(\omega S_a(\omega))}{\mathrm{IFT}(S_a(\omega))}\right), \tag{3.11}$$

where $\Re(.)$ denotes the real part. Computation of the instantaneous frequency given in (3.11) is implemented in the discrete domain as follows:

$$\phi'[n] = \frac{2\pi}{N}\Re\left(\frac{\mathrm{IDFT}(kS_a[k])}{\mathrm{IDFT}(S_a[k])}\right). \tag{3.12}$$

Here IDFT denotes the inverse discrete Fourier transform, and N is the total number of samples in the signal.

The instantaneous frequency may be interpreted as the frequency of a sinusoid which locally fits the signal under analysis. However, it has a physical interpretation only for monocomponent signals, where there is only one frequency or a narrow range of frequencies varying as a function of time. In this case, the instantaneous frequency can be interpreted as deviation of frequency of the signal from the monotone at every instant of time. The notion of a single-valued instantaneous frequency becomes meaningless for multicomponent (multiple frequency sinusoids) signals. The multicomponent signal has to be dispersed into its components for further analysis.

We propose to use a resonator to filter out from a signal a monocomponent centered around a single frequency for further analysis. A resonator is a second-order infinite impulse response (IIR) filter with a pair of complex conjugate poles in the z-plane [58]. The impulse response of a resonator is given by [58]

$$h[n] = \frac{r^n \sin[(n+1)\omega_0]}{\sin(\omega_0)}\, u[n] \tag{3.13}$$

where ω_0 determines the normalized center frequency (in radians) of the filter, radius r determines the bandwidth and $u[n]$ is the unit step function. A small value of r ($r << 1$) corresponds to a wider bandwidth, allowing a large range of frequencies, whereas a value of $r = 1$ corresponds to zero bandwidth. A value of r in the range 0.98 to 1 can be used for implementing a narrowband filter. An IIR filter was preferred over a finite

impulse response (FIR) filter, because an FIR filter requires longer filter length to realize the narrowband. Since an ideal excitation impulse is a point property in the time domain, the FIR filter smears the characteristic of the impulse, and as a result it becomes difficult to accurately extract the instant of the excitation impulse. Hence a resonator with narrow bandwidth (corresponding to a radius $r = 0.999$) was chosen to realize the narrowband filter. Ideal resonator ($r = 1$) was not used in order to avoid saturation of the filter output.

3.1.2 Illustration of instantaneous frequency for synthetic signals

When a multicomponent signal is filtered through a resonator centered around a frequency (ω_0), the output signal predominantly contains the ω_0 frequency component. Any deviation from ω_0 in frequency of the filtered output can be attributed to the impulse-like characteristics present in the multicomponent signal. In general, the analytic signal corresponding to the filtered output can be expressed as

$$s_a(t) = |s_a(t)|e^{j(\omega_0 t + \theta(t))}. \tag{3.14}$$

Hence the instantaneous phase of the filtered output (predominantly monocomponent) is given by

$$\phi(t) = \omega_0 t + \theta(t). \tag{3.15}$$

where $\omega_0 t$ is the linear phase of the resonator, and $\theta(t)$ is the time-varying phase induced in the filtered output by the multicomponent signal. The instantaneous frequency of the filtered output is given by the time derivative of the instantaneous phase, as follows:

$$\phi'(t) = \omega_0 + \theta'(t). \tag{3.16}$$

When a resonator is excited with a single impulse, it follows through its natural oscillations resulting in a signal with linear phase $\omega_0 t$ (Fig. 3.2(a), Fig. 3.2(b) and Fig. 3.2(c)). For this case $\theta(t)$ will be zero. On the other hand, when the resonator is excited by sequence of impulses, the response of the resonator due to excitation impulses at different time instants gets superposed to form a composite signal, as shown in Fig. 3.2(d). The

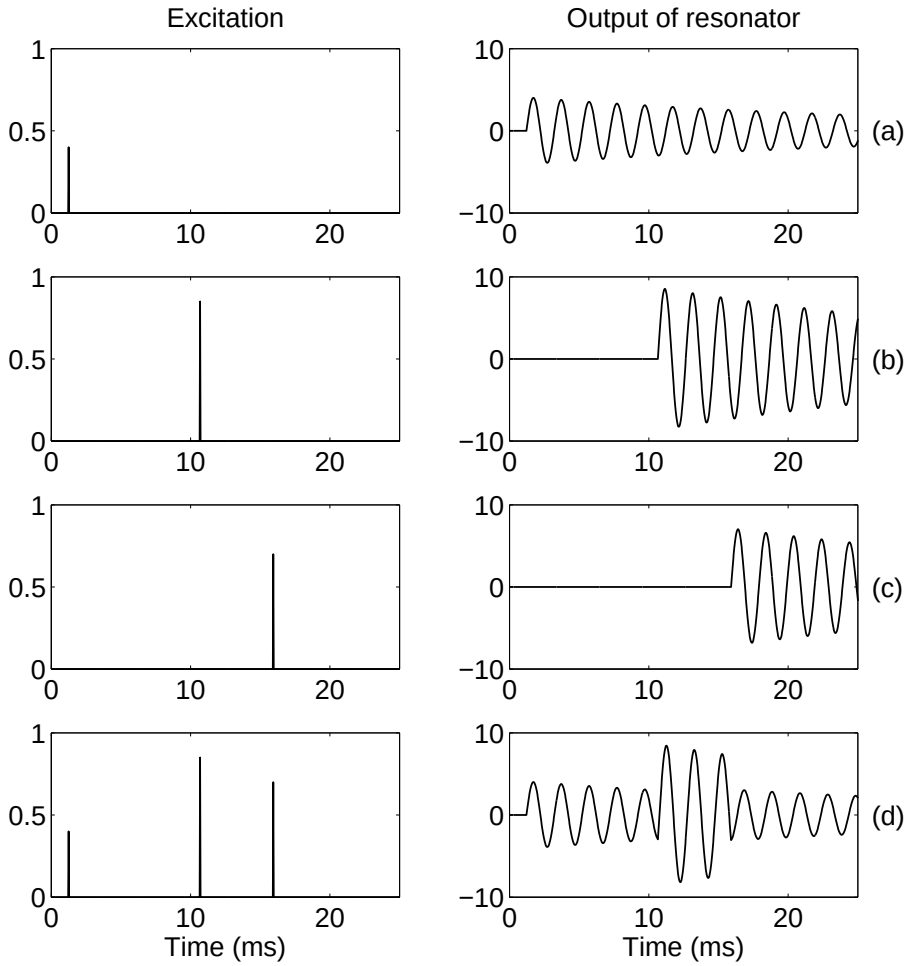


Fig. 3.2: Illustration of superposition of responses of a resonator for impulse excitations at different time instants. (a), (b) and (c) shows response of the resonator for individual impulses and (d) shows the superposed response.

composite signal deviates from the natural oscillations of the resonator at the instants of excitation impulses. The deviations from the natural oscillations reflect in the phase of the signal as deviation $\theta(t)$ from the linear phase $\omega_0 t$. The deviations from the linear phase can be better observed from the instantaneous frequency, $\phi'(t) = \omega_0 + \theta'(t)$. Fig. 3.3(a) shows a multicomponent signal in the form of a periodic impulse sequence. The signal filtered by 500 Hz resonator, and the instantaneous frequency plots of the filtered signal are also given in Fig. 3.3(b) and Fig. 3.3(c), respectively. At the instants of impulse locations, the instantaneous frequency deviates significantly from the normalized center frequency $\omega_0 = 2\pi f / f_s$, where f is the frequency of the resonator, and f_s is the sampling frequency. For a resonator frequency f= 500 Hz and sampling frequency f_s=8000, the instantaneous frequency (around $\omega_0 = 0.3927$) shows sharp peaks at the instants of excitation. Note that

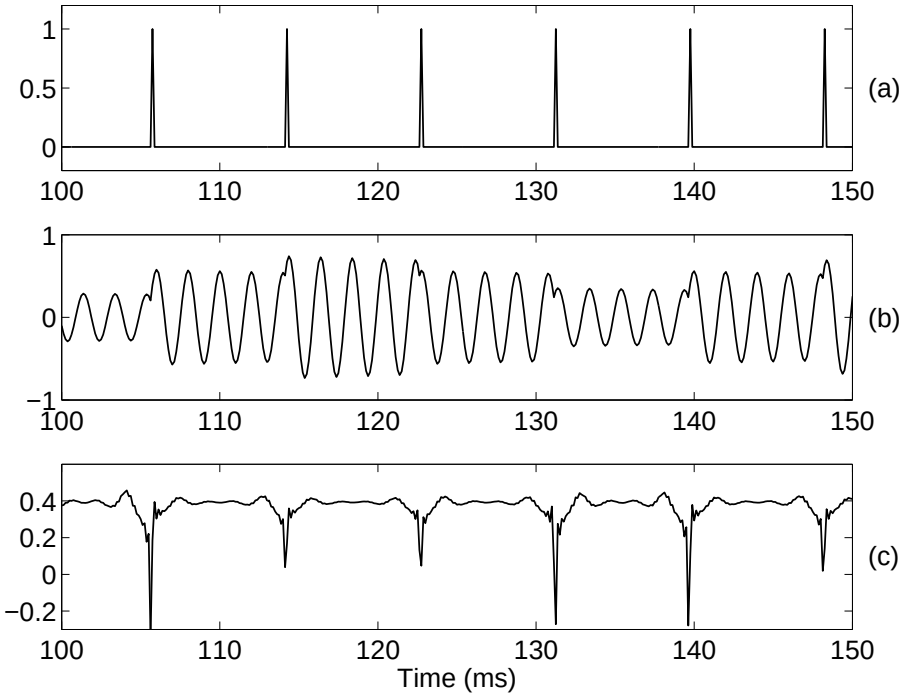


Fig. 3.3: Instantaneous frequency computed on the response of a 500 Hz resonator excited with a periodic sequence of impulses. (a) Periodic sequence of excitation impulses. (b) Output of the resonator. (c) Instantaneous frequency of the resonator output.

in the computation of the instantaneous frequency, we are not exploiting the fact that the excitation instants are periodic.

The discontinuity information can be derived from the filtered output even if the impulses are not regularly spaced, and are of arbitrary strengths. Fig. 3.4 shows a multicomponent signal in the form of a sequence of aperiodic impulses with arbitrary strengths, the filtered signal and the instantaneous frequency of the filtered signal. It is difficult to observe any discontinuity or locate the instants of excitation from the amplitude of the filtered signal. However, the instantaneous frequency (derived from phase) clearly shows sharp peaks at the instants of the excitation. The amplitudes of the peaks in the instantaneous frequency depend not only on the strengths of the impulses, but also on the phases at which the sinusoids originated at these impulses are added at the instants. This in turn depends on the locations of the impulses and the frequency of the sinusoid.

If the impulse sequence is replaced by white noise, the corresponding filtered output and the instantaneous frequency plots do not contain any significant discontinuities, as shown in Fig. 3.5. The white noise does not contain any isolated impulse-like dis-

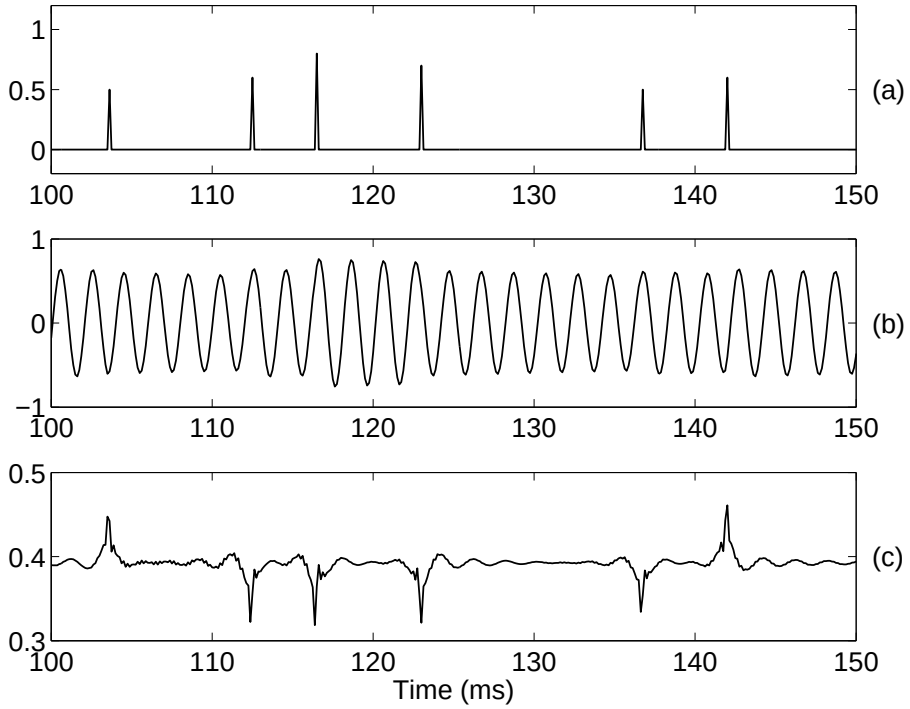


Fig. 3.4: Instantaneous frequency computed on the response of a 500 Hz resonator excited with an aperiodic sequence of impulses. (a) Aperiodic sequence of impulses with arbitrary strengths. (b) Output of the resonator. (c) Instantaneous frequency of the resonator output.

continuities. As a result, the filtered output will be a slowly varying amplitude envelope modulated by a sinusoid without any significant discontinuities in the phase. Hence the instantaneous frequency of the filtered white noise does not show any significant peaks, unlike in the case of Fig. 3.4(c). This highlights the significance of the isolated discontinuities in the impulse sequence.

Consider a situation where a synthetic speech signal is filtered through a resonator. The synthetic speech signal is generated by exciting a time-varying all-pole system by a sequence of impulses at known locations. When such a signal is filtered through a resonator, the frequency response of the all-pole system gets multiplied with the frequency response of the resonator. Hence, the frequency response of the all-pole system around the center frequency of the resonator gets selected. The filtered output carries the information about the discontinuities that are reflected in the narrow frequency band of the resonator. The instants of excitation impulses can be extracted from the filtered output using the instantaneous frequency. Fig. 3.6(b) shows a synthetic speech signal, obtained by exciting

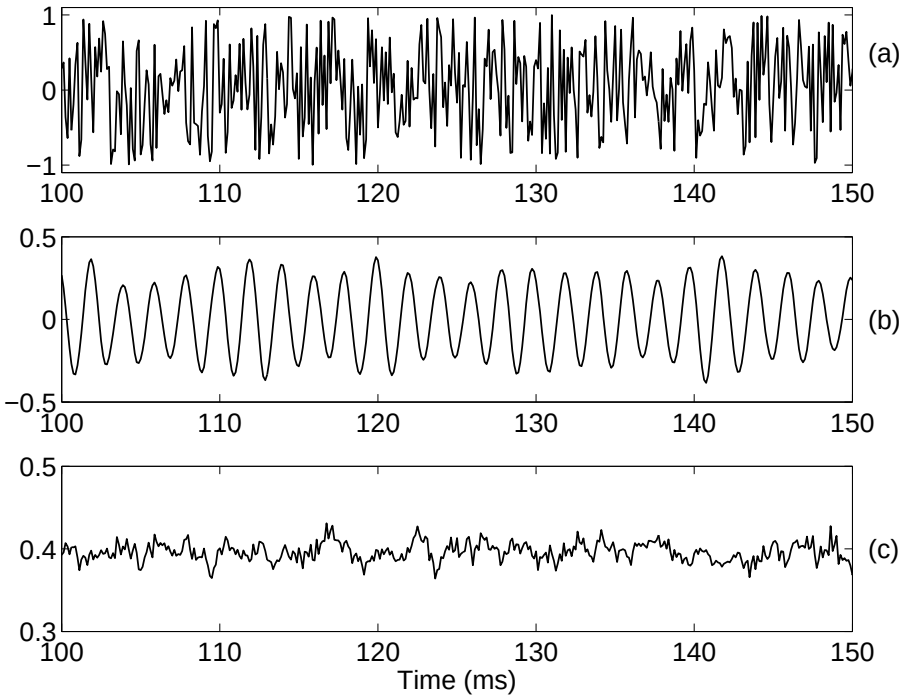


Fig. 3.5: White noise filtered through a 500 Hz resonator. (a) Segment of white noise. (b) Output of the resonator. (c) Instantaneous frequency of the resonator output.

a time-varying all-pole system with a sequence of impulses shown in Fig. 3.6(a). The instantaneous frequency (Fig. 3.6(d)) of the filtered output (Fig. 3.6(c)) shows discontinuities at the instants of excitation of the all-pole system. The locations of the discontinuities are in close agreement with the original excitation impulses.

3.2 Illustration of instantaneous frequency for speech data

Speech signal can be considered as the result of convolution of the time-varying vocal-tract transfer function and the epoch sequence due to the excitation source. The epochs are the time instants where significant excitation is delivered to the vocal-tract system. The information about the locations of the epochs is embedded in the coupling between the source and the system, though it is not evident from the speech waveform directly. It is difficult to accurately locate the time instants of excitation impulses directly from the speech waveform, because of the time-varying resonances of the vocal-tract system. To highlight the effect due to the instants of significant excitation, the speech signal is filtered

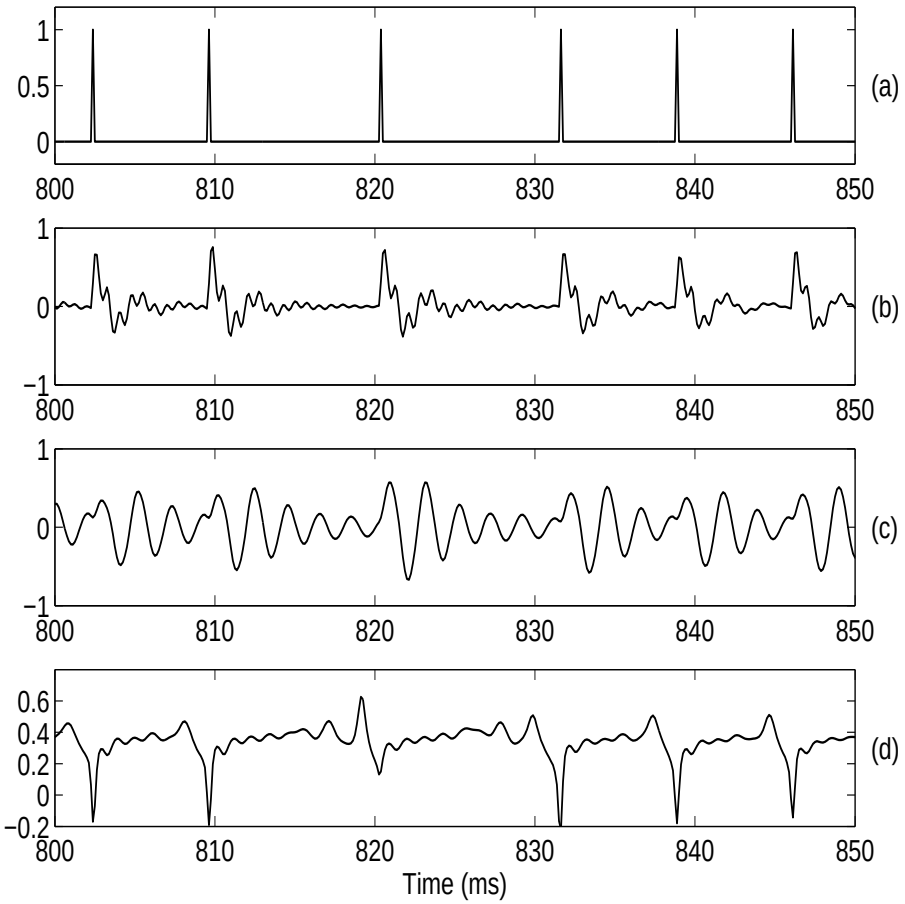


Fig. 3.6: Epoch extraction from synthetic speech signal with known epoch locations using instantaneous frequency computed around 500 Hz. (a) Sequence of excitation impulses. (b) Synthetic speech signal obtained by exciting an all-pole system with the excitation impulses. (c) Output of filtering the synthetic speech signal through a 500 Hz resonator. (d) Instantaneous frequency of the resonator output.

through a resonator centered around a chosen frequency ω_0. The significant deviations of the filtered output from the natural oscillations of the resonator can be attributed to the excitation impulses. Fig. 3.7 shows a 100 ms segment of voiced speech signal sampled at 8 kHz, and the output of the resonator at 500 Hz. The instantaneous frequency of the filtered output shows sharp peaks at the epoch locations, as shown in Fig. 3.7(c). In order to determine the accuracy of the estimated epoch locations, the differenced electroglottograph (EGG) signal is also given in Fig. 3.7(d). The peaks in the instantaneous frequency of the filtered output match well with the actual epoch locations given by the differenced EGG signal, illustrating the potential of the proposed method.

In the case of speech signal, instantaneous frequency of the filtered output also con-

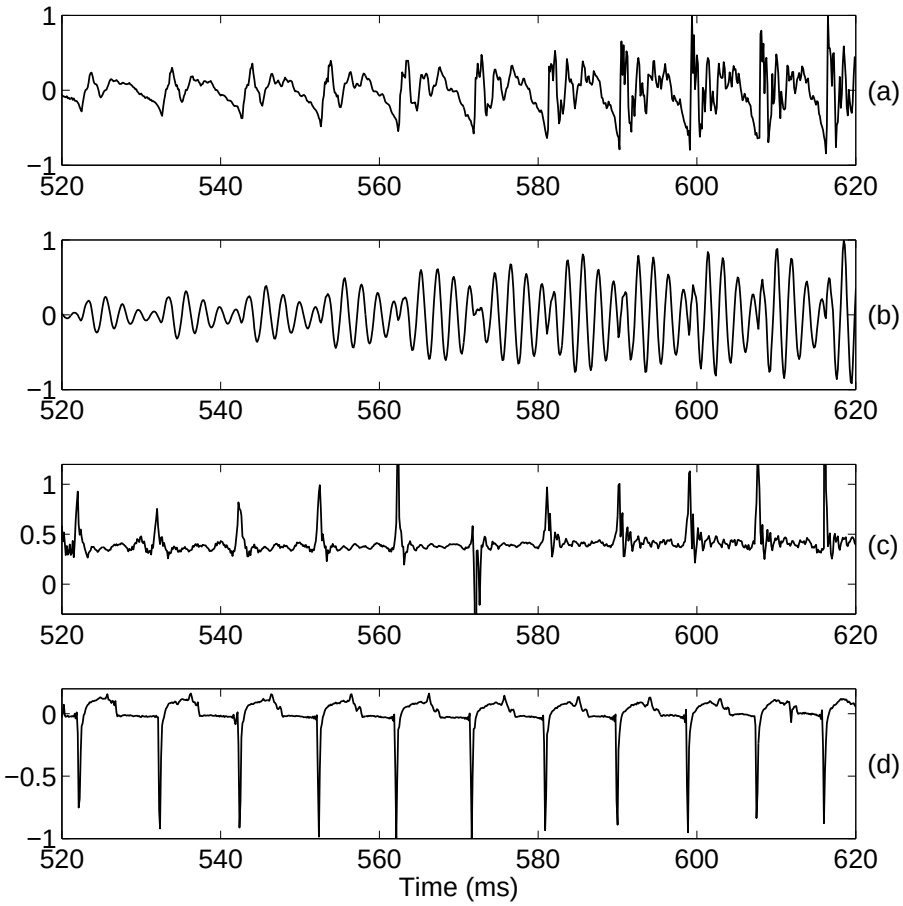


Fig. 3.7: Epoch extraction from real speech segment using instantaneous frequency. (a) A 100 ms segment of speech waveform. (b) Output of filtering the speech segment through a resonator at 500 Hz. (c) Instantaneous frequency of the resonator output. (d) Differenced EGG signal to observe the reference epoch locations.

tains the time-varying frequency changes associated with the vocal-tract transfer function, which is undesirable. As a result, though the peaks in the instantaneous frequency of the filtered output indicate the epoch locations accurately for the segment shown in Fig. 3.7, it may not be useful to extract the epoch locations unambiguously for any chosen center frequency (ω_0). Thus the method of epoch extraction using the instantaneous frequency of the filtered output depends critically on the choice of center frequency of the filter. A single center frequency may not be suitable for extracting the epoch locations of an arbitrary segment of speech. The center frequency has to be chosen based on the characteristics of the speech segment under analysis. The choice of the center frequency also depends on the distribution of energy of the speech segment in the frequency domain. To illustrate the significance of choice of the center frequency of the filter, the instantaneous frequencies

computed around four different center frequencies are shown in Fig. 3.8. The spectrogram, the speech signal and the differenced EGG signal are also shown for reference. The spectrogram in Fig. 3.8(a) shows a band of energy around 500 Hz. The instantaneous frequency computed around 500 Hz (Fig. 3.8(d)) indicates unambiguous peaks/valleys that are in close agreement with the actual epochs shown by the differenced EGG signal (Fig. 3.8(c)). In the instantaneous frequencies computed around 1000 Hz and 2000 Hz, shown in Fig. 3.8(e) and Fig. 3.8(f), respectively, the epoch locations can not be identified easily. This is because the energy of the signal in those frequency bands is very low. Since the spectrogram shows large energy in the band around 2500 Hz, the instantaneous frequency computed around 2500 Hz shows sharp peaks/valleys around the epoch locations. But, the instantaneous frequency plot in Fig. 3.8(g) shows less ambiguous peaks/valleys in the time interval 570 ms to 620 ms, than those in the time interval 520 ms to 570 ms. This is because the intensity of the 2500 Hz frequency band in the time interval 570 ms to 620 ms is greater than the intensity of the band in the time interval 520 ms to 570 ms.

Notice that the instantaneous frequencies computed around 1000 Hz and 2000 Hz also contain all the peaks/valleys corresponding to the epoch locations, but they can not be located easily due to fluctuations in the neighborhood. This is because the instantaneous frequency captures not only the discontinuities due to the excitation impulses, but also the fluctuations due to the time-varying vocal-tract system. Hence it is difficult to extract the instants of excitation from the instantaneous frequency computed around an arbitrary center frequency. The center frequency has to be chosen in such a way that the discontinuities due to the excitation impulses dominate over the fluctuations due to the time-varying vocal-tract system.

3.3 Epoch extraction using zero-frequency resonator

The discontinuity due to an impulse excitation reflects uniformly across all the frequencies including the zero-frequency. That is, even the output of a resonator at zero-frequency (0 Hz) should have the information of the discontinuities due to impulse-like excitation. The advantage of choosing the zero-frequency resonator is that the characteristics of the

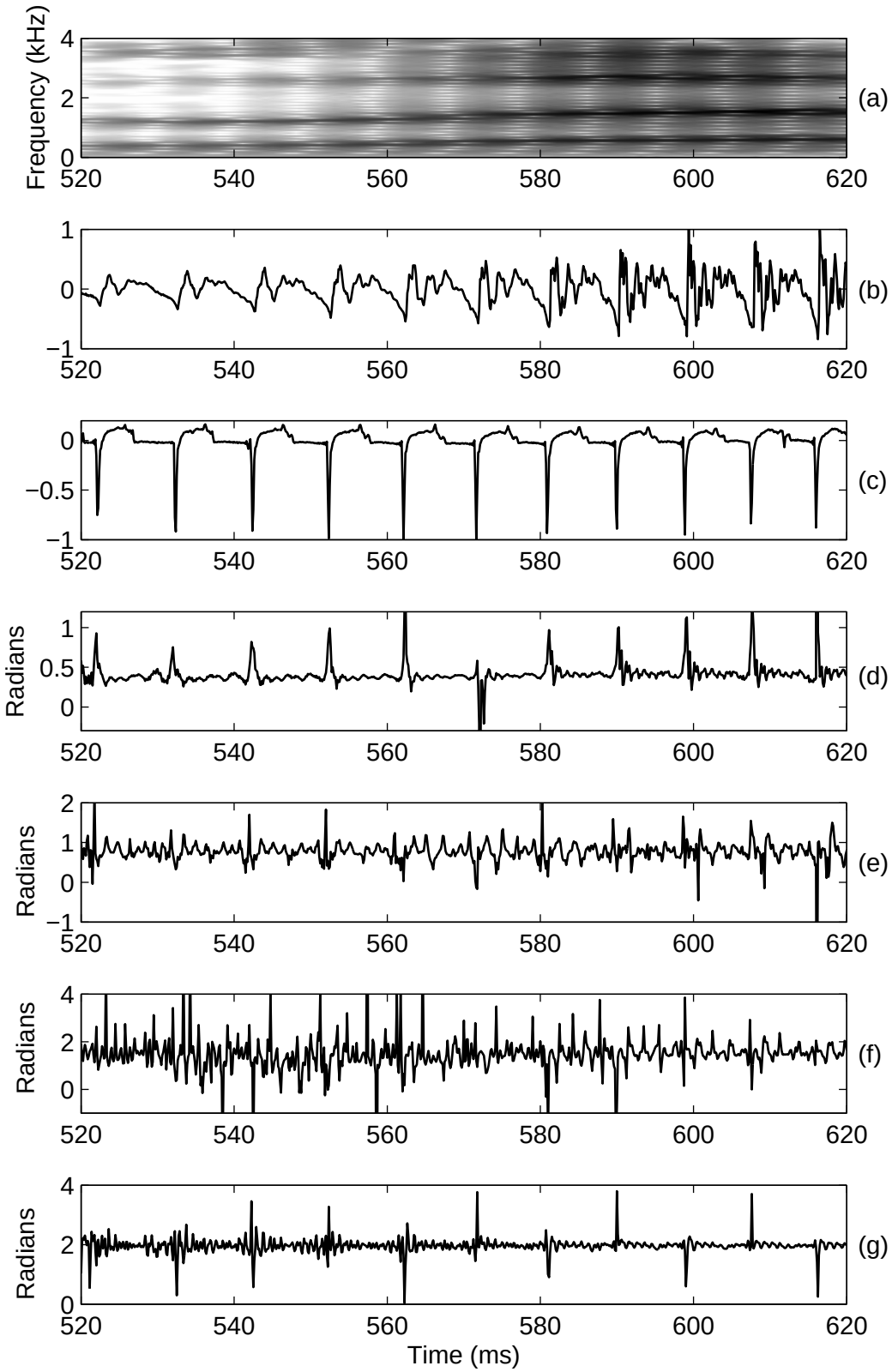


Fig. 3.8: Illustration of criticality of choice of center frequency of the resonator for epoch extraction using instantaneous frequency. (a) Spectrogram of the speech segment. (b) Speech waveform. (c) Differenced EGG signal. Instantaneous frequency plots computed around (d) 500 Hz, (e) 1000 Hz, (f) 2000 Hz, and (g) 2500 Hz.

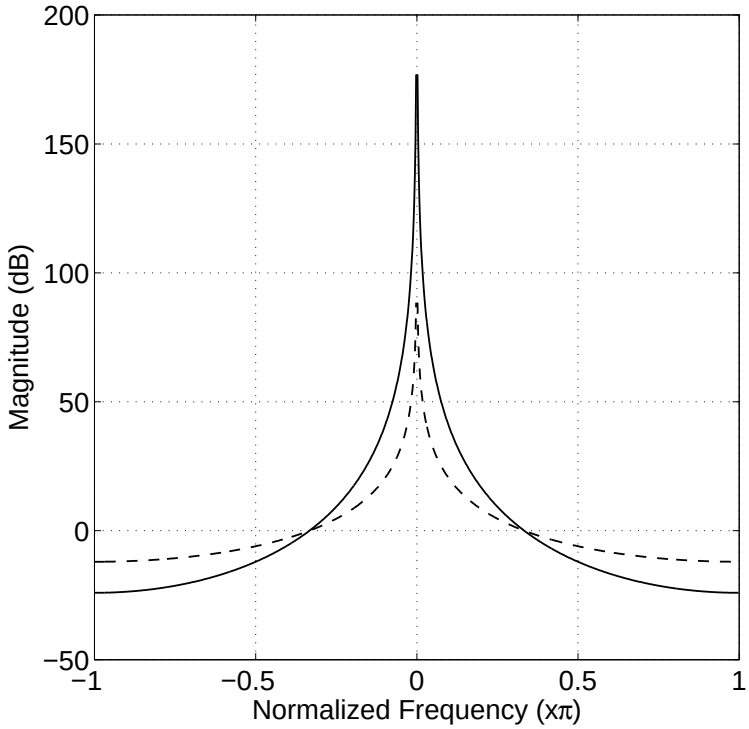


Fig. 3.9: Dotted line indicates the magnitude response of an ideal (discrete) zero-frequency resonator. Solid line indicates the magnitude response of a cascade of two ideal (discrete) zero-frequency resonators.

time-varying vocal-tract system will not affect the characteristics of the discontinuities in the resonator output. This is because the vocal-tract system has resonances at much higher frequencies than at zero-frequency. Therefore we propose that the characteristics of the discontinuities due to excitation impulses can be extracted by passing the speech signal twice through a zero-frequency resonator. The purpose of passing the speech signal twice is to reduce the effects of all (high frequency) resonances. A cascade of two zero-frequency resonators provides a sharper roll-off compared to a single zero-frequency resonator, as shown in Fig. 3.9. Since the output of the zero-frequency resonator is equivalent to double integration of the signal, passing the speech signal twice through the filter is equivalent to four times successive integration. This will result in a filtered output that grows/decays as a polynomial function of time. Fig. 3.10 shows a segment of speech signal, and its filtered output. The effect of discontinuities due to impulse sequences will be overridden by those large values of the filtered output. Hence it is difficult to compute the instantaneous frequency (deviation from zero-frequency) as in the conventional manner of computing the analytic signal of the filtered output.

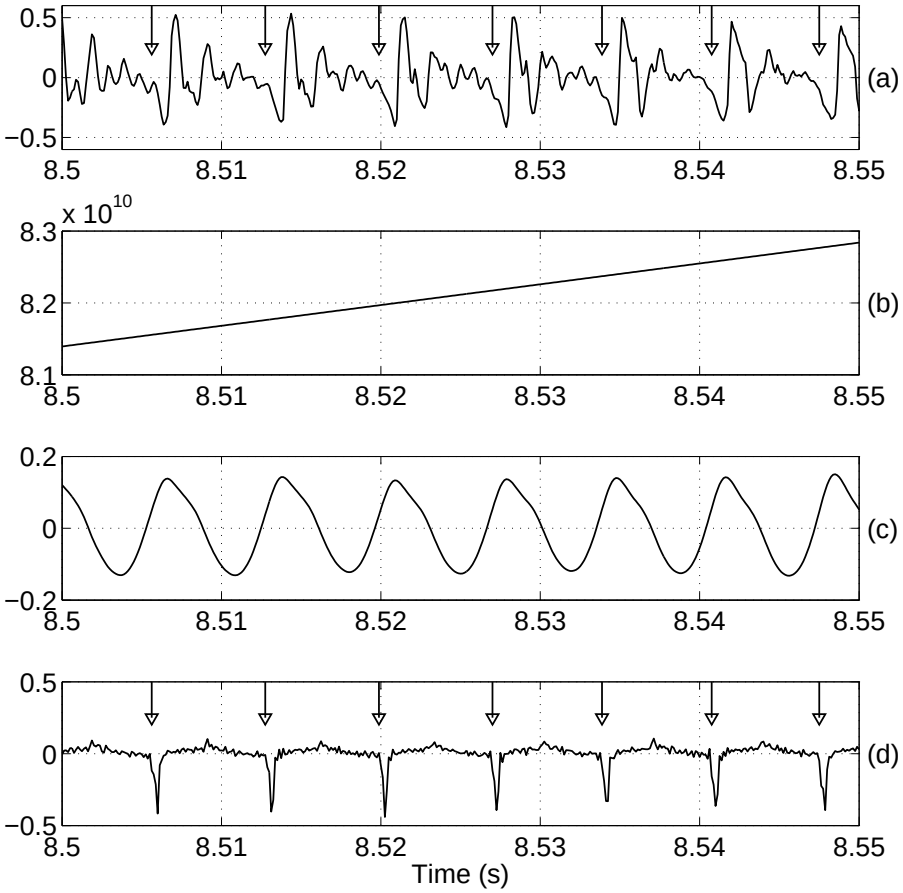


Fig. 3.10: Epoch extraction using zero-frequency resonator. A 50 ms segment of (a) Speech signal, (b) output of cascade of two 0 Hz resonators, (c) mean subtracted signal or filtered signal (d) differenced EGG signal. The arrows in (a) and (d) indicate epoch locations detected from the positive zero-crossings of the filtered signal.

We attempt to compute the deviation of the filtered output from the local mean to extract the characteristics of the discontinuities due to impulse excitation. The local mean computed over an average pitch period is subtracted from the filtered output to highlight the characteristics of the discontinuities. The resulting mean subtracted signal obtained from the filtered output in Fig. 3.10(b) is shown in Fig. 3.10(c). The mean subtracted signal is called the *zero-frequency filtered signal* or merely the *filtered signal*. The following steps are involved in processing the speech signal to derive the zero-frequency filtered signal:

(a) Difference the speech signal $s[n]$ (to remove any time-varying low frequency bias

in the signal)

$$x[n] = s[n] - s[n-1] \tag{3.17}$$

(b) Pass the differenced speech signal $x[n]$ twice through an ideal resonator at zero-frequency. That is

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n], \tag{3.18a}$$

and

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n], \tag{3.18b}$$

where $a_1 = -2$, and $a_2 = 1$. This is equivalent to successive integration by four times. But we prefer to call the process as filtering at zero-frequency.

(c) Remove the trend in $y_2[n]$ by subtracting the local mean computed over an average pitch period, at each sample. The resulting signal

$$y[n] = y_2[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_2[n+m] \tag{3.19}$$

is called the zero-frequency filtered signal, or simply the filtered signal. Here $2N+1$ corresponds to the number of samples in the interval corresponding to the average pitch period.

It was observed that the sharper zero-crossings of the filtered signal closely align with the epoch locations obtained from negative peaks of differenced EGG signals. So, the time instants of sharper zero-crossings of the filtered signal can be hypothesized as epoch locations. In Fig. 3.10(c), positive-going zero-crossings are sharper than negative-going zero-crossings, and hence indicate the epoch locations. The locations of the positive-going zero-crossings of the filtered signal in Fig. 3.10(c) coincide with the locations of the negative peaks in the differenced EGG signal as shown in Fig. 3.10(d). The sharper zero crossings of the filtered signal may either be positive-going zero-crossings or negative-going zero-crossings, depending on the polarity of the signal (typically introduced by recording devices). The polarity of the sharper zero crossings can be automatically determined by comparing the slopes of the filtered signal around the positive-going and the

negative-going zero-crossings over the entire duration of the utterance. Throughout this work, we automatically detect the polarity of the signal and compensate for the polarity so that the positive-going zero-crossings coincide with the epoch locations. In the rest of the thesis, we associate positive-going zero-crossings of the filtered signal with the hypothesized epochs. We refer to the positive-going zero-crossings as simply the positive zero-crossings.

Fig. 3.11 illustrates the performance of the proposed epoch extraction method on a creaky voice segment taken from Voqual-03 database [118]. Notice that the waveform of the speech signal (Fig. 3.11(a)) in successive glottal cycles is not periodic, making it difficult to locate the epoch locations directly from the speech signal, especially around 1.34 s to 1.36 s and 1.39 s to 1.41 s. However, the filtered signal shown in Fig. 3.11(b) clearly shows sharp positive zero-crossings around the epoch locations, which match closely with the negative peak locations of the differenced EGG signal shown in Fig. 3.11(c). It is interesting to note that even for an aperiodic sequence of impulse-like excitations, the positive zero-crossings of the filtered signal correspond to the locations of the epochs. There is no such relation between the excitation and the filtered signal for the random noise excitation of the time-varying all-pole system. Also, the filtered signal has significantly lower values for the random noise excitation compared to the impulse sequence excitation. Fig. 3.12(b) shows the filtered signal for a speech signal consisting of voiced and unvoiced segments. The unvoiced segments correspond to the random noise excitation of the vocal-tract system. The differenced EGG signal (Fig. 3.12(c)) is also given in the figure to identify the voiced and unvoiced segments.

Another important feature of the proposed approach is that it does not depend on the response of the vocal-tract system, and it does not assume quasistationarity of the vocal-tract system, unlike the conventional block processing based approaches. Since there is no assumption on quasistationarity, the proposed approach does not require block processing, and it can be applied on data segments of any length. When we apply this method on longer segments (say 0.1 s to 50 s), it is necessary to apply the trend removal operation in (3.19), successively, more than once due to rapid growth/decay of the output of the zero-frequency resonators $y_2[n]$. By applying the trend removal operation several times,

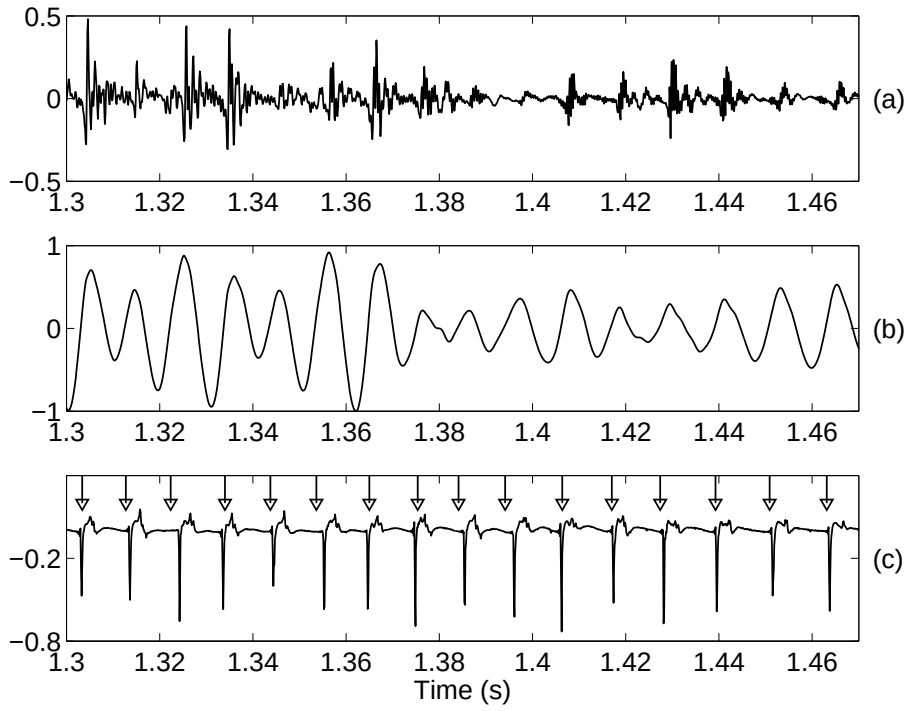


Fig. 3.11: Illustration of proposed epoch extraction method on a creaky voiced segment taken from Voqual-03 database [118]. (a) Speech waveform of a creaky voiced segment, (b) filtered signal, and (c) differenced EGG signal. The arrows in (c) indicate the epoch locations detected from the positive zero-crossings of the filtered signal.

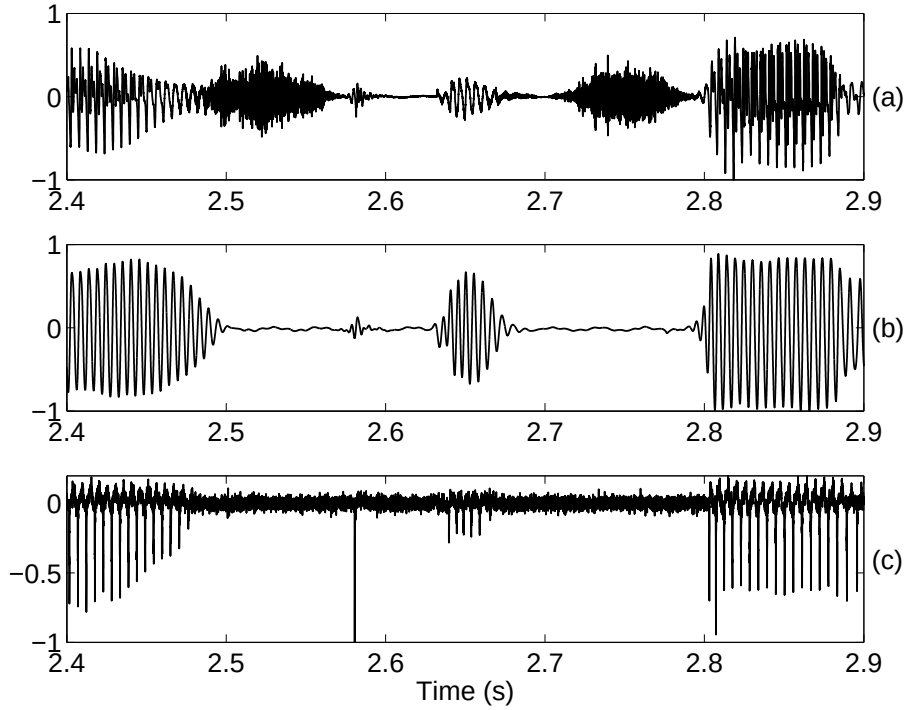


Fig. 3.12: Characteristics of filtered signal in voiced and unvoiced regions. (a) A segment of speech signal, (b) filtered signal, and (c) differenced EGG signal. The filtered output shows significantly lower values in the regions where there is no glottal activity.

the zero-crossing information does not change. Fig. 3.13 shows the effect of successive trend removal operations on the output of the zero-frequency resonators. Notice that the information in Fig 3.13(e), the signal obtained after three successive trend removals, is also present in Fig. 3.13(d), the signal obtained after two successive trend removals. But the fluctuations in Fig. 3.13(d) are overridden by a DC trend. In fact, these fluctuations are present in the output of the zero-frequency resonators also (Fig. 3.13(b)), but they are not evident because of the large DC trend arising due to filtering at zero-frequency. Throughout this thesis, the trend removal operation is applied thrice to extract the epochs.

3.3.1 Selection of window length for mean subtraction

To remove the trend in the output of the zero-frequency resonator, a suitable window length needs to be chosen to compute the local mean. The length of the window depends on the growth/decay of the output, and also on the overriding fluctuations in the output. The growth/decay in turn depends on the nature of the signal. The desired information of the overriding fluctuations depends on the intervals between impulses. If the window length is too small relative to the average duration (pitch period) between impulses, then spurious zero-crossings may occur in the filtered signal, affecting the locations of the genuine zero-crossings. If the window length is too large relative to the average pitch period, then also the genuine zero-crossings are affected in the filtered signal. Fig. 3.14 illustrates the effect of window length on the filtered signal for speech segment from a male speaker having an average pitch period of 7 ms. The filtered signal, in Fig. 3.14(b), obtained using a window length of 4 ms contains spurious (minor) zero-crossings in between the zero-crossings corresponding to the epochs. The filtered signals obtained using window lengths of 8 ms (Fig. 3.14(c)), 12 ms (Fig. 3.14(d)), and 16 ms (Fig. 3.14(e)) do not contain spurious zero-crossings. The locations of the zero-crossings across the three plots are consistent and coincide with the epoch locations. Though the filtered signal obtained using a window length of 30 ms(Fig. 3.14(f)) does not contain spurious zero-crossings, the locations of the zero-crossings are shifted arbitrarily because of the improper trend removal due to large window length. Hence, the choice of the window length for comput-

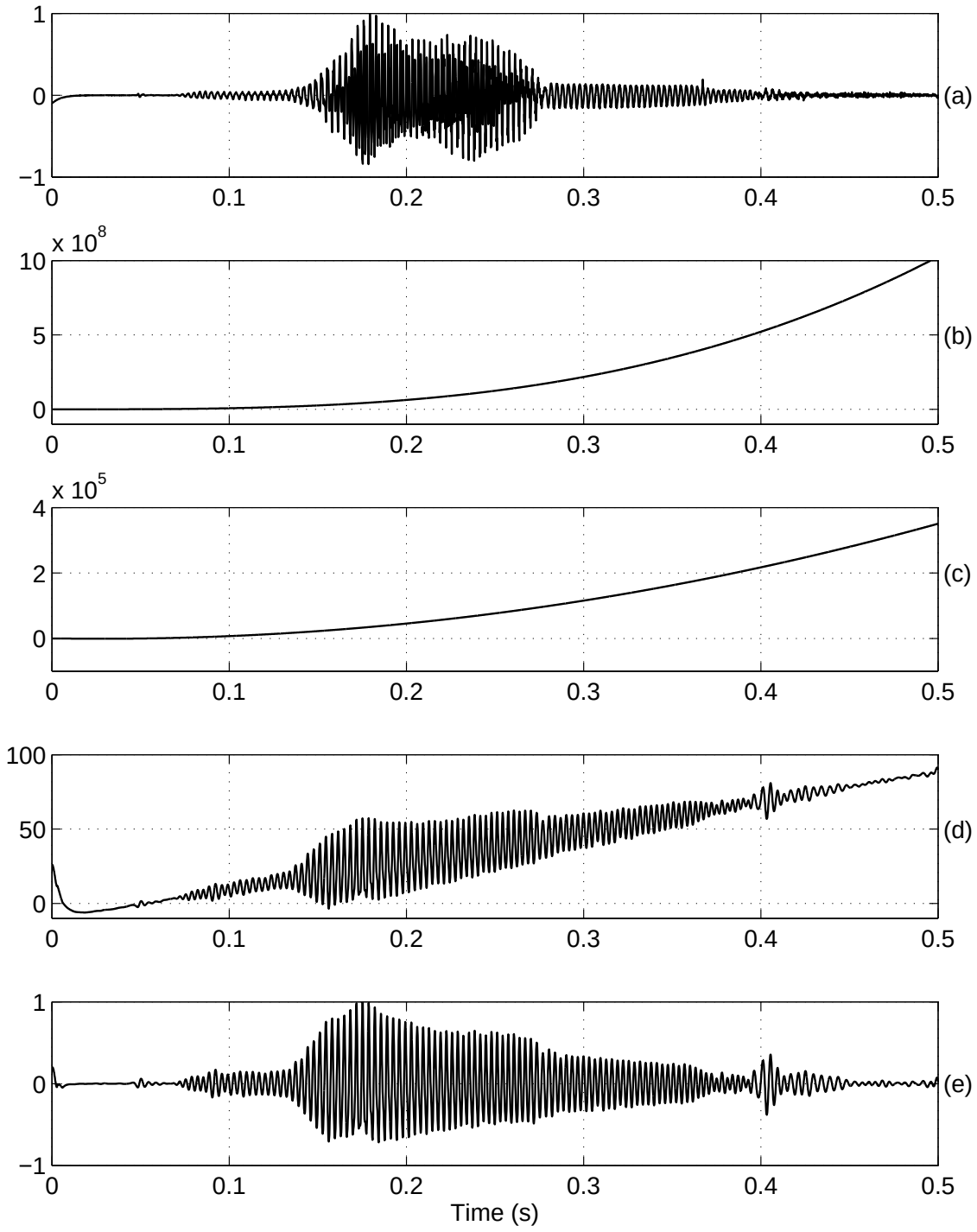


Fig. 3.13: Effect of successive trend removals from the output of the zero-frequency resonators. (a) A segment of speech signal, (b) output of the cascade of two ideal zero-frequency resonators. (c) output after first trend removal, (d) output after second trend removal, and (e) output after third trend removal, i.e., the filtered signal. The filtered signal is normalized between +1 and -1.

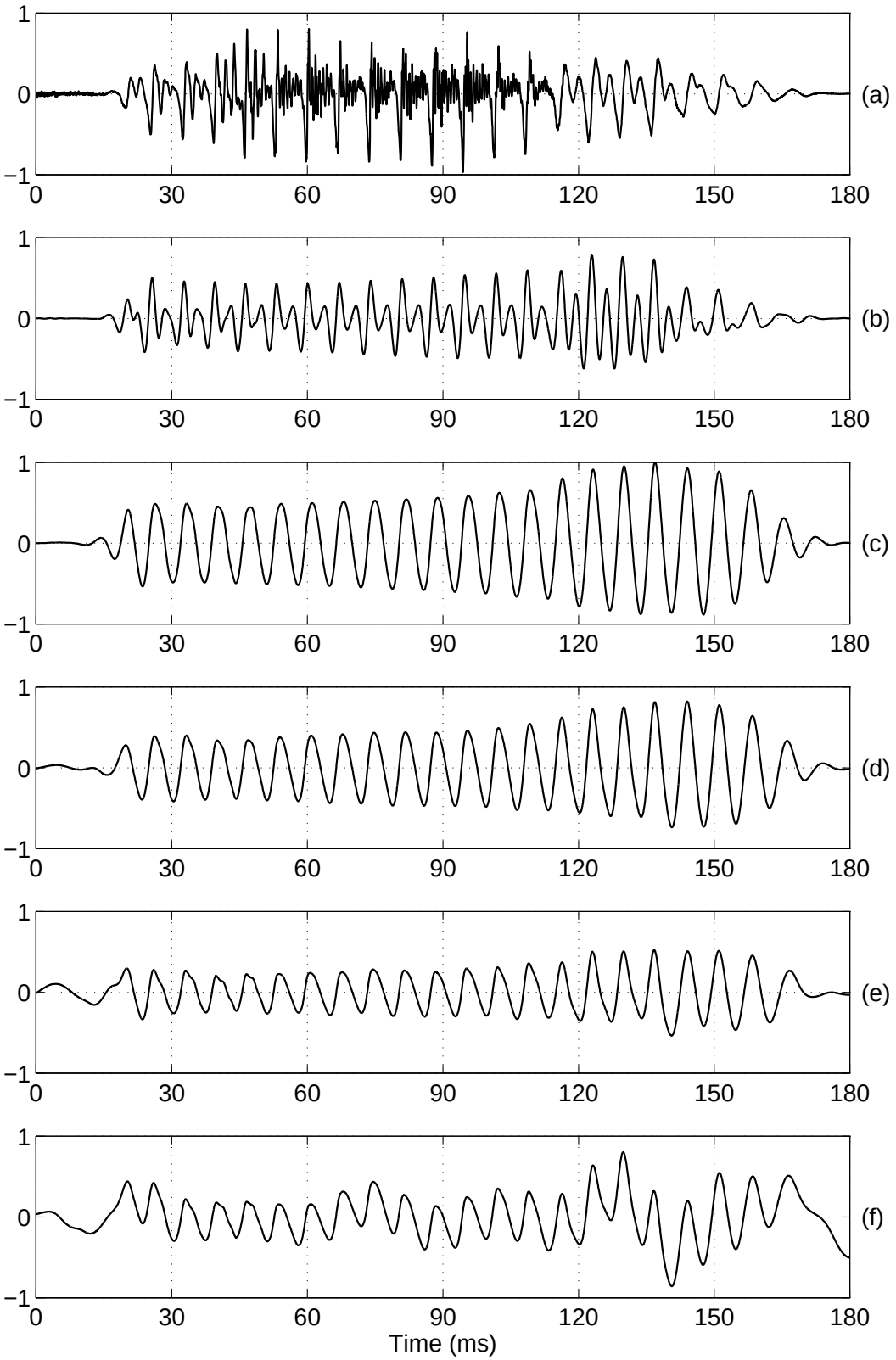
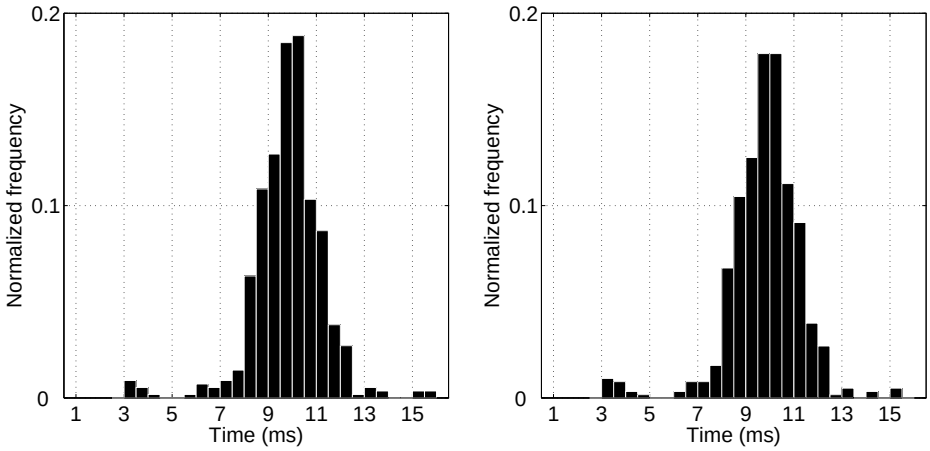


Fig. 3.14: Effect of window length for trend removal on the filtered signal. (a) A segment of speech signal. Filtered signal obtained using a window length of (b) 4 ms, (c) 8 ms, (d) 12 ms, (e) 16 ms, and (f) 30 ms,

ing the local mean is not very critical, as long as it is in the range of about 1 to 2 times the average pitch period.

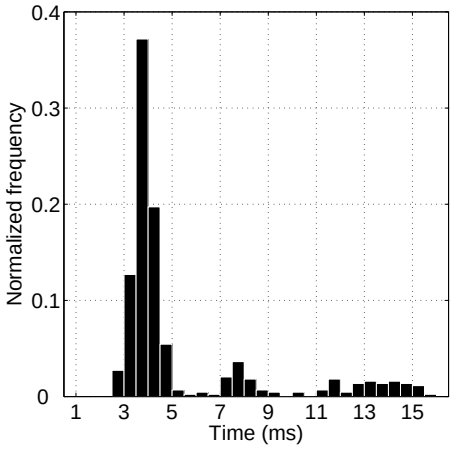
The average pitch period information can be derived in several ways. One way is to use the autocorrelation function of short (30 ms) segments of differenced speech, and determine the pitch period from the locations of the strongest peak in the interval 2 ms to 15 ms (normal range of pitch period). The histogram of the pitch periods is plotted. The pitch period value corresponding to the peak in the histogram can be chosen as the window length.

The average pitch period can be estimated using the histogram method even from degraded speech as shown in Fig. 3.15 for a male speech and a female speech at two different SNRs. The location of the peak does not change significantly even under noisy conditions. Hence the average pitch period can be estimated reliably. Fig. 3.16 shows the speech waveform, the filtered signal and the derived epoch locations and the differenced EGG signals for an utterance of a female voice. The epoch locations coincide with the locations of the large negative peaks in the differenced EGG signal (Fig. 3.16(c)). Similar illustration for a male voice is given in Fig. 3.17.

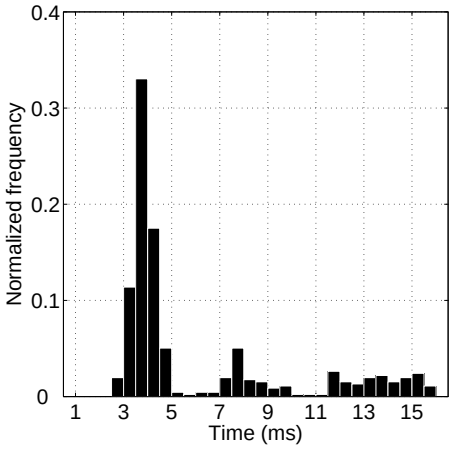


(a) Male Speaker (Clean)

(b) Male Speaker (0 dB)



(c) Female Speaker (Clean)



(d) Female Speaker (0 dB)

Fig. 3.15: Histogram of the locations of the pitch peak in the autocorrelation function for (a) clean speech signal from a male speaker, (b) speech signal from the same male speaker at 0 dB SNR, (c) clean speech signal from a female speaker, and (d) speech signal from the same female speaker at 0 dB SNR. Note that the location of the peak in the histogram plot is not affected by white noise.

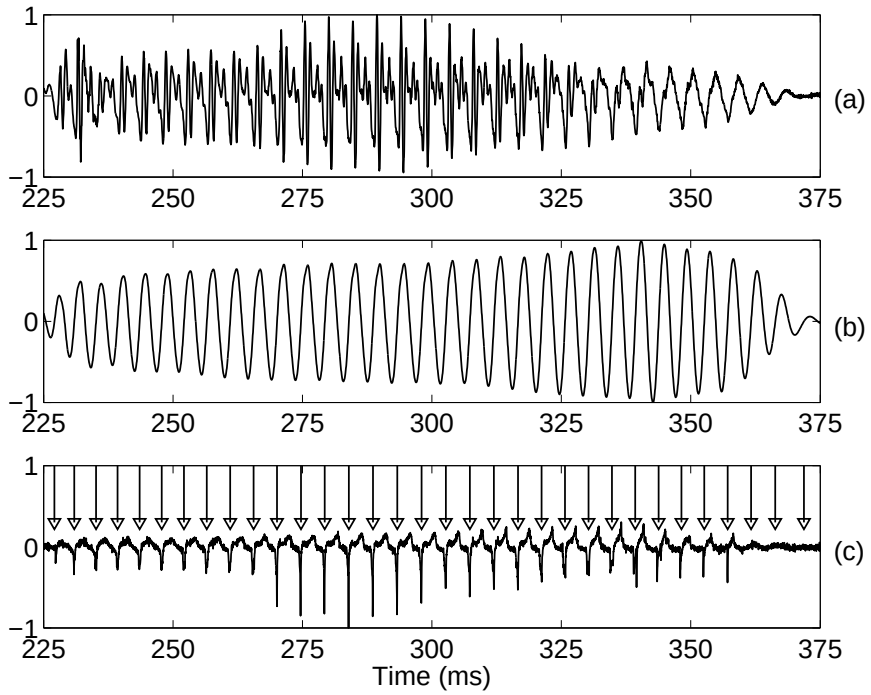


Fig. 3.16: Illustration of the proposed method of epoch extraction for female speaker. (a) Speech signal, (b) filtered signal, and (c) differenced EGG signal. Arrows in (c) indicate the detected epochs. Note that the filtered output brings out even the epochs not picked up by the EGG signal (in the interval 360 ms to 375 ms).

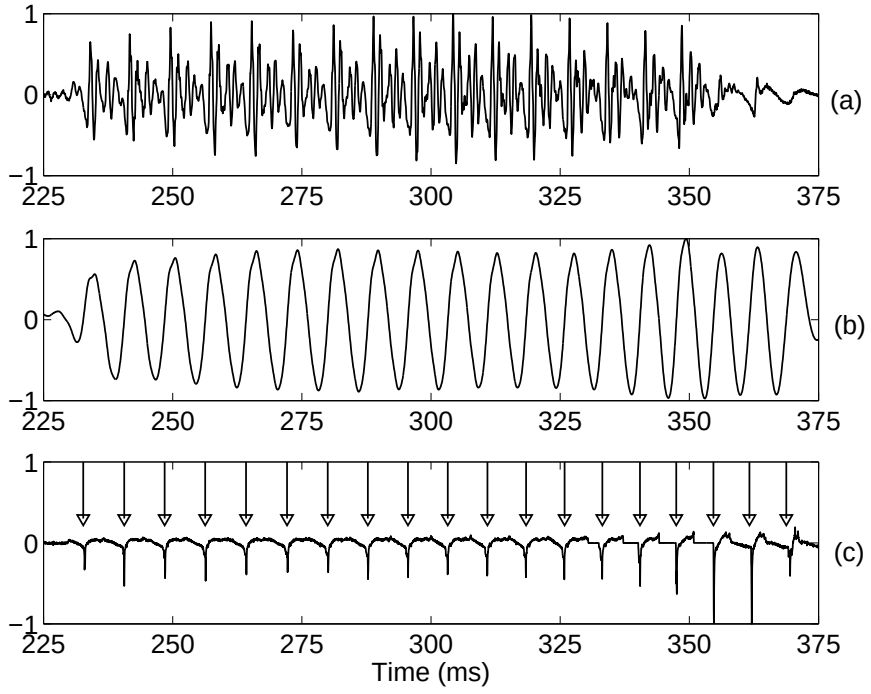


Fig. 3.17: Illustration of the proposed method of epoch extraction for male speaker. (a) Speech signal, (b) filtered signal, and (c) differenced EGG signal. Arrows in (c) indicate the detected epochs.

3.4 Comparison of proposed epoch extraction with other methods

In this section the proposed method of epoch extraction is compared with three existing methods in terms of identification accuracy and in terms of robustness against degradation. The three methods chosen for comparison are the Hilbert envelope based (HE-based) method [119], the group-delay based (GD-based) method [3] and the DYPSA algorithm [36]. Initially, the performance of the algorithms was evaluated on the clean data. Subsequently, we have evaluated robustness of the proposed method and the three existing methods at different levels of degradations. A brief discussion on the implementation details of the three chosen methods for comparison is given below.

3.4.1 Description of existing epoch extraction methods

Hilbert envelope based method [119]: During voicing, the strength of excitation at the epoch is large and impulse-like. Though this can be observed from the LP residual, it can not be extracted unambiguously because of multiple peaks of random polarity around the instant of excitation. Ideally, it is desirable to derive an impulse-like signal around the instant of significant excitation. A close approximation to this is possible by using the Hilbert envelope of the LP residual. Even though the real and imaginary parts of an analytic signal have positive and negative samples, the Hilbert envelope of a signal is a positive function, giving the envelope of the signal. For example, the HE of a unit sample sequence or its derivative has a peak at the same instant. Thus the properties of the HE can be exploited to derive approximate epoch locations. The evidence for epoch locations can be obtained by convolving the HE with a Gabor filter (modulated Gaussian pulse), as suggested in [119]. In the present work, the evidence for epoch locations is obtained by convolving the HE with a modulated Gaussian pulse,

$$g[n] = \frac{(n - N/2)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(n - N/2)^2}{2\sigma^2}\right), \quad n = 1, 2, \ldots, N,$$

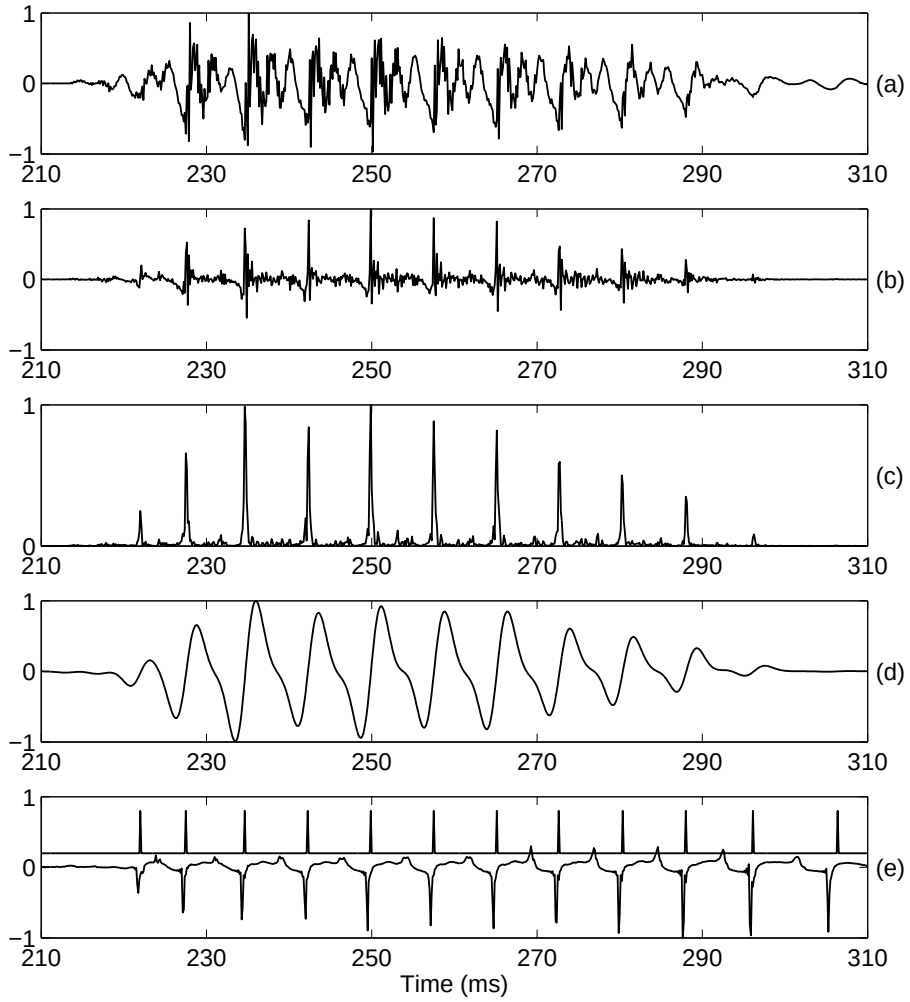


Fig. 3.18: Illustration of Hilbert envelope based method for epoch extraction [119]. (a) Speech signal, (b) LP residual, (c) Hilbert envelope of LP residual, (d) epoch evidence plot, and (e) differenced EGG signal. The pulses in (e) indicate the detected epoch locations.

where σ defines the spatial spread of the Gaussian, and N is the length of the filter. For this evaluation, the values of $\sigma = 10$, and $N = 80$ (number of samples equivalent to a duration of 10 ms, at 8 kHz sampling frequency) are used. The Hilbert envelope of the LP residual is convolved with the modulated Gaussian pulse to obtain the epoch evidence plot shown in Fig. 3.18(d). The instants of positive zero-crossings in the epoch evidence plot correspond approximately to the locations of the instants of significant excitation.

Group delay based method [3]: This method is based on the global phase characteristics of minimum phase signals. The average slope of the unwrapped phase of the short-time Fourier transform of LP residual is computed as a function of time. The average slope

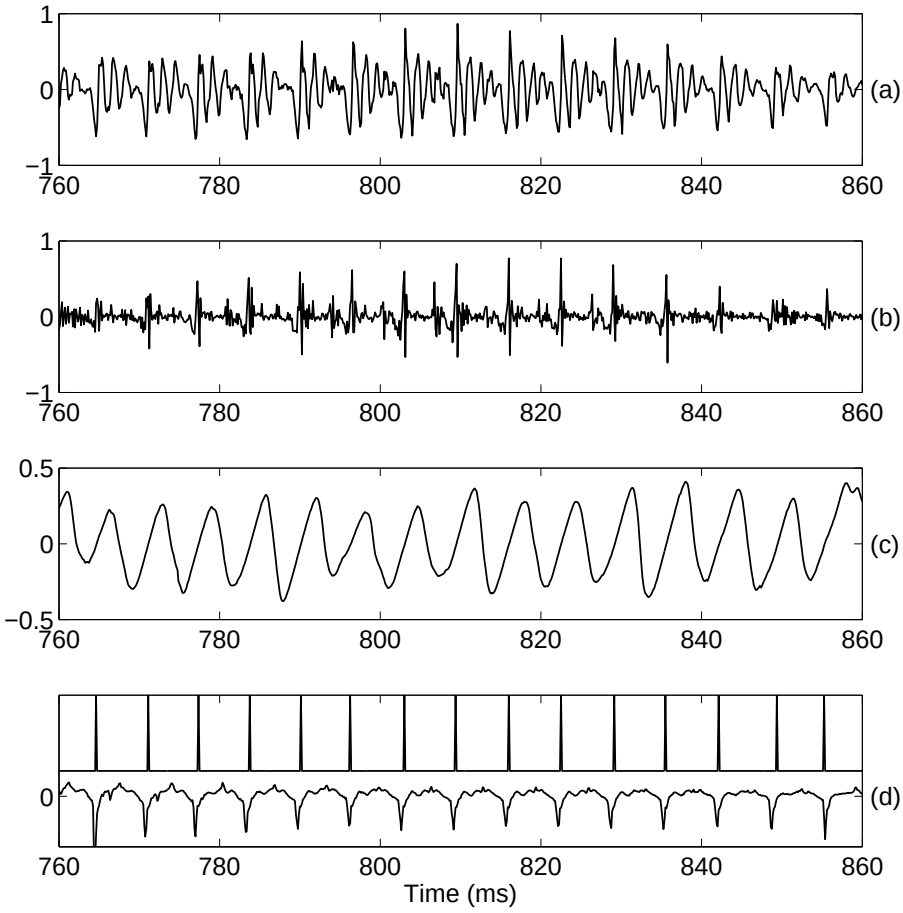


Fig. 3.19: Illustration of group-delay based method for epoch extraction [3]. (a) Speech signal, (b) LP residual, (c) phase-slope function, and (d) differenced EGG signal. The pulses in (d) indicate the detected epoch locations.

obtained as a function of time is termed as phase-slope function. Instants where the phase-slope function makes a positive zero-crossing are identified as epochs. Fig. 3.19 shows a speech utterance, its LP residual, the phase-slope function and the extracted instants. For this evaluation, we have used a 10^{th} order LP analysis to derive the LP residual, and an 8 ms window for computing the phase-slope function.

The DYPSA algorithm [36]: The DYPSA algorithm is an automatic technique for estimating the epochs in voiced speech from the speech signal alone. There are three components in the algorithm. The first component generates candidate epochs using zero-crossings of the phase-slope function. The energy weighted group-delay was used as a measure to derive the phase-slope function. The second component employs a novel phase-slope projection technique to recover candidates for which the phase-slope func-

tion does not include a zero-crossing. These two components detect almost all the true epochs, but they also generate a large number of false alarms. The third component of the algorithm uses dynamic programming to identify the true epochs from the set of hypothesized candidates by minimizing a cost function. For evaluating this technique, the MATLAB implementation of the DYPSA available in [120] was used.

3.4.2 Database for evaluation of epoch extraction methods

The CMU-Arctic database [121][122] was used to evaluate the proposed method of epoch extraction, and to compare the results with the existing methods. The Arctic database consists of 1132 phonetically balanced English sentences spoken by two male and one female speakers. The duration of each utterance is approximately 3 s, which makes the duration of the entire database to be around 2 hours 40 minutes. The database was collected in a sound proof booth, and digitized at a sampling frequency of 32 kHz. In addition to the speech signals, the Arctic database contains simultaneous recordings of EGG signals collected using an electroglottograph. The speech and EGG signals were time-aligned to compensate for the larynx-to-microphone delay, determined to be approximately 0.7 ms. Reference locations of the epochs were extracted from the voiced segments of the EGG signals by finding peaks in the differenced EGG signal. The performance of the algorithms was evaluated only in the voiced segments (detected from EGG signal) between the reference epoch locations and the estimated epoch locations. The database contains a total of 792249 epochs in the voiced regions.

3.4.3 Performance evaluation

The performance of the epoch detection methods was evaluated using the measures defined in [36]. Fig. 3.20 shows the characterization of epoch estimates showing each of the possible decisions from the epoch detection algorithms. The following measures were defined to evaluate the performance of the epoch detection algorithms:

(a) *Larynx cycle:* The range of samples $\frac{1}{2}(l_{r-1} + l_r) \leq n \leq \frac{1}{2}(l_r + l_{r+1})$, given an epoch reference at sample l_r with preceding and succeeding epoch references at samples l_{r-1} and l_{r+1}, respectively.

(b) *Identification rate* (IDR): The percentage of larynx cycles for which exactly one epoch is detected.

(c) *Miss rate* (MR): The percentage of larynx cycles for which no epoch is detected.

(d) *False alarm rate* (FAR): The percentage of larynx cycles for which more than one epoch is detected.

(e) *Identification error ζ:* The timing error between the reference epoch location and the detected epoch location in larynx cycles for which exactly one epoch is detected.

(f) *Identification accuracy σ* (IDA): The standard deviation of the identification error ζ. Small values of σ indicate high accuracy of identification.

Table 3.1 shows the performance results on Arctic database for identification rate, miss rate, false alarm rate, and identification accuracy for the three existing methods, HE-based, GD-based and DYPSA algorithm, as well as for the proposed method. Fig. 3.21 shows the histograms of the timing errors ζ in detecting the epoch locations, averaged over the entire Arctic database. The spread of the timing errors for the proposed method is relatively less compared to the exiting methods. From Table 3.1, it can be concluded that the DYPSA algorithm performed best among the three existing techniques, with an identification rate of 96.66%. The proposed method of epoch extraction gives even better identification rate as well as identification accuracy, compared to the DYPSA algorithm.

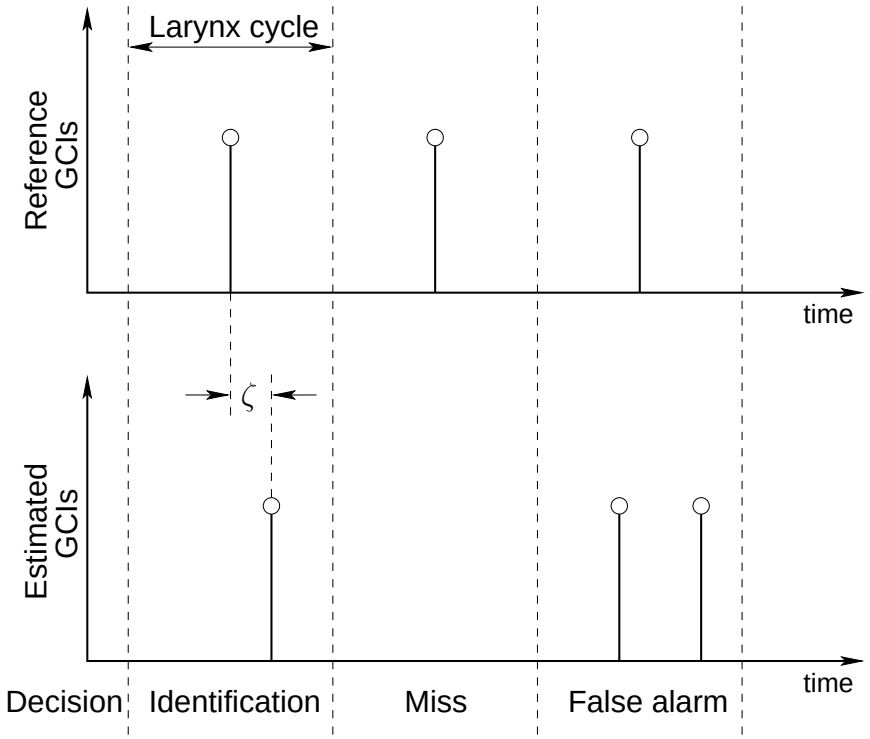
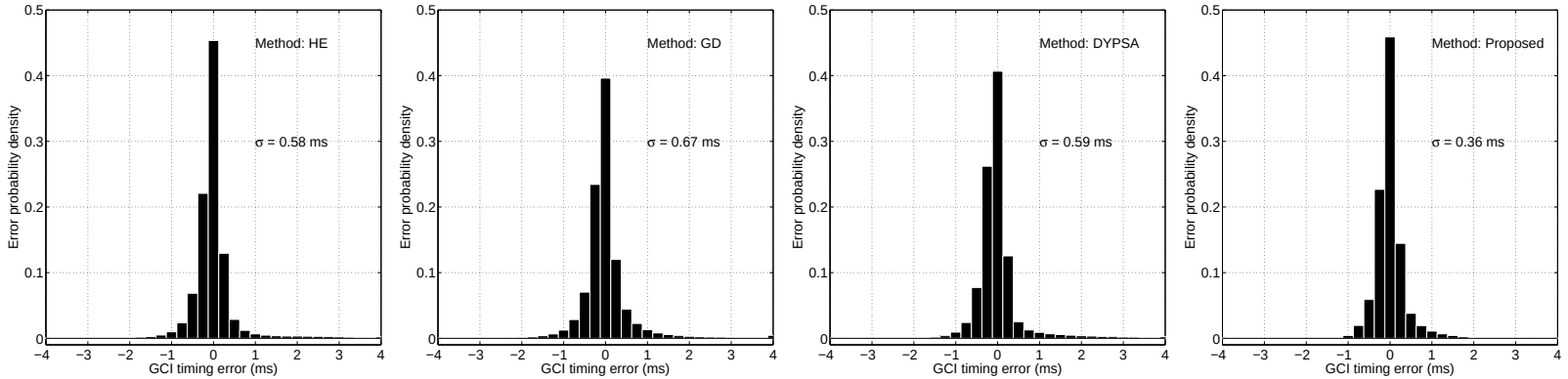


Fig. 3.20: Characterization of epoch estimates showing 3 larynx cycles with examples of each possible outcome from epoch extraction [36]. Identification accuracy is measured as standard deviation of ζ.

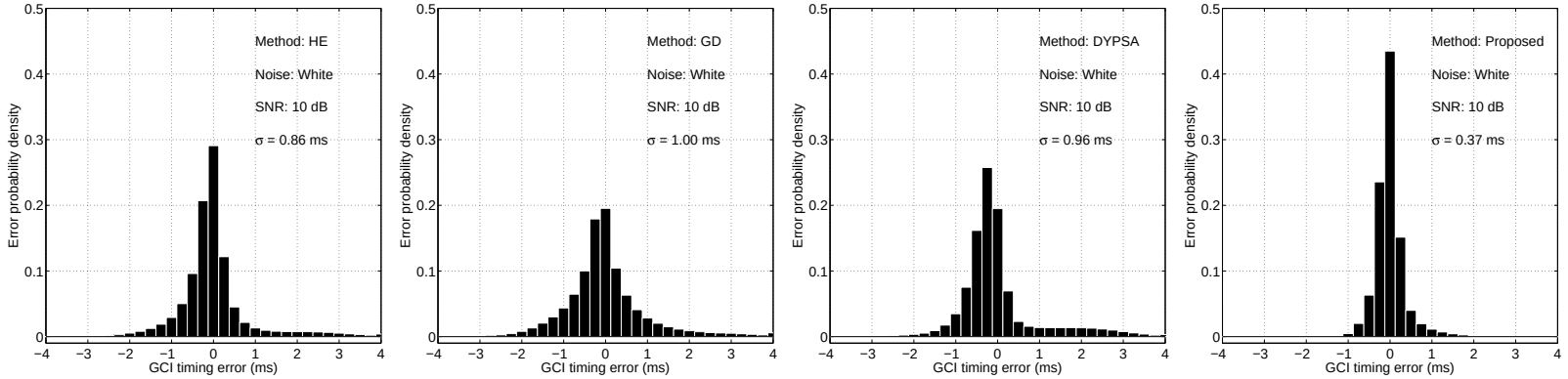
Table 3.1: Performance comparison of epoch extraction methods on CMU-Arctic database. IDA - Identification rate, MR - Miss rate, FAR - False alarm rate, IDA - Identification accuracy.

Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)
HE-based	89.86	1.43	8.71	0.58
GD-based	92.80	4.01	3.18	0.67
DYPSA	96.66	1.76	1.58	0.59
Proposed	99.04	0.18	0.77	0.36



(a) HE-based (b) GD-based (c) DYPSA (d) Proposed

Fig. 3.21: Histogram of the epoch timing errors for clean speech. (a) HE-based method, (b) GD-based method, (c) DYPSA algorithm and (d) proposed method.



(a) HE-based (b) GD-based (c) DYPSA (d) Proposed

Fig. 3.22: Histogram of the epoch timing errors for speech signals, degraded by white noise, at an SNR of 10 dB. (a) HE-based method, (b) GD-based method, (c) DYPSA algorithm and (d) proposed method.

3.5 Effect of noise on performance of the proposed method of epoch extraction

In this section we study the effect of (moderate levels of) noise on the accuracy of the epoch detection methods. The existing methods and the proposed method are evaluated on artificially generated noisy speech data. Several noise environments at varying signal-to-noise ratio (SNR) were simulated to evaluate the robustness of the epoch detection methods. The noise samples were taken from NOISEX-92 database [123]. The database consists of white, babble, high frequency (HF) channel, and vehicle noise. The noise samples from the NOISEX-92 database were added to the clean speech utterances from Arctic database to generate noisy speech data at different levels of degradation. The utterances are appended with silence so that the total amount of silence in each utterance is constrained to be about 60% of data, including the pauses in the utterances. Including different noise environments and SNRs, the database consists of 33 hours of noisy speech data.

Table 3.2 shows the comparative performance of epoch extraction methods for different types of degradations at varying SNRs. Fig. 3.22 shows the distribution of the timing errors ζ in detecting the epoch locations, for white noise environment at an SNR of 10 dB. The proposed method consistently performs better than the existing techniques even under degradation. The improved performance of the proposed method may be attributed to the following reasons: (a) There is no block processing involved in this method. Hence there are no effects of the size and the shape of the window. The entire speech signal is processed at once to obtain the filtered signal. (b) The proposed method is not dependent on the energy of the signal. This method detects the epoch locations even in weakly voiced regions like voice-bar. (c) There is only one parameter involved in the proposed method, i.e., the length of the window for removing the trend from the output of zero-frequency resonator, the choice of which is not very critical. (d) There are no critical thresholds or costs involved in identifying the epoch locations.

Table 3.2: Performance comparison for epoch detection methods for various SNRs and noise environments. IDR - Identification rate, MR - Miss rate, FAR - False alarm rate, IDA - Identification accuracy

Environment		**HE Based**				**GD Based**			
Noise	SNR (dB)	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDR (%)	MR (%)	FAR (%)	IDA (ms)
White	20 dB	84.56	1.58	13.86	0.686	87.34	3.82	8.85	0.812
White	15 dB	82.26	1.9	15.85	0.761	84.65	4.15	11.2	0.891
White	10 dB	79.45	2.39	18.16	0.864	81.07	4.79	14.14	0.907
Babble	20 dB	86.73	1.54	11.73	0.674	89.45	3.99	6.56	0.782
Babble	15 dB	84.88	1.77	13.35	0.743	87.27	4.28	8.45	0.855
Babble	10 dB	82.51	2.17	15.32	0.842	84.32	4.77	10.91	0.956
HF Channel	20 dB	84.23	1.87	13.91	0.738	86.54	4.36	9.10	0.849
HF Channel	15 dB	82.04	2.26	15.69	0.822	83.87	4.84	11.29	0.934
HF Channel	10 dB	79.24	2.85	17.91	0.927	80.13	5.53	14.34	1.040
Vehicle	20 dB	89.75	1.40	8.85	0.584	92.67	3.95	3.38	0.674
Vehicle	15 dB	89.58	1.39	9.03	0.585	92.49	3.92	3.59	0.679
Vehicle	10 dB	89.25	1.37	9.38	0.591	92.18	3.88	3.95	0.689

Environment		**DYPSA**				**Proposed Method**			
Noise	SNR (dB)	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDR (%)	MR (%)	FAR (%)	IDA (ms)
White	20 dB	92.12	1.41	6.47	0.738	99.04	0.19	0.77	0.363
White	15 dB	85.33	1.24	13.43	0.841	99.06	0.19	0.75	0.365
White	10 dB	75.95	1.09	22.96	0.957	99.05	0.23	0.72	0.371
Babble	20 dB	96.42	1.8	1.79	0.621	99.02	0.19	0.79	0.366
Babble	15 dB	96.14	1.82	2.05	0.647	98.99	0.21	0.80	0.374
Babble	10 dB	95.48	1.78	2.74	0.69	98.83	0.30	0.87	0.405
HF Channel	20 dB	95.89	1.77	2.33	0.654	99.04	0.19	0.77	0.363
HF Channel	15 dB	94.99	1.66	3.35	0.702	99.05	0.19	0.76	0.363
HF Channel	10 dB	92.4	1.56	6.01	0.775	99.06	0.21	0.73	0.368
Vehicle	20 dB	96.67	1.76	1.57	0.589	99.06	0.20	0.73	0.372
Vehicle	15 dB	96.6	1.78	1.62	0.596	98.93	0.37	0.70	0.397
Vehicle	10 dB	96.64	1.76	1.61	0.597	97.83	1.53	0.64	0.460

3.6 Summary

In this chapter we proposed a method for epoch extraction that does not depend on the characteristics of the vocal-tract system. The method exploits the impulse-like excitation of the vocal-tract system. The method uses the output of speech from a zero-frequency resonator. The positive zero-crossings of the filtered signal correspond to epochs. The identification rate and identification accuracy are evaluated using the CMU-Arctic database, where the speech signal and the corresponding EGG signals are available. The epoch information derived from the EGG signals is used as reference. The performance of the proposed method is compared with the results from the DYPSA and two other methods. The proposed method gives a significantly better performance in terms of identification rate and identification accuracy. It is also interesting to note that the proposed method is robust against degradations such as white noise, babble, high frequency channel and vehicle noise.

There are many novel features in the proposed method of epoch extraction. The method does not use any block processing as most other signal processing methods do. The performance of the method does not depend on the energy of the segment of speech signal, and hence the method works equally well for all types of voiced sound units. In addition, there are no parameters to control, and no arbitrary thresholding in the identification of epochs.

The method performs well for speech collected with close-speaking microphone, even with the addition of degradations. But the method is not likely to work well when the degradations produce additional impulse-like sequences in the collected speech data as in the case of reverberation. The method is also not likely to work well when there is interference of speech from other speakers. Some of these issues are addressed in Chapter 6 using speech signals collected over a pair of spatially separated microphones. The proposed method of epoch extraction may not work well on speech data collected over telephone channels and high pass filtered speech signals where the low frequency components are deliberately attenuated.

Chapter 4

Characterization of Glottal Activity

The primary mode of excitation of the vocal-tract system during speech production is due to vibration of vocal folds (glottal activity) at the glottis. The strength of excitation during the glottal activity is determined mostly by the rate of closure of the vocal folds in each glottal cycle. Detecting the regions of glottal activity and the strength of excitation in each glottal cycle from the speech signal is a challenging task, as it is difficult to suppress the response of the time-varying vocal-tract system in the speech signal. Several methods have been suggested in the literature, which involve estimating the characteristics of the time-varying vocal-tract system, followed by some form of inverse filtering of speech to highlight the characteristics of the excitation source [1]. Linear prediction (LP) analysis is one such method in which the LP coefficients are used to inverse filter the speech signal to derive the LP residual [4]. The LP residual has noise-like characteristics in the regions of nonglottal activity. In the regions of glottal activity, corresponding to the vocal fold vibration, the LP residual shows regions of large amplitude at regular intervals. The large energy region corresponds mostly to the closing phase of each glottal cycle. The effectiveness of detecting glottal activity from the LP residual depends on the accuracy of the LP model, and also on the nature and quality (degradation) of the speech signal.

In this chapter, we propose a method based on the zero-frequency filtered signal to detect the regions of glottal activity, and to estimate the strength of excitation in each glottal cycle. In Section 4.1, we present a method to estimate the strength of excitation at

epoch locations from the speech signals. Section 4.2 discusses a method to automatically detect the regions of glottal activity, and its performance evaluation. In Section 4.3 we summarize the contributions of this chapter.

4.1 Estimation of strength of excitation

The manner in which vocal folds vibrate influences the glottal airflow that serves as an excitation source for the vocal-tract filter. Vocal intensity may be increased by sharply truncating the expiratory airflow (sharper closure of the vocal folds), and thereby increasing the rate of glottal airflow [124]. Some of these features are manifested well in the EGG signals. The negative peak amplitude in the differenced EGG signal indicates the rate of glottal closure. However, the vocal-tract is known to absorb a variable amount of acoustic energy, and the degree of mouth opening affects the acoustic pressure level detected at the microphone [125]. Hence, the acoustic pressure level as picked up by the microphone does not provide a reliable cue for the strength of excitation or rate of glottal closure.

In this study, we exploit the narrowband nature of the zero-frequency resonator to measure the strength of excitation at each instant. Since the effect due to an impulse is spread uniformly across the frequency range, the relative strengths of impulses can be derived from a narrowband around any frequency, including the zero-frequency. Hence, the information about the strength of excitation can also be derived from the zero-frequency resonator. It is observed that the slope of the zero-frequency filtered signal around the zero-crossings corresponding to the epoch locations gives a measure of the strength of excitation. Fig. 4.1(a) and Fig. 4.1(b) show a sequence of randomly spaced impulses with arbitrary strengths, and the zero-frequency filtered signal, respectively. The filtered signal (Fig. 4.1(b)) shows sharper zero-crossings at the impulse locations, and the slopes of filtered signal around those zero-crossings are proportional to the actual impulse strengths as shown in Fig. 4.1(c). The scatter plot between the strengths of impulses and the slopes of the filtered signal shown in Fig. 4.2 clearly shows a linear trend, indicating that the estimated strengths are proportional to the actual impulse strengths.

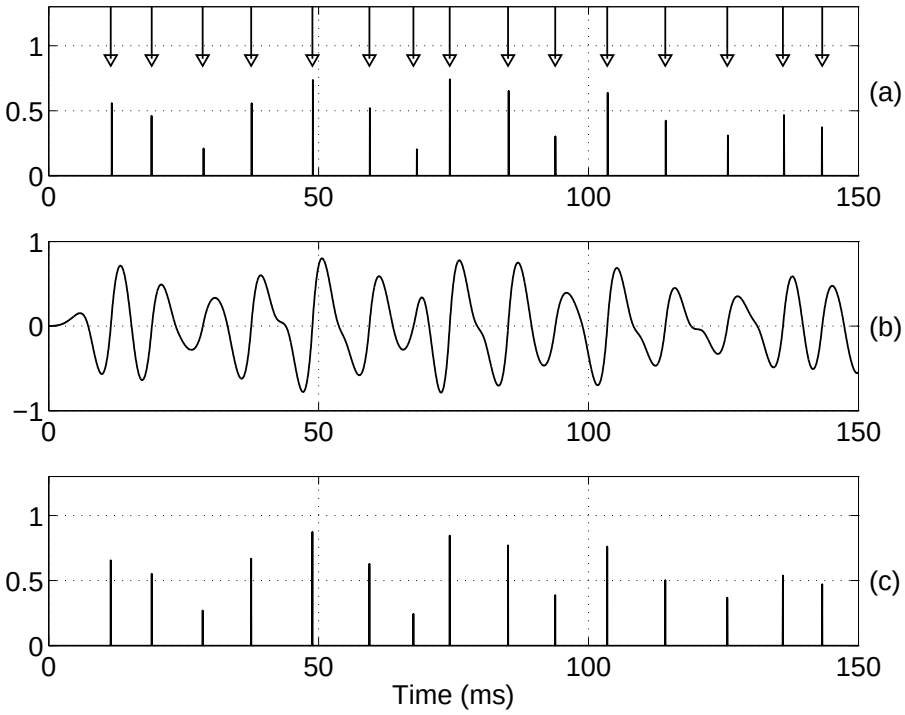


Fig. 4.1: Estimation of strength of randomly spaced impulses using zero-frequency resonator. (a) Sequence of randomly spaced impulses, (b) Zero frequency filtered signal. (c) Slope of signal around the positive (sharper) zero-crossings. Arrows in (a) indicate detected impulse locations.

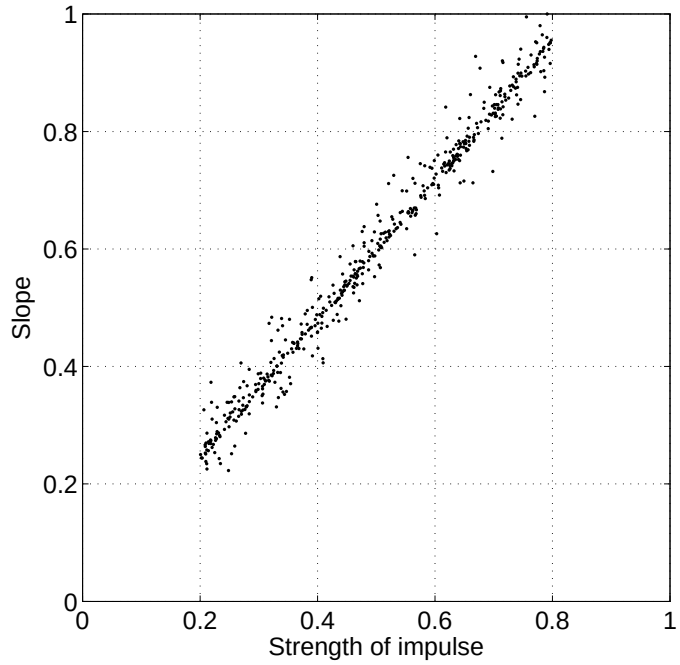


Fig. 4.2: Scatter plot of strength of impulse vs. slope of the filtered signal

This method of quantifying the strengths of impulses is valid even for speech signals. In the case of speech signals, the significant contribution at the zero-frequency is due to the impulse-like excitation. The vocal-tract system has resonances at much higher frequencies than at zero-frequency. Hence the contribution of the time-varying vocal-tract system at zero-frequency is significantly low compared to the contribution due to the impulse-like excitation. Hence the slope of the filtered signal around the epoch location reflects predominantly the strength of excitation. Fig. 4.3(d) shows the estimated strengths of excitation at the epoch locations for the speech signal shown in Fig. 4.3(a). Notice that the amplitude of the speech signal (Fig. 4.3(a)) around 0.5 s is low, though the strength of the excitation as reflected in the differenced EGG signal (Fig. 4.3(b)) is high. The strength of excitation derived from the filtered signal of speech shows similar trend as that of the differenced EGG signal. Fig. 4.4(a) shows a scatter plot between the strength of excitation derived from the differenced EGG signal and the absolute maximum amplitude of the speech signal around the epoch location. Fig. 4.4(b) shows a scatter plot between the strength of differenced EGG signal and the strength of excitation estimated from the filtered signal of speech. The scatter plot in Fig. 4.4(b) shows a better linear orientation indicating that the estimated strength of excitation is proportional to the actual strength of excitation observed from EGG signal. This behavior is not present in Fig. 4.4(a), indicating that the strength of excitation can not be directly observed from the speech signal.

4.2 Glottal activity detection (GAD)

The strength of excitation of the vocal-tract system can be considered to be significant in the regions of the vocal fold vibration (glottal activity). In the absence of vocal fold vibration, the vocal-tract system can be considered to be excited by random noise, as in the case of frication. The energy of the random noise excitation is distributed both in time and frequency domains. While the energy of an impulse is distributed uniformly in the frequency domain, it is highly concentrated in the time-domain. As a result, the filtered signal exhibits significantly lower amplitude for random noise excitation compared to

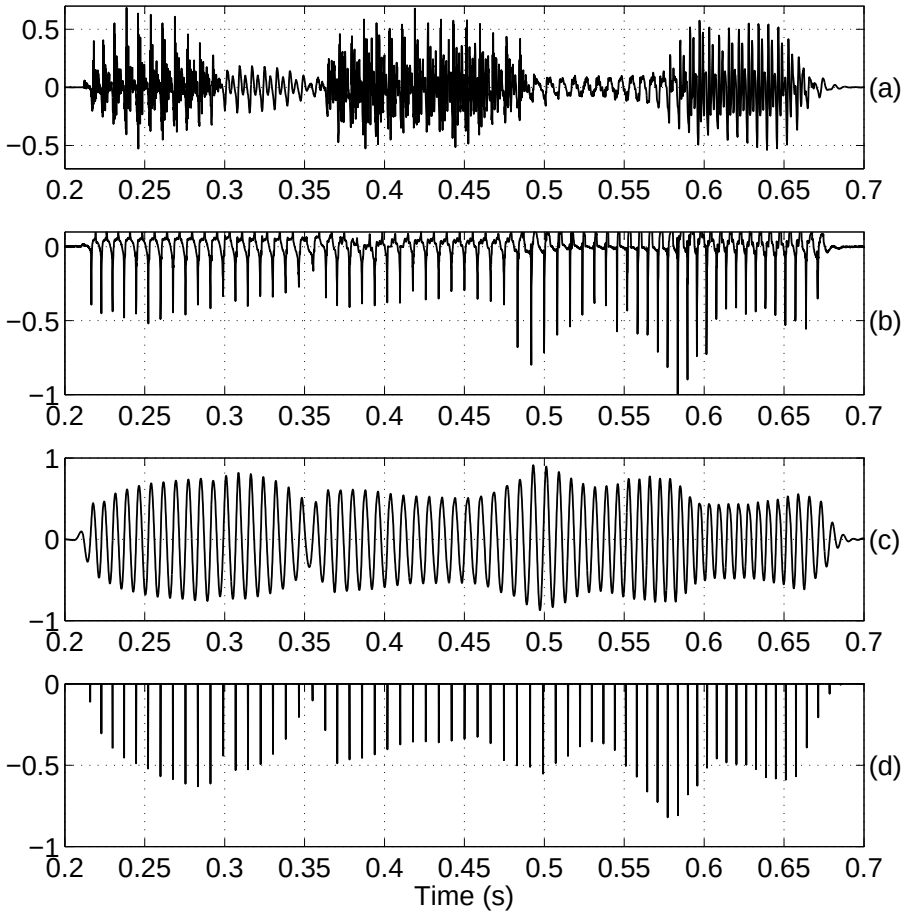
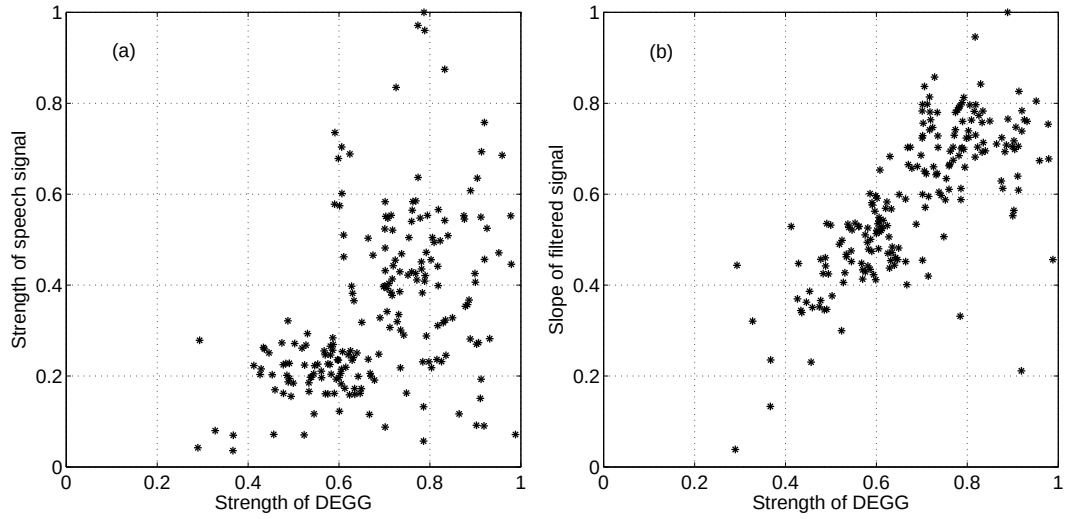


Fig. 4.3: Estimation of the strengths of excitation of the epochs from speech signal. (a) A segment of speech signal. (b) Differenced EGG signal. (c) Filtered signal. (d) Slopes of the filtered signal around detected epoch locations (sharper zero-crossings). The slopes are plotted as negative in order to compare with the differenced EGG signal.

the impulse-like excitation. Hence the filtered signal can be used to detect the regions of glottal activity (vocal fold vibration) as illustrated in Fig. 4.5. Fig. 4.5(a) shows a segment of speech signal with regions of glottal activity, marked by dotted lines, obtained from the differenced EGG signal in Fig. 4.5(b). The filtered signal of speech shown in Fig. 4.5(c) clearly indicates the regions of glottal activity, and they match well with those obtained from the differenced EGG signal in Fig. 4.5(b). Notice that the unvoiced regions around 0.6 s and 1.2 s in the speech signal (Fig. 4.5(a)) have very low amplitude in the filtered signal (Fig. 4.5(c)). Hence the short term energy of the filtered signal computer over 20ms frames, shown in Fig. 4.5(d), can be used for glottal activity detection (GAD). The short term energy of the filtered signal shows a clear indication of glottal activity even in noisy



(a) (b)

Fig. 4.4: Scatter plot of (a) negative peak amplitude of differenced EGG vs. absolute maximum amplitude of speech signal around the epoch location and (b) negative peak amplitude of differenced EGG vs. slope of the filtered signal at the epoch location.

speech signals. Fig. 4.6 shows a segment of speech signal degraded by babble noise at 5 dB SNR. It is difficult to identify the glottal activity around 0.4 s and 1.1 s directly from the degraded speech signal. However, the filtered signal shown in Fig. 4.6(c) enhances the regions of glottal activity over the unvoiced and noise regions. The short term energy of the filtered signal shown in Fig. 4.6(d) clearly shows large amplitude in the regions of glottal activity marked with dashed lines. The reference regions of glottal activity are manually marked by observing the differenced EGG signal shown in Fig. 4.6(b).

4.2.1 Performance evaluation of the proposed GAD

The proposed GAD method was evaluated under different noisy environments at varying levels of degradation. A subset of CMU-Arctic database [121] consisting of 100 randomly selected sentences from each of the 3 speakers was used to evaluate the proposed GAD method. The entire dataset was samplewise labeled for glottal activity using the simultaneously recorded EGG signals available with the database. All the signals were downsampled to 8 kHz.

To study the effect of noise on the proposed method for GAD, the method was eval-

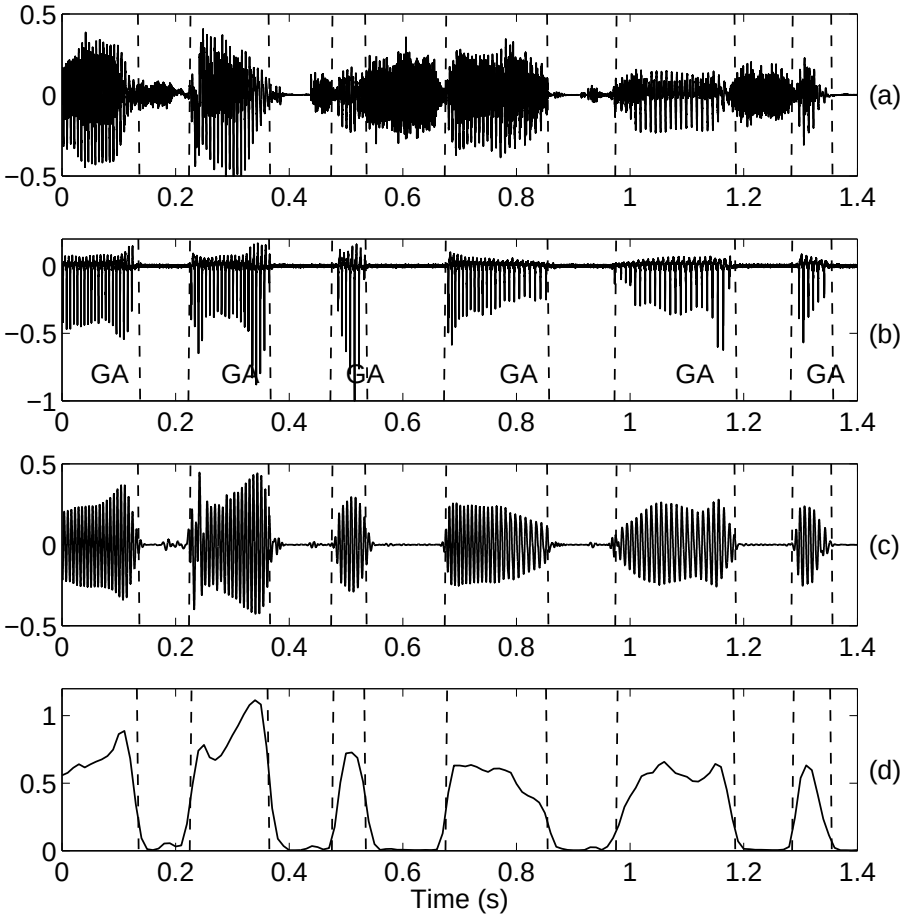


Fig. 4.5: Glottal activity detection from the filtered signal. (a) Speech signal. (b) Differenced EGG signal. (c) Filtered signal. (d) Energy computed over 20 ms segments of the filtered signal. Regions marked with GA in (b) indicate regions of glottal activity.

uated on artificially generated noisy speech data. Several noise environments at varying levels of degradation were simulated by adding noise taken from Noisex-92 database [123]. The utterances were appended with silence so that total duration of silence in each utterance is restricted to be about 60% of the data including pauses in the utterances. The database consists of speech signals under white, babble and vehicle noise environments at signal-to-noise ratio (SNR) ranging from 20 dB to 0 dB. The speech signals were processed using the proposed zero-frequency resonator to obtain the filtered signal. The energy of the filtered signal for every frame of 20 ms with 10 ms shift is used to detect the glottal activity.

The performance of the proposed GAD method was evaluated using the detection error tradeoff (DET) curves [126], which show the tradeoff between false alarm rate (FAR)

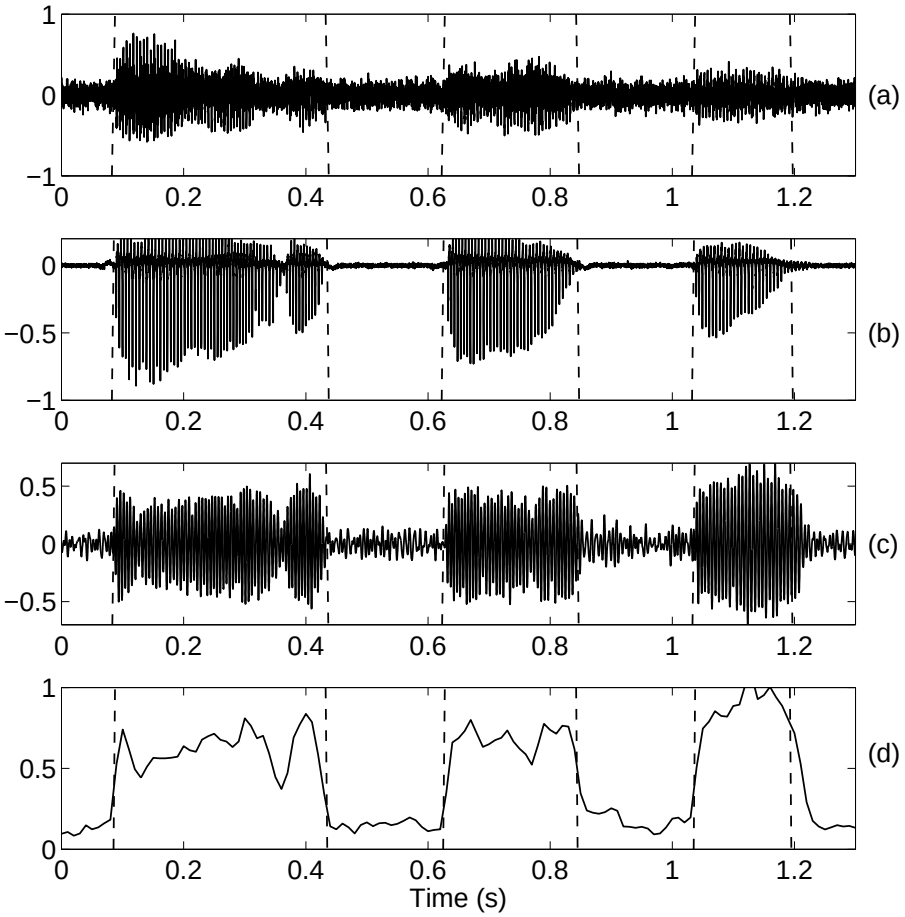


Fig. 4.6: Glottal activity detection under degraded conditions. (a) A segment of speech signal degraded by babble noise at 5 dB SNR. (b) Differenced EGG signal. (c) Filtered signal. (d) Energy of the filtered signal.

and false rejection rate (FRR). The FAR represents the percentage of nonglottal activity frames that were detected as glottal activity, and FRR represents the percentage of glottal activity frames that were detected as nonglottal activity. The performance of the system is expressed in terms of equal error rate (EER), the point at which FAR and FRR are equal. The lower the EER value, the higher is the accuracy of the GAD method. Fig. 4.7 shows the DET curves obtained for the proposed GAD algorithm under different noise environments at an SNR of 0 dB. The performance of GAD at varying levels of degradation is listed in Table 4.1 using the reference derived from the EGG signals.

The proposed method achieved an EER of 3.54% on the clean data, and exhibits a gradual degradation under noisy conditions. The performance of the method under babble noise and vehicle noise is inferior to that under white noise, because the babble noise

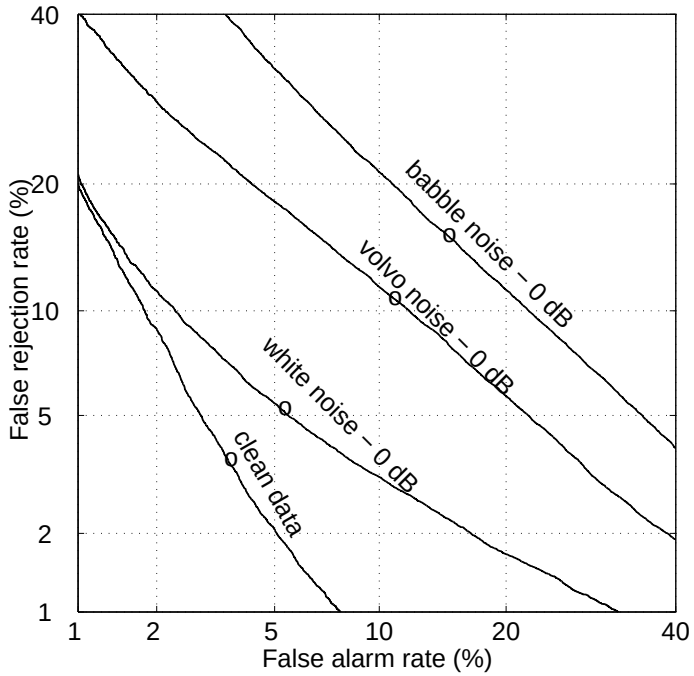


Fig. 4.7: DET curves indicating the performance of proposed GAD method under different noise environments.

Table 4.1: Performance of GAD in EER (%) under different noise environments at varying levels of degradation. Reference is derived from EGG signals.

Noise Type	20 dB	15 dB	10 dB	5 dB	0 dB
White	3.56	3.56	3.60	3.78	5.24
Babble	3.56	3.64	4.62	7.95	15.10
Vehicle	3.56	3.58	4.09	6.28	10.83

contains impulse-like excitations arising from epochs of other speakers, and the vehicle noise introduces high degradations in low frequency region. The errors on clean speech may be attributed to the errors in the reference which are a result of inability of the EGG signals in capturing the weak voiced regions. Fig. 4.8(a) and Fig. 4.8(b) show a segment of weak voiced region and the corresponding differenced EGG signal, respectively. The differenced EGG signal in Fig. 4.8(b) does not show prominent peaks around the epoch locations in the region from 1.26 s to 1.32 s, whereas the filtered signal in Fig. 4.8(c) clearly shows the glottal activity in that region, and the positive zero-crossings approximately coincide with the epoch locations. Similar observations can be made from the weak voiced segment from a female speaker shown in Fig. 4.9. Hence the proposed method can be

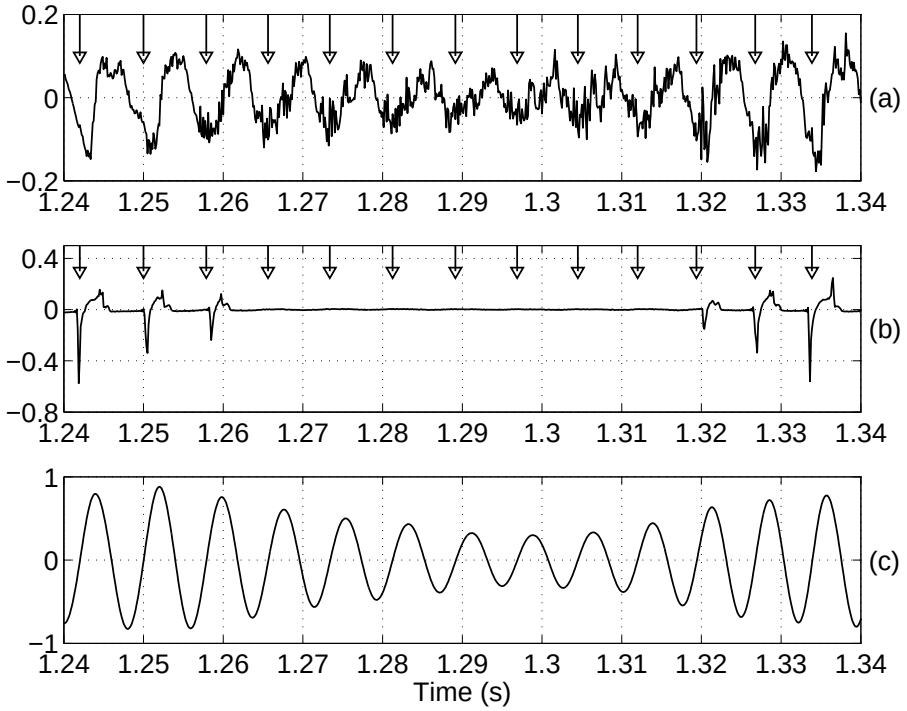


Fig. 4.8: Illustration of potential of proposed method in identifying weak voiced regions for a male speaker. (a) A segment of speech signal. (b) Differenced EGG signal. (c) Filtered signal. Arrows in (a) and (b) indicate the detected epoch locations.

Table 4.2: Performance of GAD in EER (%) under different noise environments at varying levels of degradation. Reference is derived from clean speech signals.

Noise Type	20 dB	15 dB	10 dB	5 dB	0 dB
White	0	0	0.003	0.41	2.77
Babble	0	0.23	1.81	6.13	14.14
Vehicle	0	0.006	1.08	4.22	9.66

effectively used to detect the glottal activity even in the weak voiced regions. The performance of the proposed GAD under different noisy environments is evaluated with the reference derived from the clean speech. Table 4.2 gives the performance of the proposed GAD at varying levels of degradation using the reference derived from the clean speech data. The results show that the performance of the proposed method for GAD is robust against different types of degradation.

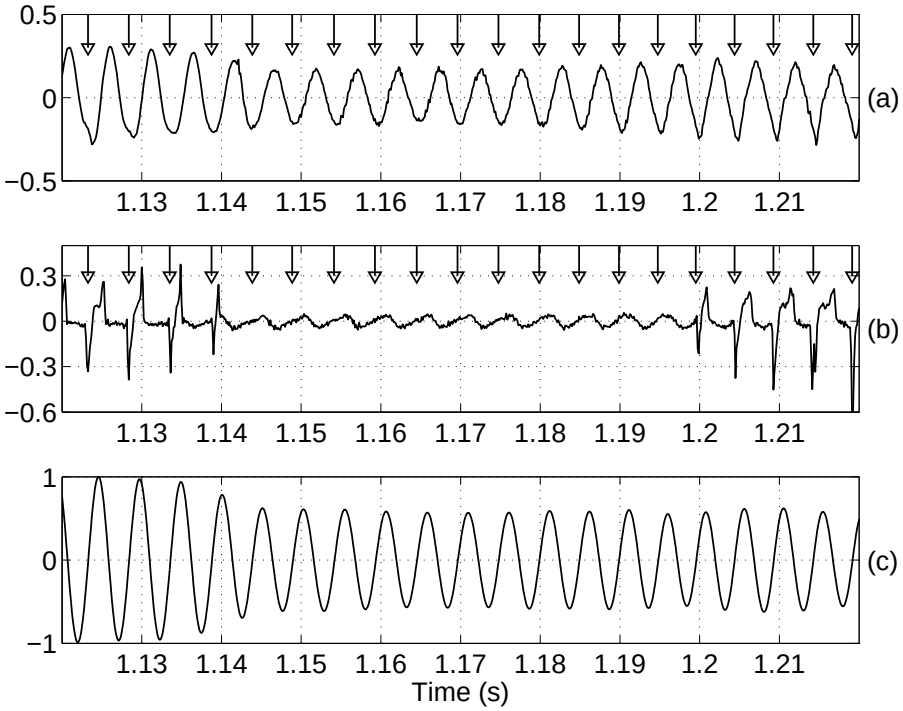


Fig. 4.9: Illustration of potential of proposed method in identifying weak voiced regions for a female speaker. (a) A segment of speech signal. (b) Differenced EGG signal. (c) Filtered signal. Arrows in (a) and (b) indicate the detected epoch locations.

4.3 Summary

In this chapter, we have proposed a method for detecting the regions of glottal activity and estimating the strength of excitation within each glottal cycle. The methods proposed in this chapter exploit the impulse-like characteristic of the excitation which is extracted using the zero-frequency resonator. The proposed method for estimating strength of excitation does not depend on estimating the vocal-tract response. Unlike conventional voicing detection methods, the proposed method for GAD does not assume periodicity of speech waveform in successive glottal cycles. The epoch location along with its strength of excitation form important features of a glottal pulse. These features may be useful in representing the excitation information in speech signal for speech coding and speech synthesis. The estimated strength of excitation may be useful in defining shimmer which is known to be a speaker-specific characteristic.

Chapter 5

Instantaneous Fundamental Frequency Estimation

Voiced sounds are produced from the time-varying vocal-tract system excited by a sequence of events caused by vocal fold vibrations. The vibrations of the vocal folds result in a sequence of glottal pulses with major excitation taking place around the instant of glottal closure (GCI). The rate of vibration of the vocal folds determines the fundamental frequency (F_0), and contributes to the perceived pitch of the sound produced by the vocal-tract system. Though the usage of the term "rate of vibration" gives an impression that the vibrations of the vocal folds are periodic, in practice the vocal fold vibrations at the glottis may or may not be periodic. Even a periodic vibration of the vocal folds at the glottis may produce a speech signal that is less correlated in successive cycles because of the time-varying vocal-tract system that filters the glottal pulses. Sometimes, the vocal fold vibrations at the glottis themselves may show aperiodic behavior, as in the case of changes in the shape of the glottal flow waveform (for example, the changes in the duty cycles of open/closed phases), or the intervals where the vocal fold vibration reflects several superposed periodicities (diplophony) [60], or where glottal pulses occur without obvious regularity in the time (glottalization, vocal fry or creaky voice) [127]. In practice, the rate of vibration of the vocal folds changes from one glottal cycle to the next cycle. Hence, it is more appropriate to define instantaneous fundamental frequency of excitation source for every glottal cycle. In this work, we propose an event-based approach to ac-

curately estimate the instantaneous fundamental frequency from speech signals. Epochs derived using zero-frequency resonator are used as anchor points within each glottal cycle for pitch estimation.

This chapter is organized as follows: In Section 5.1, the basis for the proposed method of fundamental frequency estimation is discussed. In Section 5.2, a method for pitch extraction from the speech signals is developed. In Section 5.3 the proposed method is compared with some standard methods for pitch extraction on standard databases, for which the ground truth is available in the form of electroglottograph (EGG) waveforms. The performance of the proposed method is also evaluated for different cases of simulated degradations in speech. Finally in Section 5.4, a summary of the ideas presented in this chapter is given along with some issues that need to be addressed while dealing with speech signals in practical environments.

5.1 Basis for the proposed method of pitch estimation

As mentioned earlier, voiced speech is the output of the time-varying vocal-tract filter excited by a sequence of glottal pulses caused by vocal fold vibrations. The vocal-tract system modulates the excitation source by formant frequencies, which depend on the sound unit being generated. The formant frequencies together with the fundamental frequency form important features of the voiced speech. There is an important distinction in the production of a formant frequency and in the production of the fundamental frequency. Formant frequencies are due to resonances of the vocal-tract system. The frequency of the resulting damped sinusoids are controlled by the size and the shape of the vocal-tract through the movement of the articulators. Because of the sinusoidal nature of the resonance, the formant frequency appears as a single impulse in the frequency domain. However, the fundamental frequency or pitch is produced as a result of vibration of the vocal folds, producing a sequence of regularly spaced impulses over short intervals of time. Periodic sequence of impulses in the time domain results in a periodic sequence of impulses in the frequency domain also. Hence, unlike the formant frequency, the information about the fundamental frequency is spread across the frequency range. This redundancy of in-

formation about the fundamental frequency in the frequency domain makes it a robust feature for speech analysis. For example, this redundancy helps us in perceiving the pitch even when the fundamental frequency is not present in the speech signal (as in the case of telephone speech).

Speech production mechanism is designed in such a way that the energy in the higher (> 300 Hz) frequencies is produced in the form of formants, whereas the perception of low (< 300 Hz) frequencies is due to the sequence of glottal cycles. Note that it is physically impossible for a human being to produce a resonance frequency of 200 Hz or less because of the limited length of the vocal-tract. In fact, the perception of low frequency (< 200 Hz) is felt more due to the intervals between the impulses rather the presence of any low frequency components in the form of sinusoids. In other words, it is the strong discontinuities at these impulse locations in the sequence that is producing the low frequency effect in perception. Moreover, the information about the discontinuities is spread across all the frequencies including the zero-frequency. In this work, we use the method based on the zero-frequency resonator to derive the information about the impulse-like discontinuity in each glottal cycle. The derived sequence of impulse locations is used for estimating the fundamental frequency for each glottal cycle.

5.2 Fundamental frequency estimation from epochs

Fundamental period is the time elapsed between two successive glottal cycles, the reciprocal of which is referred to as fundamental frequency. Measurement of fundamental period requires identification of a well specified point within each glottal cycle to mark the starting point of the cycle. Since the instant of glottal closure is the most abrupt event in a glottal cycle, it is the most commonly used anchor point for measuring fundamental period [19]. In this work, we use the instants of glottal closure (epochs) extracted from the zero-frequency filtering, discussed in Chapter 3, as anchor points for measuring fundamental period. The fundamental period is measured as the time interval between two successive epochs, and its reciprocal is used as fundamental frequency. The proposed approach is based only on the point property of the epoch and it does not involve any block

processing. As a result, the proposed approach can measure the finer period-to-period variation in the fundamental frequency, which is an important source of naturalness in speech synthesis and voice conversion systems. Hence we call the measured quantity as instantaneous fundamental frequency as opposed to 'mean pitch' estimated by conventional periodicity-based block processing methods.

Fundamental frequency estimation methods are often associated with a voicing decision that is used to eliminate the unvoiced regions. In this method, the glottal activity detection discussed in Chapter 4 is used to detect the regions of vocal fold vibration. Since the glottal activity detection is based on the strength of excitation in each glottal cycle, accurate end-points of the voiced regions can be obtained by this method. Fig. 5.1 illustrates the proposed method for fundamental frequency estimation on a Mandarin utterance (collected from a female speaker) with fast pitch variations. The speech signal is passed through a cascade of two ideal zero-frequency resonators, and the local mean computed over the average pitch period is subtracted from the resonator output. Fig. 5.1(b) shows the resulting filtered signal for the speech signal shown in Fig. 5.1(a). The positive zero-crossings of the filtered signal indicate the epochs, and slopes of the filtered signal around the epochs give their strengths of excitation. The locations of the epochs along with their strengths are shown as pulses in Fig. 5.1(c). Notice that the strengths of the excitation are significantly high compared to the slopes of the spurious zero-crossings occurring in the unvoiced regions. Hence the strengths of the excitation are used to detect the voiced regions, and the time-interval between two successive epochs in the voiced regions is used to measure the fundamental frequency. Fig. 5.1(d) shows the fundamental frequency measured from the epoch locations in the voiced regions. The proposed method is able to measure the fast varying changes in the fundamental frequency accurately. The finer variations are due to cycle-to-cycle variations in pitch, which may be a speaker-specific characteristic.

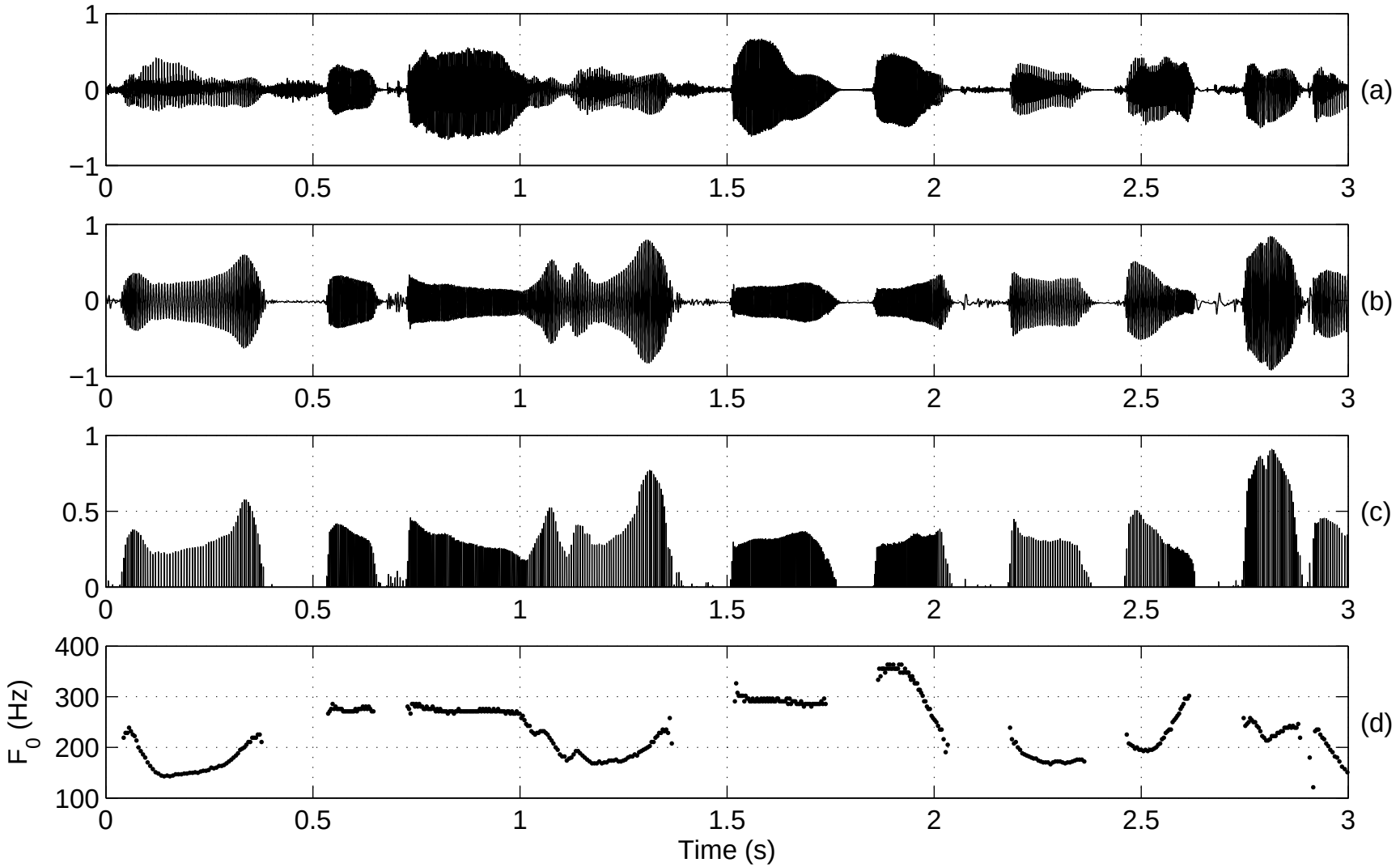


Fig. 5.1: Illustration of proposed method of fundamental frequency estimation on a Mandarin utterance with fast pitch variations. (a) Speech signal. (b) Filtered signal. (c) Epoch locations and their strengths of excitation. (d) Fundamental frequency measured from epoch locations. The unvoiced regions are eliminated using the GAD method discussed in Chapter 4.

5.2.1 Validation of F_0 estimates using Hilbert envelope

In the process of measuring the instantaneous fundamental period from the intervals of successive positive zero-crossings of the filtered signal, there could be errors due to spurious zero-crossings which occur mainly if there is another impulse in between two glottal closure instants. To reduce the effects due to spurious zero-crossings, the knowledge that the strength of the impulse is the strongest at the epoch in each glottal cycle may be used. In order to exploit the strength of impulses in the excitation for reducing the effects due to spurious zero-crossings, the Hilbert envelope of speech signal is computed. The Hilbert envelope $h[n]$ is computed from the speech signal $s[n]$ as follows:

$$h[n] = \sqrt{s^2[n] + s_h^2[n]}, \tag{5.1}$$

where $s_h[n]$ is the Hilbert transform of $s[n]$, and is given by

$$s_h[n] = \text{IDFT}[S_h(\omega)], \tag{5.2}$$

where

$$S_h(\omega) = \begin{cases} +jS(\omega), & \omega < 0 \\ -jS(\omega), & \omega > 0, \end{cases} \tag{5.3}$$

and

$$S(\omega) = \text{DFT}[s[n]]. \tag{5.4}$$

Here DFT and IDFT refer to the discrete Fourier transform and inverse discrete Fourier transform, respectively.

The Hilbert envelope contains a sequence of strong impulses around the glottal closure instants, and may also contain some spurious impulses at other places due to the formant structure of the vocal-tract, and the secondary excitations in the glottal cycles. But, the amplitude of the impulses around the glottal closure instants dominate over those of the spurious impulses in the computation of the filtered signal. Hence, the filtered signal of the Hilbert envelope mainly contains the zero-crossings around the instants of glottal closure. However, the zero-crossings derived from the filtered signal of Hilbert envelope deviate

slightly (around 0.5 ms to 1 ms) from the actual locations of the instants of glottal closure. In other words, the zero-crossings derived from the filtered signal of Hilbert envelope are not as accurate as those derived from the filtered signal of speech signal. Hence, the accuracy of the zero-crossings derived from the filtered signal of speech, and the robustness of the zero-crossings derived from the Hilbert envelope are used in conjunction to obtain an accurate and robust estimate of the instantaneous fundamental frequency.

The instantaneous pitch frequency contour obtained from the filtered signal of speech is used as the primary pitch contour, and the errors in the contour are corrected using the pitch contour derived from the Hilbert envelope of the speech signal. The pitch frequency contours are obtained from the zero-crossings of the filtered signals for every 10 ms. The value of 10 ms is chosen for comparison with the results from other methods. Let $p_s[m]$ and $p_h[m]$ be the pitch frequency contours derived, respectively, from the speech signal and the Hilbert envelope of the speech signal. The following logic is used to correct the errors in $p_s[m]$:

$$p[m] = \begin{cases} p_h[m], & \text{if } p_s[m] > 1.5 p_h[m] \\ p_s[m], & \text{otherwise,} \end{cases} \tag{5.5}$$

where m is the frame index for every 10 ms and $p[m]$ is the corrected pitch contour. The factor 1.5 is used mainly to reduce the pitch doubling errors in $p_s[m]$ due to spurious zero-crossings. Any value between 1.3 to 1.8 is adequate to perform this correction.

The significance of using the pitch contour $p_h[m]$ to correct the errors in the contour $p_s[m]$ is illustrated in Fig. 5.2. The filtered signal shown in Fig. 5.2(c) is obtained from the speech segment shown in Fig. 5.2(a). It contains spurious zero-crossings around 0.1 s to 0.2 s due to small values of the strength of excitation in this region. On the other hand, the pitch derived from the Hilbert envelope gives the correct zero-crossings. The main idea of this logic is to correct the errors due to spurious zero-crossings occurring in the filtered signal derived from the speech signal. Steps involved in measuring instantaneous fundamental frequency from speech signals are given in Table 5.1.

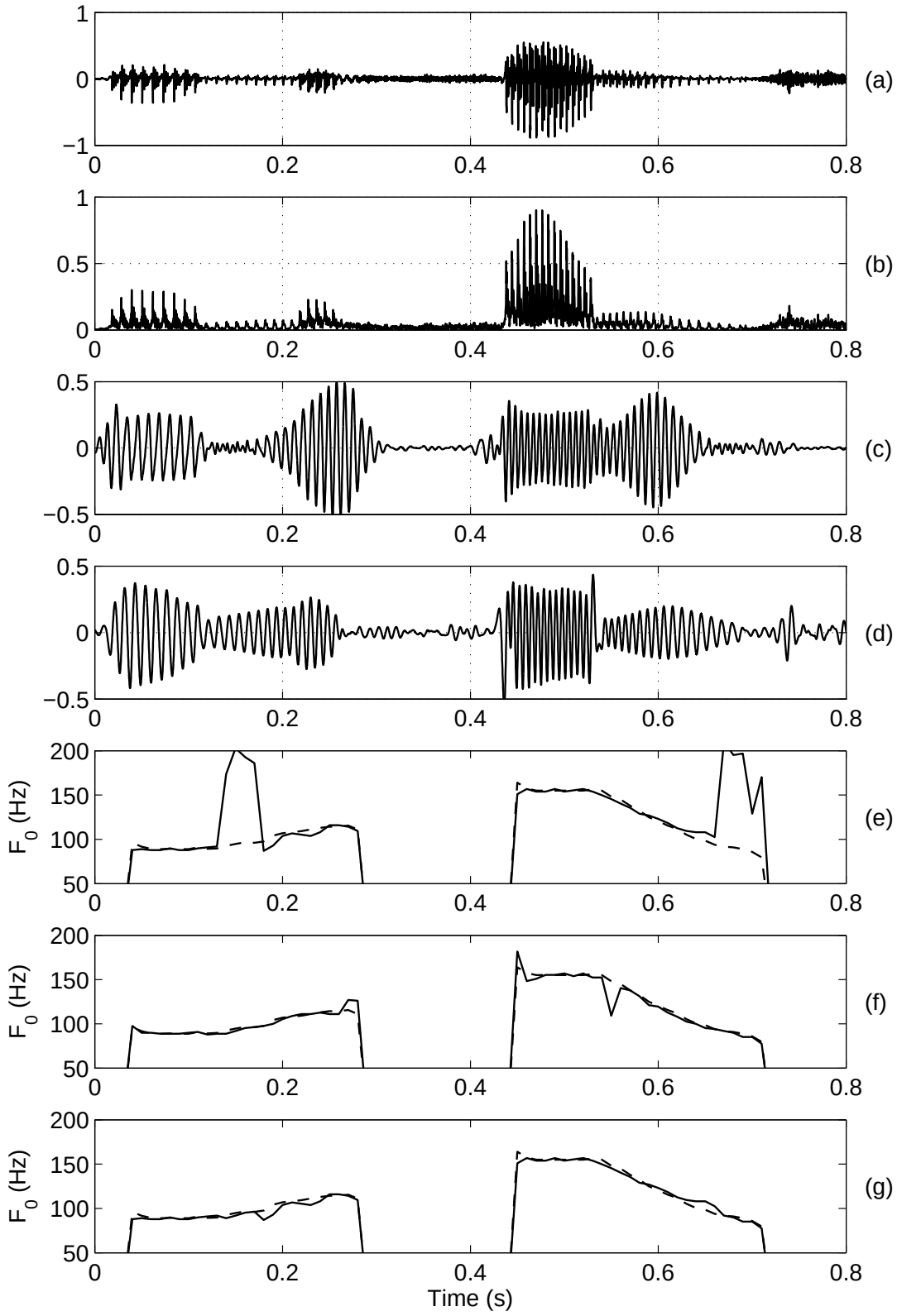


Fig. 5.2: Correcting the pitch contour obtained from speech signal using the pitch contour obtained from Hilbert envelope. (a) Speech signal. (b) Hilbert envelope of the speech signal. Zero-frequency filtered signal derived from (c) speech signal, and (d) Hilbert envelope. Fundamental frequency derived from (e) filtered speech signal, (f) filtered Hilbert envelope, (g) correction suggested in (5.5). The dashed lines in the figures indicate the ground truth given by the EGG signals.

Table 5.1: Steps in computation of instantaneous fundamental frequency from speech signals

(a) Compute the differenced speech signal $x[n]$.

(b) Compute the average pitch period using the histogram of the pitch periods estimated from autocorrelation of 30 ms speech segments.

(c) Compute the output $y_2[n]$ of the cascade of two zero-frequency resonators.

(d) Compute the filtered signal $y[n]$ from $y_2[n]$ using a window length corresponding to the average pitch period.

(e) Compute the instantaneous fundamental (pitch) frequency from the positive zero-crossings of the filtered signal.

(f) Obtain the pitch contour $p_s[m]$ for every 10 ms from the instantaneous pitch frequency.

(g) Compute the Hilbert envelope $h[n]$ of speech signal $s[n]$.

(h) Compute the pitch contour $p_h[n]$ from the filtered signal of $h[n]$.

(i) Replace the value in $p_s[m]$ with $p_h[m]$ whenever $p_s[m] > 1.5p_h[m]$.

5.3 Performance evaluation and comparison with other pitch extraction methods

In this section, the proposed method of extracting the instantaneous fundamental frequency from the speech signals is compared with four existing methods in terms of accuracy in estimation and in terms of robustness against degradation. The four methods chosen for comparison are Praat's autocorrelation method [64], crosscorrelation method [128], subharmonic summation [77], and a fundamental frequency estimator (YIN) [127]. Initially the fundamental frequency estimation algorithms are evaluated on clean data. Subsequently, the robustness of the proposed method and the four existing methods are evaluated at different levels of degradation by white noise, babble noise and vehicle noise. A brief description of the implementation details of the four methods chosen for comparison is given below. The software program codes for implementing these methods are available at the respective websites, and are used in this study for evaluation.

5.3.1 Existing methods for fundamental frequency estimation

Praat's autocorrelation method (AC) [64]: The Praat's algorithm performs an acoustic periodicity detection on the basis of an accurate autocorrelation method. This method is more accurate and robust than the cepstrum-based methods and original autocorrelation-based method [64]. It was pointed out that sampling and windowing the data cause problems in determining the peak corresponding to the fundamental period in the autocorrelation function. In this method, the autocorrelation of the original signal segment $r_x[\tau]$ is computed by dividing the autocorrelation of the windowed signal $r_a[\tau]$ with the autocorrelation of the window $r_w[\tau]$. That is,

$$r_x[\tau] = \frac{r_a[\tau]}{r_w[\tau]}. \tag{5.6}$$

This correction does not let the autocorrelation function $r_x[\tau]$ taper to zero as the lag increases, which helps in identification of the peak corresponding to the fundamental period. To overcome the artifacts due to sampling, the algorithm employs a *sinc* interpolation around the local maxima. The interpolation provides an estimation of the fundamental period. The software code for implementation of this algorithm is available at `http://www.fon.hum.uva.nl/praat/` [129].

Crosscorrelation method (CC) [128]: In the computation of the autocorrelation function, fewer samples are included as the lag increases. This effect can be seen as the roll-off of the autocorrelation values for the higher lags. The values of the autocorrelation function at higher lags are important, especially for low-pitched male voices. For a 50 Hz pitch, the lag between successive pitch pulses is 200 samples at a sampling frequency of 10 kHz. To overcome this limitation in the computation of the autocorrelation function, a crosscorrelation function that operates on two different data windows is used. Each value of the crosscorrelation function is computed over the same number of samples. A software implementation of this algorithm is available with the Praat system [129].

Subharmonic summation (SHS) [77]: Subharmonic summation performs pitch analysis based on a spectral compression model. Since a compression on a linear scale corresponds to a shift on a logarithmic scale, the spectral compression along the linear fre-

quency abscissa can be substituted by shifts along the logarithmic frequency abscissa. This model is equivalent to the concept that each spectral component activates not only those elements of the central pitch processor, but also those elements that have a lower harmonic relation with this component. For this reason, this method is referred to as the subharmonic summation method. The contributions of various components add up, and the activation is the highest for the frequency sensitive element that is most activated by its harmonics. Hence, the maximum of the resulting sum spectrum gives an estimate of the fundamental frequency. A software implementation of this algorithm is available with the Praat system [129].

The fundamental frequency estimator, YIN [127]: The fundamental frequency estimator, YIN [127], was developed by Alain de Cheveigne and Hideki Kawahara, is named after the oriental yin-yang philosophical principle of balance. In this algorithm, the authors attempt to balance between the pitch peak in the autocorrelation function and cancellation of the secondary peaks due to harmonics. The difficulty with autocorrelation-based methods is that the peaks occur at multiples of the fundamental period also, and it is sometimes difficult to determine which peak corresponds to the true fundamental period. The YIN method attempts to solve these problems in several ways. It is based on a difference function, that attempts to minimize the difference between the waveform and its delayed duplicate, instead of maximizing the product as in autocorrelation. The difference function is given by

$$d[\tau] = \sum_{n=1}^{N} (s[n] - s[n+\tau])^2 \tag{5.7}$$

In order to reduce the occurrence of subharmonic errors, YIN employs a cumulative mean function which deemphasizes higher period valleys in the difference function. The cumulative mean function is given by

$$\hat{d}[\tau] = \begin{cases} 1, & \tau = 0 \\ \frac{d[\tau]}{\frac{1}{\tau}\sum_{k=1}^{\tau} d[k]}, & \text{otherwise.} \end{cases} \tag{5.8}$$

The YIN method also employs a parabolic interpolation of the local minima, which has the effect of reducing the errors when the estimated pitch period is not a factor of the window length. The Matlab code for implementation of this algorithm is available at

http://www.auditory.org/postings/2002/26.html [130].

5.3.2 Databases for evaluation

Keele database: The Keele pitch extraction reference database [131][132] is used to evaluate the proposed method, and to compare with the existing methods. The database includes speech data from five male and five female speakers, each speaking a short story of about 35 s duration. All the speech signals were sampled at a rate of 20 kHz. This database provides a reference pitch for every 10 ms, which is obtained from a simultaneously recorded EGG signal, and is used as the *ground truth*. Pitch values are provided at a frame rate of 100 Hz using a 25.6 ms window. Unvoiced frames are indicated with zero pitch values, and negative values are used for uncertain frames.

CSTR database: The CSTR database [133] [134] consists of fifty sentences, each read by one adult male and one adult female, both with non-pathological voices. The database contains approximately five minutes of speech. The speech is recorded simultaneously with a close-talking microphone and a electroglottograph in an anechoic chamber. The database is biased towards utterances containing voiced fricatives, nasals, liquids and glides. Since some of these phones are aperiodic in comparison to vowels, standard pitch estimation methods find them difficult to analyze. In this database, the reference pitch values are provided at the instants of glottal closure.

5.3.3 Evaluation procedure

The performance of the existing as well as the proposed pitch estimation algorithms is evaluated on both Keele database and CSTR database. All the signals are downsampled to 8 kHz for this evaluation. All the methods are evaluated using a search range of 40 Hz to 600 Hz (typical pitch frequency range of human beings). The post-processing and voicing detection mechanisms of the existing algorithms are disabled (wherever applicable) in this evaluation.

The accuracy of pitch estimation methods is measured according to the following

criteria [60]:

- *Gross error (GE):* It is percentage of voiced frames with an estimated F_0 value that deviates from the reference value by more than 20%.
- *Mean absolute error (M):* It is the mean of the absolute value of the difference between the estimated and the reference pitch values. Gross errors are not considered in this calculation.
- *Standard deviation (SD):* It is the standard deviation of the difference between estimated and reference pitch values. Gross errors are not considered in this calculation.

The reference estimates as provided in the databases are used for evaluating the pitch estimation algorithms. The reference estimates are time-shifted and aligned with the estimates of each of the methods. The best alignment is determined by taking the minimum error, over a range of time-shifts, between the estimates derived from the speech signal and the ground truth [127]. This compensation for time-shift is required due to acoustic propagation delay from glottis to microphone, and/or due to the differences in the implementations of the algorithms.

The gross estimation errors, the mean absolute errors and the standard deviation of errors of different algorithms for fundamental frequency estimation are given in Table 5.2. In the table, the performances of pitch contours derived from $p_s[m]$, $p_h[m]$ and $p[m]$ are also given. Most of the time, the percentage gross errors for the proposed method are significantly lower than the percentage gross errors for other methods. The results clearly demonstrate the effectiveness of the proposed method over other methods. Note that the proposed method is based on the strength of the impulse-like excitation, and it does not depend on the periodicity of the signal in successive glottal cycles. The method does not use any averaging or smoothing of the estimated values over a longer segment consisting of several glottal cycles.

The potential of the proposed method in estimating the instantaneous fundamental frequency from the speech signals is illustrated in Fig. 5.3. The segment of voiced

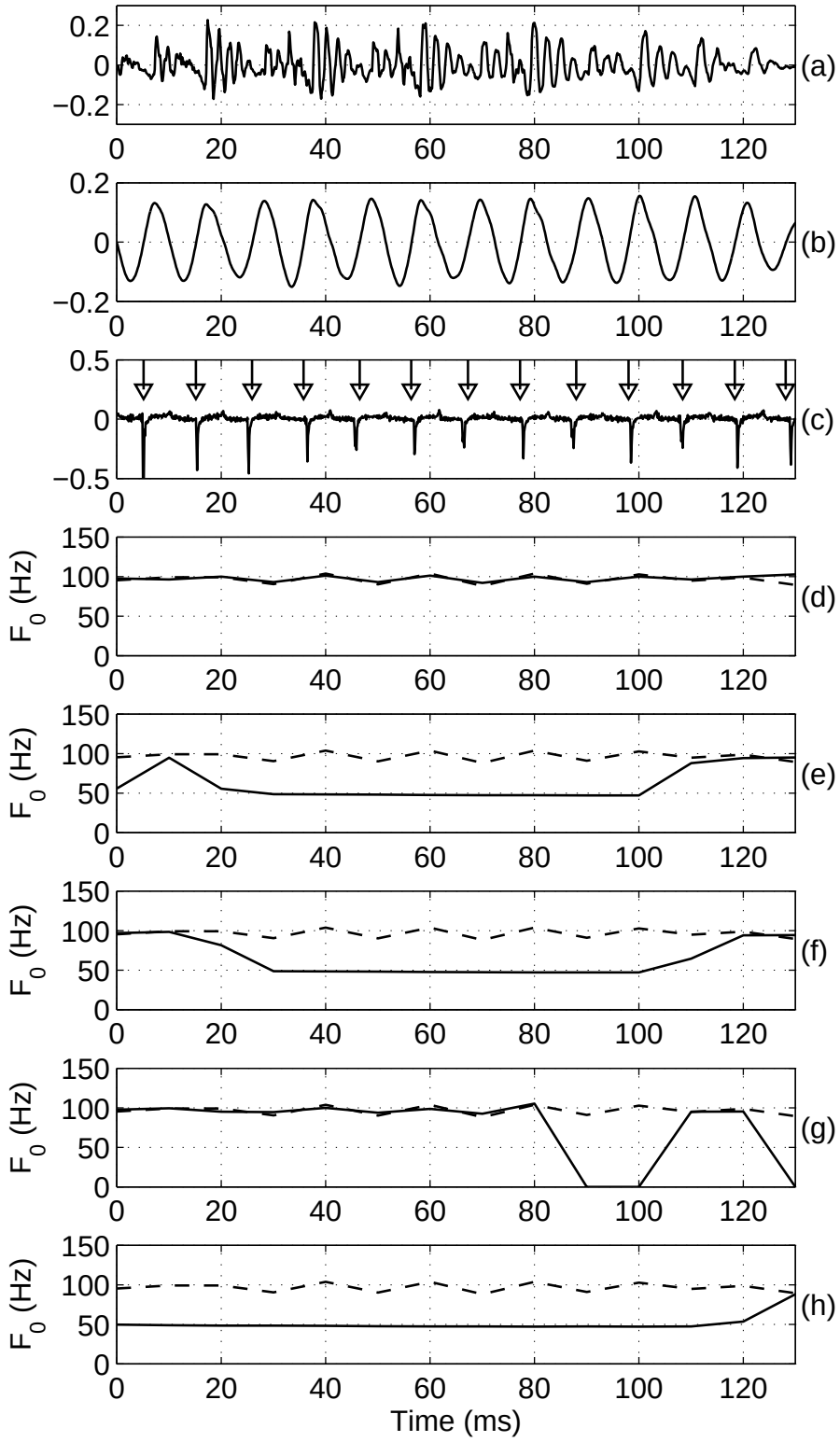


Fig. 5.3: Potential of the proposed method in estimating the instantaneous fundamental frequency. (a) Speech signal. (b) Zero frequency filtered signal. (c) Differenced EGG signal. Arrows indicate the positive zero-crossings of the zero-frequency filtered signal. Fundamental frequency derived from (d) proposed method, (e) Praat's autocorrelation method, (f) cross correlation method, (g) subhormonic summation, and (h) YIN method. The dashed lines in the figures correspond to the reference pitch contour (i.e., ground truth).

Table 5.2: Performance of algorithms for fundamental frequency estimation on clean data. $p_s[m]$ denotes the pitch contour derived from filtered speech signal. $p_h[m]$ denotes the pitch contour derived from filtered Hilbert envelope. $p[m]$ denotes the pitch contour obtained by combining evidence from $p_s[m]$ and $p_h[m]$ (5.5).

	Keele Database			**CSTR Database**		
Method	GE (%)	M (Hz)	SD (Hz)	GE (%)	M (Hz)	SD (Hz)
AC	5.345	2.656	3.694	5.238	4.777	6.820
CC	6.891	2.201	3.371	6.818	5.108	6.730
YIN	3.219	2.165	2.906	3.073	4.922	6.584
SHS	10.774	1.868	2.398	8.938	4.108	5.864
$p_s[m]$	2.935	3.198	4.555	3.394	5.459	6.974
$p_h[m]$	5.647	4.562	6.381	4.157	5.699	6.886
$p[m]$	2.603	3.207	4.473	1.943	5.367	6.801

speech in Fig. 5.3(a) is not periodic. The signal shows more similarity between alternate periods, than between adjacent periods. It is only through the analysis of the differenced EGG signal (Fig. 5.3(c)), the actual pitch periods could be observed. The correlation-based methods fail to estimate the actual fundamental frequency of the speech segment in these cases. On the other hand, the positive zero-crossings of the filtered signal clearly show the actual glottal closure instants.

5.3.4 Evaluation under noisy conditions

In this section we study the effect of noise on the accuracy of pitch estimation algorithms. The existing methods and the proposed method were evaluated on an artificially generated noisy speech database. The noisy environment conditions were simulated by adding noise to the original speech signal at different signal-to-noise ratios (SNRs). The noise signals were taken from Noisex-92 database [123]. Three noise environments, namely, white Gaussian noise, babble noise and vehicle noise, were considered in this study. The utterances were appended with silence so that the total amount of silence in each utterance is constrained to be about 60% of data, including the pauses in the utterances. The resulting data consist of about 40% speech samples, which is the amount of speech activity in

a typical telephone conversation. The noise from Noisex-92 database is added to both Keele database and CSTR database to create the noisy data at SNR levels ranging from 20 dB to -5 dB.

Table 5.3 shows the gross estimation errors for different pitch estimation algorithms on the Keele database and CSTR database at varying levels of degradation by white noise. The performance of the correlation-based methods is similar, and is reasonable at low noise levels (upto an SNR of 10 dB). But for higher levels of degradation, the estimation errors increase dramatically for all the systems, except for the proposed method, where the degradation in performance is somewhat gradual. Robustness of the proposed method to noise can be attributed to the impulse-like nature of the glottal closure instants in the speech signal. The energy of white noise is distributed both in time and frequency domains. While the energy of an impulse is distributed across the frequency range, it is highly concentrated in the time domain. Therefore the zero-crossing due to an impulse is unaffected in the output of the zero-frequency resonator even in the presence of high levels of noise. Fig. 5.4 illustrates the robustness of the proposed method in estimating the instantaneous fundamental frequency under noisy conditions. Fig. 5.4(a) and Fig. 5.4(b) show the waveforms of a weakly voiced sound under clean and degraded (by white noise at 0 dB SNR) conditions, respectively. Fig. 5.4(c) and Fig. 5.4(d) show the zero-frequency filtered signals derived from the clean (Fig. 5.4(a)) and the noisy signals (Fig. 5.4(b)), respectively. Though the individual periods can be observed from the clean signal in Fig. 5.4(a), it is difficult to observe any periodicity in the noisy signal shown in Fig. 5.4(b). But the zero-crossings of the filtered signal derived from the noisy waveform, remain almost the same as those derived from the clean signal, illustrating the robustness of the proposed method.

Fig. 5.5 illustrates the performance of the proposed method under noisy conditions, compared to the performance of the other methods. A segment of speech signal, degraded by white noise, at 0 dB SNR is shown in Fig. 5.5(a). The estimated pitch contour from the proposed method is given in Fig. 5.5(d), where the estimated values match well with the reference pitch values or ground truth (shown by dashed curves). The errors in the estimated pitch (deviation from the ground truth) can be seen clearly in all the other four

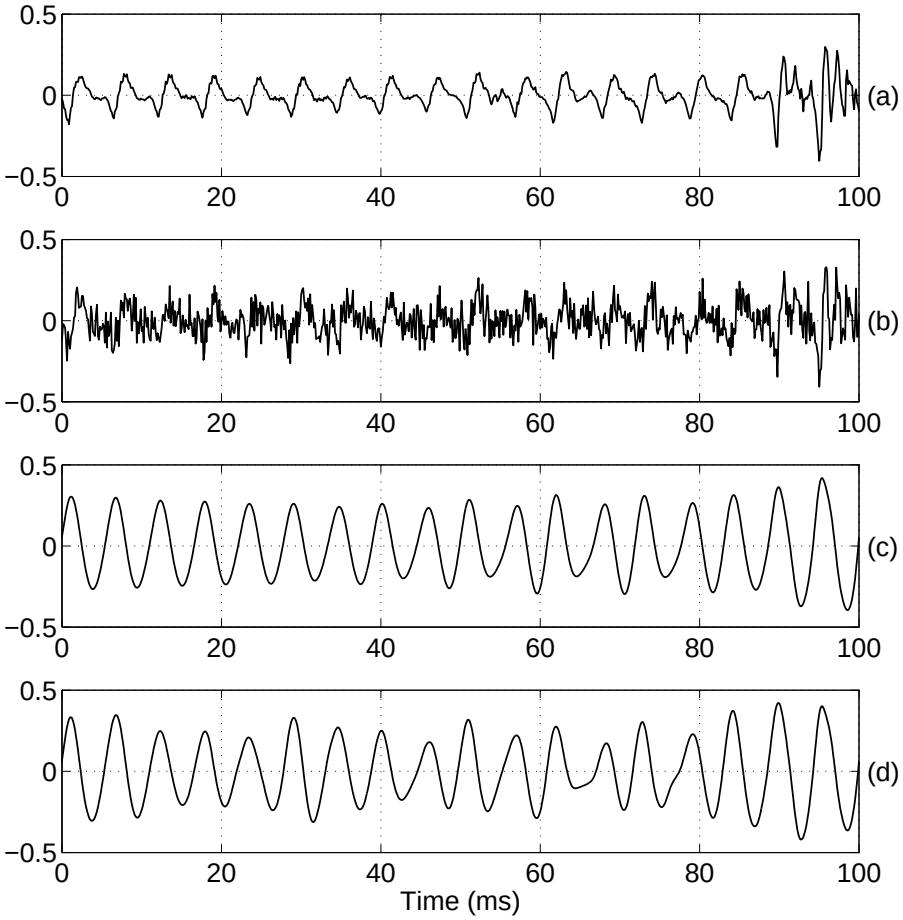


Fig. 5.4: Comparison of filtered signals derived from clean and degraded speech signals. (a) Speech signal of a weakly voiced sound. (b) Speech signal degraded by white noise at 0 dB SNR. (c) Filtered signal derived from clean signal in (a). (d) Filtered signal derived from noisy signal in (b).

methods used for comparison . Since the other methods depend mostly on the periodicity of the signal in successive glottal cycles, the periodicity of the signal waveform is affected by noise and hence the accuracy. Even for clean signal, there may be regions where the signal is far from periodic in successive glottal cycles, and hence there are more errors in comparison to the proposed method as can be seen in Table 5.2. Note that the proposed method does not use any knowledge of the periodicity of the speech signal, nor assumes regularity of the glottal cycles. Therefore there is scope for further improvement in the accuracy of the pitch estimation by combining the proposed method with methods based on autocorrelation.

Table 5.4 and Table 5.5 show the performance of all the five pitch estimation methods under speech-like degradation as in babble noise and low frequency degradation as in

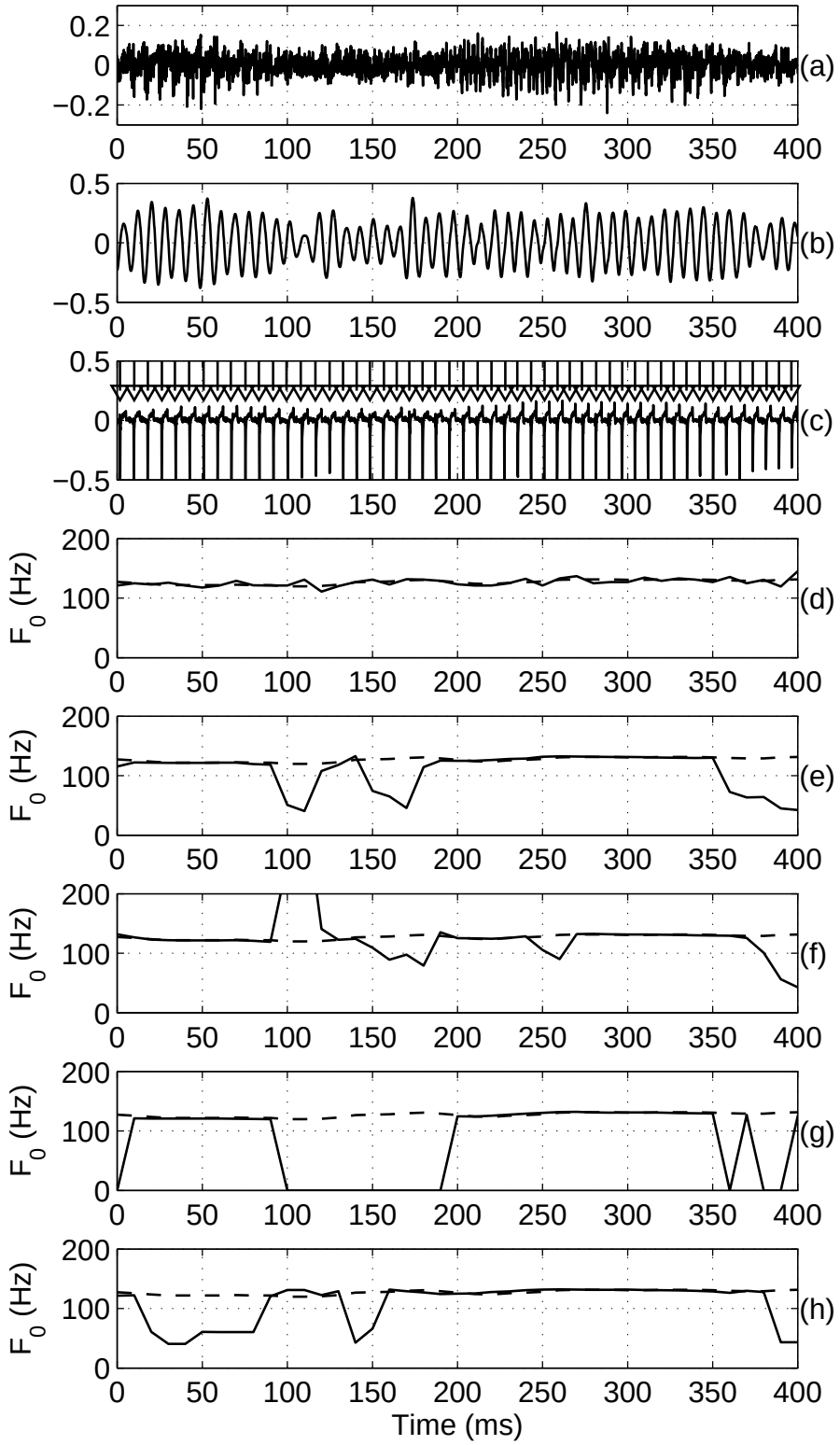


Fig. 5.5: Robustness of fundamental frequency estimation algorithms under noisy conditions. (a) Speech signal degraded by white noise at 0 dB SNR, (b) Zero frequency filtered signal, (c) Differenced EGG signal, arrows indicate the positive zero-crossings of the filtered signal in (b). F_0 derived from (d) Proposed method, (e) Praat's autocorrelation method, (f) crosscorrelation method, (g) Subharmonic summation and (h) YIN method. The dashed lines in the figures correspond to the reference pitch contour.

Table 5.3: Gross estimation errors (in %) for different pitch estimation algorithms at varying levels of degradation by white noise.

	Keele Database					**CSTR Database**				
SNR	AC	CC	YIN	SHS	Proposed	AC	CC	YIN	SHS	Proposed
Clean	5.345	6.891	3.219	10.774	2.603	5.238	6.818	3.073	8.938	1.943
20 dB	5.580	7.012	3.352	11.366	2.832	5.319	6.900	3.081	9.432	1.959
15 dB	5.756	7.320	3.400	12.085	3.116	5.626	7.131	3.139	9.981	2.211
10 dB	6.655	9.065	4.058	14.313	3.346	5.972	8.100	3.366	11.462	2.256
5 dB	9.173	13.462	5.955	19.562	3.907	6.249	12.287	4.933	14.868	3.069
0 dB	15.340	21.85	12.876	30.994	5.768	14.505	21.191	12.885	22.820	5.019
-5 dB	28.373	36.043	26.223	50.115	10.188	26.809	34.876	28.582	40.691	10.530

vehicle noise. The performance of the proposed method is comparable to or better than the other methods even for these two types of degradation.

The performance of the proposed method under babble noise (Table 5.4) and vehicle noise (Table 5.5) is inferior to its performance under white noise (Table 5.3). This is because the effect of degradation due to white noise is uniformly distributed in the frequency domain, and does not introduce any impulse-like discontinuities in the time domain. The degradation due to vehicle noise is mostly concentrated in the low frequency region (0–300 Hz), from which the epoch information is derived using the proposed method. Hence the vehicle noise affects the locations of zero-crossings in the filtered signal, resulting in a performance degradation. In the case of babble noise, the impulse-like degradations due to epochs of the background speakers introduces spurious zero-crossings in the filtered signal. The spurious zero-crossings lead to high gross errors in the proposed method.

5.4 Summary

In this chapter, we have proposed a method for extracting the instantaneous fundamental frequency from the speech signal. The proposed method exploits the impulse-like characteristic of excitation in the glottal vibrations for producing voiced speech. Since an impulse sequence has energy at all frequencies, the zero-frequency resonator proposed in Chapter 3 was used to derive the instant of significant excitation in each glottal cycle.

Table 5.4: Gross estimation errors (in %) for different pitch estimation algorithms at varying levels of degradation by babble noise.

	Keele Database					**CSTR Database**				
SNR	AC	CC	YIN	SHS	Proposed	AC	CC	YIN	SHS	Proposed
Clean	5.345	6.891	3.219	10.774	2.603	5.238	6.818	3.073	8.938	1.943
20 dB	5.635	7.501	3.624	12.061	3.147	5.597	7.238	3.233	10.026	2.268
15 dB	6.613	8.860	4.705	13.921	3.781	6.653	8.938	3.629	11.713	2.640
10 dB	9.246	12.900	7.356	17.895	5.158	10.513	14.438	7.007	15.330	3.720
5 dB	16.155	21.579	15.745	26.35	8.618	19.438	24.400	18.947	24.177	7.205
0 dB	29.086	35.795	31.852	42.559	16.149	36.072	41.879	41.788	41.232	15.038
-5 dB	45.114	50.211	48.714	62.840	28.530	54.854	60.430	63.685	62.307	30.141

Table 5.5: Gross estimation errors (in %) for different pitch estimation algorithms at varying levels of degradation by vehicle noise.

	Keele Database					**CSTR Database**				
SNR	AC	CC	YIN	SHS	Proposed	AC	CC	YIN	SHS	Proposed
Clean	5.345	6.891	3.219	10.774	2.603	5.238	6.818	3.073	8.938	1.943
20 dB	5.333	6.891	3.358	11.215	2.941	5.040	6.607	3.060	9.169	2.046
15 dB	5.550	7.428	3.708	12.067	3.104	5.069	6.372	3.184	9.701	2.281
10 dB	6.504	8.763	4.457	14.102	3.920	5.164	6.479	3.514	11.099	3.007
5 dB	9.886	13.196	7.893	18.227	6.081	6.756	8.191	5.576	14.147	5.551
0 dB	17.689	21.669	14.246	25.583	10.509	10.695	13.091	10.867	20.770	10.884
-5 dB	32.564	35.934	27.956	39.950	20.304	19.904	23.431	23.909	34.402	18.89

The method does not depend on the periodicity of glottal cycles, nor does it rely on the correlation of the speech signal in successive pitch periods. Thus the method extracts the instantaneous fundamental frequency given by the reciprocal of the interval between successive glottal closure instants. Errors occur when the strength of excitation around the instant of glottal closure is not high. To correct these errors, the pitch period information derived from the zero-frequency resonator output is modified based on the pitch period information derived from the Hilbert envelope of the speech signal using the proposed method. The method gives a better accuracy in comparison with standard algorithms for pitch estimation. Moreover, the method was shown to be robust even under low signal-to-noise ratio conditions.

The proposed method depends only on the impulse-like excitation in each glottal cycle, and hence the intervals between successive glottal cycles are obtained without using the periodicity property in the time domain, or the harmonic structure in the frequency domain. Since the correlation of the speech signal in successive glottal cycles is not used, the method is robust even when there are rapid changes in the successive periods of excitation, and also when there are rapid changes in the vocal-tract system, as in dynamic sounds. It may be possible to improve the performance of the proposed method by exploiting additionally the periodicity and correlation properties of the glottal cycles and speech signals, respectively.

Chapter 6

Processing Multimicrophone Data Using Excitation Source Information

In many modern (hands-free) communication applications, speech signals are obtained in enclosed spaces such as meeting rooms with talkers situated at a distance from microphone. Moreover, in a meeting room scenario there is a possibility of more than one talker speaking at the same time. In these conditions, the observed speech signal is degraded by room reverberation, background noise and speech of the competing speakers. Reverberation degrades the speech signal [135], acting adversely on many speech processing applications including speech analysis, speech recognition and speaker recognition.

In the presence of room reverberation, a microphone signal is the mixture of the source speech signal and its delayed/attenuated copies. As a result, the microphone signal contains spurious impulse-like excitations due to reflected components of the actual source signal. Moreover, the amplitude of the direct component itself is low as the microphone is located at a distance from the source. Because of these factors, the speech signal collected in reverberant environment is different from the speech signal recorded through a close speaking microphone. Fig. 6.1 illustrates the effect of reverberation on the speech signal collected by a microphone placed at a distance from the speaker. Fig. 6.1(a) shows the speech signals collected by a close speaking microphone, and Fig. 6.1(b) shows the speech signal collected by a microphone placed at a distance of 2 m from the speaker.

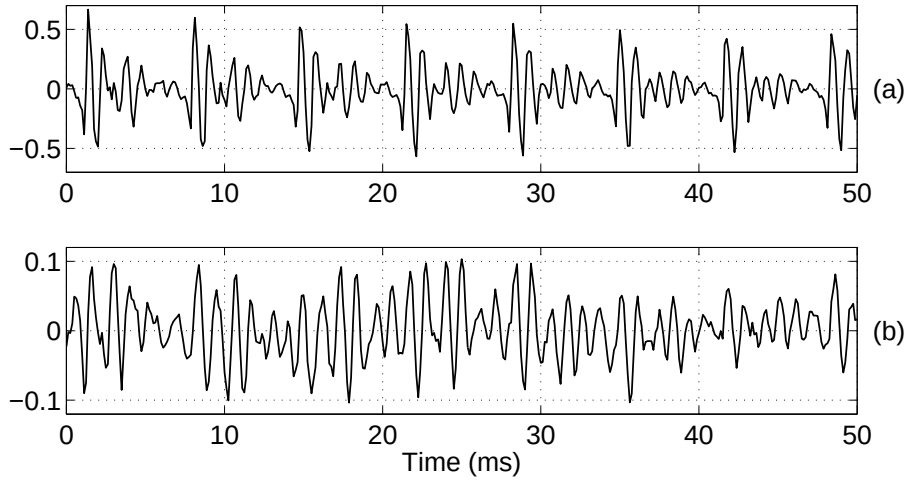


Fig. 6.1: Illustration of effect of reverberation on speech signal collected at a distance. Segment of speech signal collected by (a) close speaking microphone, and (b) microphone placed at a distance of 2 m from the speaker.

Both the signals are recorded simultaneously, and are time-aligned by compensating for the delay due to acoustic propagation. The amplitude of the clean speech signal in the closed glottis region is significantly high compared to its amplitude in open glottis region. Hence, the individual pitch periods can clearly be observed from the clean speech signal in Fig. 6.1(a). On the other hand, these observations are not evident from the reverberant speech signal shown in Fig. 6.1(b), because of the following reasons: (a) The amplitude of the direct component is low because of the attenuation suffered along the acoustic path. (b) The reflected components of the high amplitude signal in closed glottis region spread into the open glottis region making it difficult to unambiguously detect the individual pitch periods. As a result, the performance of the algorithms for epoch extraction and pitch determination, inevitably, degrades in the presence of reverberation.

In the meeting room scenario, in addition to the effect of reverberation, often more than one talker speak at the same time. In such a case, the signal collected at the microphone is a mixture of convolution of source signals with the impulse responses of the acoustic paths. As a result, the multispeaker speech signal contains epochs due to all the speakers (and their reflected components), which makes it difficult to observe the individual pitch periods.

The issues involved in both reverberant environment and multispeaker environment can be addressed when the speech signals collected from multiple microphones are avail-

able for processing. Microphone arrays are known to be useful in reverberant environments [136] due to spatial diversity of the room transfer function. Recent developments in speech analysis have made use of microphone arrays for epoch extraction [137, 138] and pitch detection [139, 140, 139] from reverberant speech using a delay-and-sum beamformer.

In this chapter, we discuss methods for pitch extraction from speech signals collected in reverberant environment, and in multispeaker environment. In both the cases, the speech signals are collected from a pair of spatially separated microphones in a real room environment. Spatial separation of microphones results in a fixed time-delay of arrival of speech signals from a given speaker at the pair of microphones. Except for the time-delay, the relative locations of the instants of significant excitation of the vocal-tract system remain unchanged in the direct components of the speech signals at the microphones. The time-delay of arrival between the pair of microphones is estimated using the excitation source characteristics of the speech signal. By compensating for the estimated time-delay of arrival, the speech signals from the pair of microphones are coherently processed to emphasize the epoch information, while minimizing the effect due to reverberation. In a multispeaker case, the differences in the time-delays for different speakers are exploited to separate the epochs due to individual speakers. This chapter is organized as follows: A method for estimation of the time-delay using the Hilbert envelope of the LP residual is presented in Section 6.1. In Section 6.2, we discuss a method for pitch estimation in reverberant environment using multimicrophone data. A method for multipitch estimation from multispeaker multimicrophone data is presented in Section 6.3. Section 6.4 summarizes the contributions discussed in this chapter.

6.1 Time-delay estimation

In a multispeaker multimicrophone scenario, assuming that the speakers are stationary with respect to the microphones, there exists a fixed time-delay of arrival of the speech signals (between every pair of microphones) from a given speaker. The time-delays corresponding to different speakers can be estimated using the crosscorrelation function of

the multispeaker signals. Positions of dominant peaks in the crosscorrelation function should ideally correspond to the time-delays due to all the speakers at the pair of microphones. However, the crosscorrelation function of the multispeaker signals does not show prominent peaks at the time-delays. This is mainly because of the damped-sinusoid-like components in the speech signal due to resonances of the vocal tract, and also because of the effects of reverberation and noise. These effects can be reduced by exploiting the characteristics of the excitation source of speech [141].

The impulse-like excitations during the production of voiced speech occur at the epoch locations. In the vicinity of these impulses, the speech signal exhibits a high SNR relative to the other regions. In order to highlight the high SNR regions in the speech signal, LP residual is derived from the speech signal using the autocorrelation method [4]. The LP residual reduces the second order correlations among the samples of the signal, and produces large amplitude fluctuations around the instants of significant excitation. The LP residual corresponds to an estimate of the excitation source of the speech signal. Note that the LP analysis of a multispeaker speech signal also produces uncorrelated samples in the LP residual, where large amplitude residual samples approximately correspond to the excitation part in the multispeaker signal. The crosscorrelation function of the LP residual signals from the two microphone multispeaker speech signals is not likely to yield strong peaks because of the large amplitude fluctuations of random polarity around the epoch locations, as shown in Fig. 6.2(b). The high SNR regions around the epoch locations can be highlighted by computing the Hilbert envelope of the LP residual [2]. The Hilbert envelope $h[n]$ of the LP residual is computed as the amplitude envelope of the analytic signal derived from the LP residual, as discussed in Section 5.2.1. The Hilbert envelope of the LP residual in Fig. 6.2(b) is shown in Fig. 6.2(c).

The crosscorrelation function of the Hilbert envelope of the LP residual signals derived from the multispeaker signals is used to estimate the time-delays [9]. Apart from the large amplitudes around the instants of significant excitation, the Hilbert envelope contains a large number of smaller positive values also, which may result in spurious peaks in the crosscorrelation function. Therefore, the regions around the instants of significant excitation can be further emphasized by dividing the square of each sample of the Hilbert

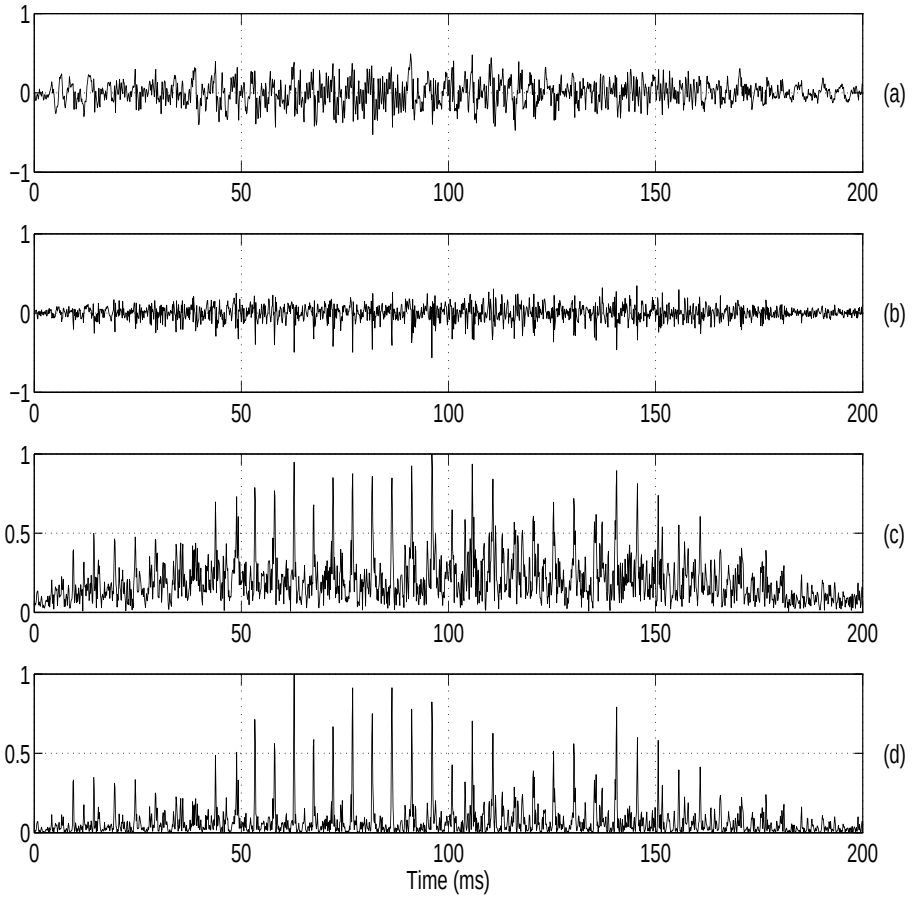


Fig. 6.2: Highlighting the high SNR regions around the epoch locations. (a) A 200 ms segment of multispeaker speech signal, (b) its LP residual, (c) Hilbert envelope of LP residual, (d) Hilbert envelope after emphasizing the epochs. The plots in (c) and (d) are normalized with their respective maximum values.

envelope by the moving average of the Hilbert envelope computed over a short window (about 4 ms, i.e., less than the average pitch period) around the sample. The computation of the preprocessed Hilbert envelope is as follows [141]:

$$g_i[n] = \frac{h_i^2[n]}{\frac{1}{2M+1}\sum_{m=n-M}^{n+M} h_i[m]}, \qquad i = 1, 2, \tag{6.1}$$

where $h_i[n]$ is the Hilbert envelope of the LP residual of the multispeaker signal collected at the i^{th} microphone, $g_i[n]$ is the corresponding preprocessed Hilbert envelope, and M is the number of samples corresponding to the duration of 2 ms (16 samples at 8 kHz). The effect of emphasizing the regions around the instants of significant excitation is shown in Fig. 6.2(d) for the Hilbert envelope shown in Fig. 6.2(c). The crosscorrelation function

$r_{12}[l]$ between the preprocessed Hilbert envelopes $g_1[n]$ and $g_2[n]$ is computed as [141]

$$r_{12}[l] = \frac{\sum_{n=z}^{N-|k|-1} g_1[n]g_2[n-l]}{\sqrt{\sum_{n=z}^{N-|k|-1} g_1^2[n] \sum_{n=z}^{N-|k|-1} g_2^2[n-l]}}, \qquad l = 0, \pm 1, \pm 2, \ldots, \pm L, \tag{6.2}$$

where $z = l$, $k = 0$, for $l \geq 0$, and $z = 0$, $k = l$, for $l < 0$, and N is the length of the segments of the Hilbert envelope. Here, both the vectors are normalized to unit magnitude for every sample shift before computing the crosscorrelation. The crosscorrelation function is computed over an interval of $2L + 1$ lags, where $2L + 1$ corresponds to an interval greater than the largest expected time-delay. The largest expected time-delay can be estimated from the approximate positions of the speakers and the microphones. The locations of the peaks with respect to the origin (zero lag) of the crosscorrelation function correspond to the time-delays between the microphone signals for all the speakers. The number of prominent peaks should correspond to the number of speakers. However, in practice, this is not always true because of the following reasons: (a) All the speakers may not contribute to voiced sounds in the segments used for computing the crosscorrelation function. (b) There could be spurious peaks in the crosscorrelation function, which may not correspond to the time-delay due to a speaker. Hence we rely only on the delay due to the most prominent peak in the crosscorrelation function. This delay is computed from the crosscorrelation function of successive frames of 50 ms duration shifted by 5 ms. Since different regions of the speech signal may provide evidence for the delays corresponding to different speakers, the number of frames corresponding to each delay is accumulated over the entire data. This helps in determining the number of speakers as well as their respective delays. Thus by collecting the number of frames corresponding to each delay over the entire data, there will be a large evidence for the delays corresponding to the individual speakers. Fig. 6.3 shows the percentage of the frames for each delay, for single-speaker recordings. Similarly, Fig. 6.4 shows the percentage of the frames for each delay, for two-speakers recordings. The histogram plots obtained by using the crosscorrelation of speech signals are also shown for comparison. The plots (for example Fig. 6.4(b) and Fig. 6.4(f)) show that emphasizing the regions around the significant excitation using the

Hilbert envelope gives better estimation of the time-delays. The locations of the peaks in the histogram indicate the time-delays due to different speakers [141]. Thus the number of prominent peaks in the histogram indicates the number of speakers in the conversation. The estimation of time-delays is based on the assumption that each speaker speaks at least for reasonable percentage of time. The minor peaks are due to random peaks in the correlation functions, and their occurrence is usually small (< 5%).

The accuracy of the time-delay estimation is evaluated using the speech signals collected from a single-speaker and from two speakers. Speech data was collected simultaneously using a pair of microphones separated by about 1 m in a laboratory environment with an average (over the frequency) reverberation time of about 0.5 s. All the recordings for this study were made under the following practical conditions:

(a) The speakers were seated at an average distance of 1.5 m from the microphones. The speakers were seated such that their heads and the microphones were approximately in the same plane.

(b) While collecting the two-speakers data, the speakers were positioned in such a way that the time delay is different for different speakers. In fact, any arbitrary placement of the speakers with respect to the microphones satisfies this requirement.

(c) It is assumed that the level of the direct component of speech from each speaker at the microphones is significantly high relative to the noise and reverberation components in the room.

(d) While recording the two-speakers data, both the speakers were stationary and they spoke simultaneously during the entire duration of recording, resulting in significant overlap.

The speech signals were sampled at 8 kHz. During each recording, the distances of the speakers from both the microphones were measured. The actual time-delay τ of arrival of the speech signals at *Mic-1* and *Mic-2* located at distances d_1 and d_2, respectively, from a speaker is given by

$$\tau = \frac{(d_1 - d_2)}{c} \tag{6.3}$$

Table 6.1: Comparison of estimated time-delays $\hat{\tau}$ with reference time-delays τ for four single-speaker recordings. Reference values are computed from the measured distances d_1 and d_2.

S. No.	d_1 (m)	d_2 (m)	τ (ms)	$\hat{\tau}$ (ms)
1	1.22	1.64	-1.27	-1.25
2	1.20	0.99	0.636	0.625
3	1.33	1.43	-0.303	-0.25
4	1.75	1.40	1.060	1

where c is the speed of sound in air. A negative time-delay (lead) indicates that the speaker is nearer to *Mic-1* relative to *Mic-2*.

The multimicrophone signals were processed using the proposed method to obtain the time-delays. A 10^{th} order LP analysis was used for deriving the LP residual. The crosscorrelation function of the preprocessed Hilbert envelopes of the LP residuals of the multimicrophone signals was used to estimate the time-delays. The percentage of frames for each delay (in ms) for single-speaker cases and two-speakers cases are shown in Fig. 6.3 and Fig. 6.4, respectively. The locations of peaks in the time-delay histogram correspond to the time-delays due to different speakers. Table 6.1 lists the actual time-delay τ obtained from the measured distances d_1 and d_2, and the estimated time-delay obtained from the histogram for four single-speaker recordings. Similar comparison of time-delays for four two-speakers recordings is provided in Table 6.2. The actual and the estimated time-delays are in close agreement, thus indicating the effectiveness of the method in determining the time-delay of arrival from multimicrophone data. Minor deviations between the actual and estimated time-delays could be attributed to the following: (a) Errors in measuring the actual distances, (b) the time-resolution that can be achieved at the sampling frequency of 8 kHz (multiples of 0.125 ms), and (c) movement of the speakers during a recording session.

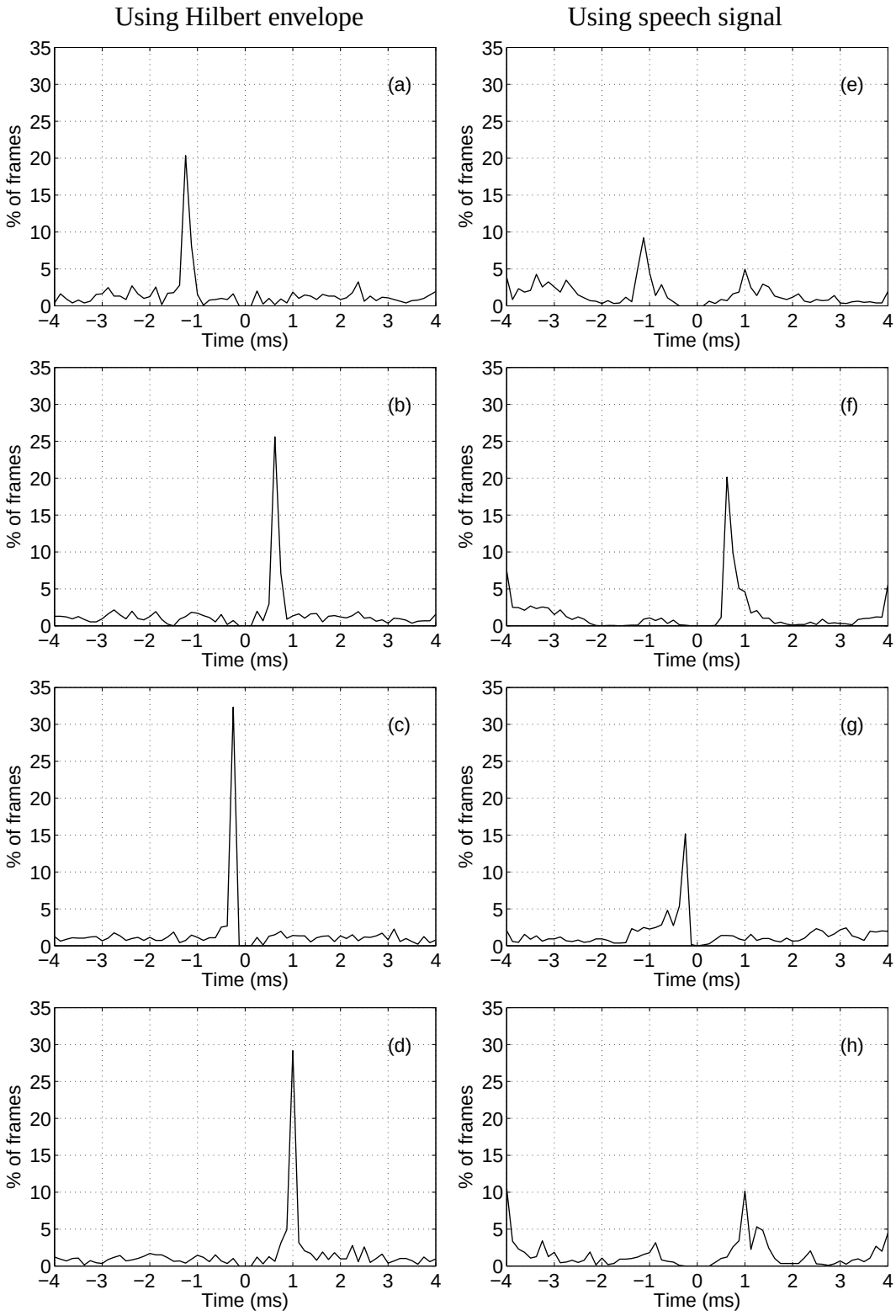


Fig. 6.3: Time-delay histograms for four single-speaker recordings. The plots in the first column are obtained from the preprocessed Hilbert envelope, and those in second column are obtained directly from the speech signal. Each row of plots corresponds to an entry in Table 6.1, in that order.

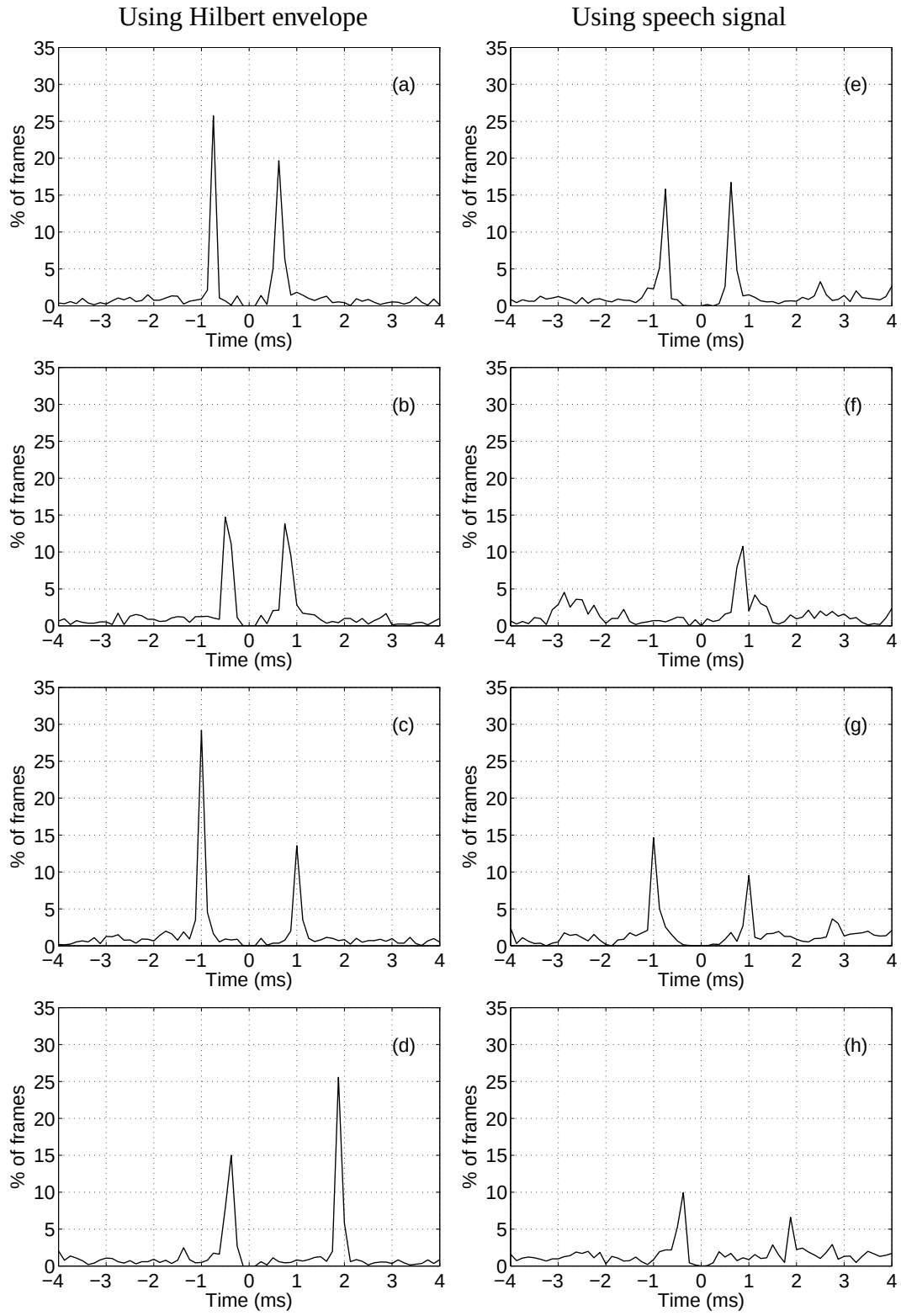


Fig. 6.4: Time-delay histograms for four two-speakers recordings. The plots in the first column are obtained from the preprocessed Hilbert envelope, and those in second column are obtained directly from the speech signal. Each row of plots corresponds to an entry in Table 6.2, in that order.

Table 6.2: Comparison of estimated time-delays $\hat{\tau}$ with reference time-delays τ for four two-speakers recordings. Reference values are computed from the measured distances d_1 and d_2.

S. No	*Spkr-j*	d_{1j} (m)	d_{2j} (m)	τ_j (ms)	$\hat{\tau}_j$ (ms)
1	*Spkr-1*	0.85	110	-0.757	-0.75
	Spkr-2	1.27	1.07	0.61	0.625
2	*Spkr-1*	1.24	1.38	-0.424	-0.5
	Spkr-2	1.20	0.94	0.787	0.75
3	*Spkr-1*	1.10	1.43	-1	-1
	Spkr-2	1.50	1.16	1.030	1
4	*Spkr-1*	1.16	1.29	-0.393	-0.375
	Sprk-2	1.52	0.90	1.878	1.875

6.2 Pitch estimation in reverberant environment

When the speech signal is collected in an acoustical environment like a meeting room, it will be degraded by background noise and reverberation. The speech signal collected at a distance from the microphone may be expressed as

$$s_d[n] = s[n] + \sum_{i=1}^{N} \alpha_i s[n - \tau_i] + v[n], \tag{6.4}$$

where $s_d[n]$ is the degraded signal, $s[n]$ is the source signal and $v[n]$ is the background noise component, α_i is the amplitude of the reflected component arriving after a delay of τ_i samples and N is the number of reflections. The background noise component is independent of speech, whereas, the reflected component is dependent on the previous speech signal. The effect of the reflected components on the speech signal is not predictable. The reflected components of the speech signal arriving at arbitrary time-delays get superposed with the direct component causing discontinuities in between the actual epoch locations. Because of the spurious discontinuities resulting from the reflected components, the pitch estimation algorithm presented in Chapter 5 can not be applied directly on the reverberant

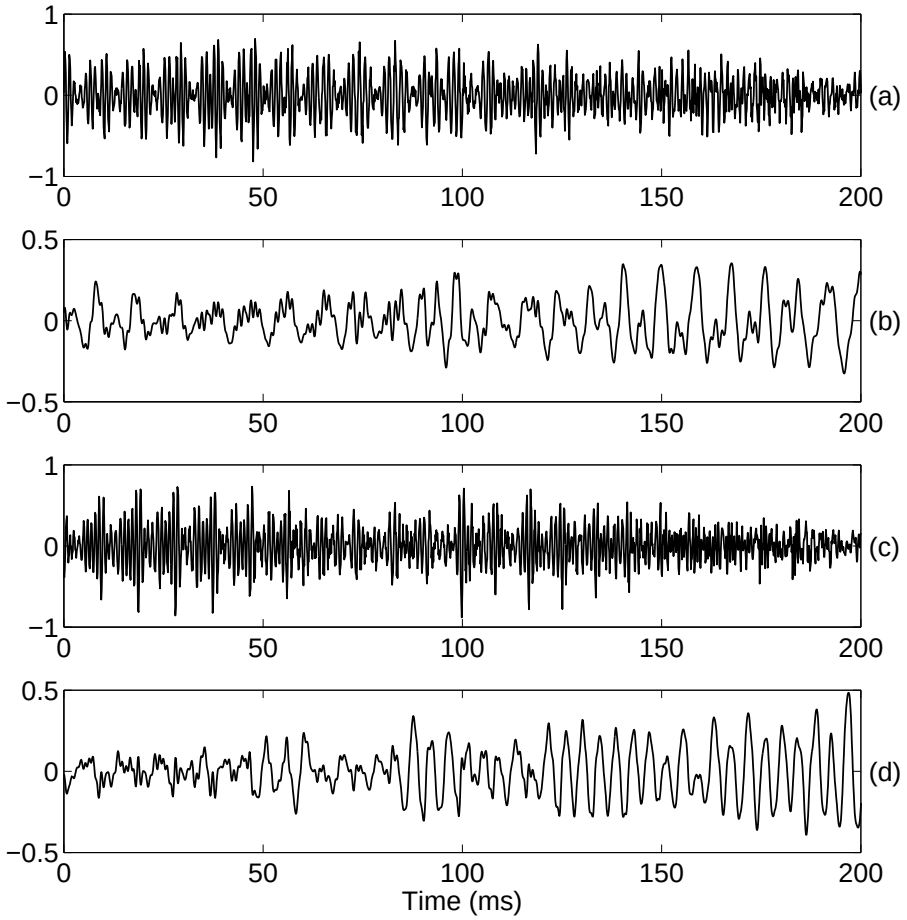


Fig. 6.5: Effect of reverberation on the filtered signal. (a) Speech signal collected at a distance of 1.8 m from the speaker, and (b) its filtered signal.(c) Speech signal collected at a distance of 1.5 m from the speaker, and (d) its filtered signal.

speech signal. The discontinuities caused by the reflected components introduce spurious zero-crossings in zero-frequency filtered speech signal, and it is difficult to distinguish the zero-crossings due to the epochs. Fig. 6.5(a) and Fig. 6.5(c) show a segment of speech signal collected by a pair of microphones placed at distances 1.8 m and 1.5 m, respectively, from the speaker. The zero-frequency filtered signals obtained from the speech signals are shown in Fig. 6.5(b) and Fig. 6.5(d). The zero-frequency filtered signals contains several spurious zero-crossings that are difficult to distinguish from the zero-crossings due to the actual epochs. Though the individual evidences from the filtered signals of speech collected using a pair of microphones do not provide true zero-crossings, and thereby pitch, both the speech signals can be processed coherently to estimate the pitch in reverberant conditions.

6.2.1 Emphasizing epochs over reverberant components

The excitation source of voiced speech segments consists of impulse-like excitations around the epoch locations. The impulse-like excitation characteristics are captured using preprocessed Hilbert envelope of LP residual of voiced speech. The impulse-like excitation is robust in the sense that the relative spacing of the epochs due to direct component remains unchanged at different microphone locations. That is, the epochs corresponding to the direct components will be *coherent* at different microphone positions. This can be observed from Fig. 6.6, where the Hilbert envelopes of LP residuals of speech signals collected from a pair of microphones are time-aligned and displayed. The voiced speech segments chosen for this illustration are the same as those used in Fig. 6.5. Fig. 6.6(d) and Fig. 6.6(e) show the Hilbert envelopes of LP residuals of voiced speech segments shown in Fig. 6.6(a) and Fig. 6.6(b), respectively. The Hilbert envelopes contain several spurious peaks along with the peaks at the epoch locations. But the locations of the spurious peaks are not in coherence, whereas, the locations of the peaks due to epochs are in coherence. Hence, by coherently adding Hilbert envelopes of both the LP residuals, the coherent epochs can be enhanced over the incoherent spurious peaks as shown in Fig. 6.6(f). The coherently added Hilbert envelope shows significantly less spurious peaks as compared to the individual Hilbert envelopes. Notice that the pitch period information is observed more clearly in Fig. 6.6(f), compared to the signal shown in Fig. 6.6(c), which is obtained by coherent addition of speech signals.

The coherently added Hilbert envelope obtained from the speech signals collected using a pair of microphones is used for pitch estimation in reverberant environment. The interval between two successive significant peaks in the coherently added Hilbert envelope corresponds to the pitch period. But peak detection from the coherently added Hilbert envelope is not a trivial task. As the amplitudes of the peaks have a large dynamic range, it is not possible to set a fixed threshold to detect them. In this work, the peaks are detected using the zero-frequency filtered signal of the coherently added Hilbert envelope. The zero-frequency filtered signal of the coherently added Hilbert envelope exhibits sharper zero-crossings at the peak locations. Fig. 6.7(b) shows the zero-frequency filtered signal of the coherently added Hilbert envelope shown in Fig. 6.7(a). The positive zero-

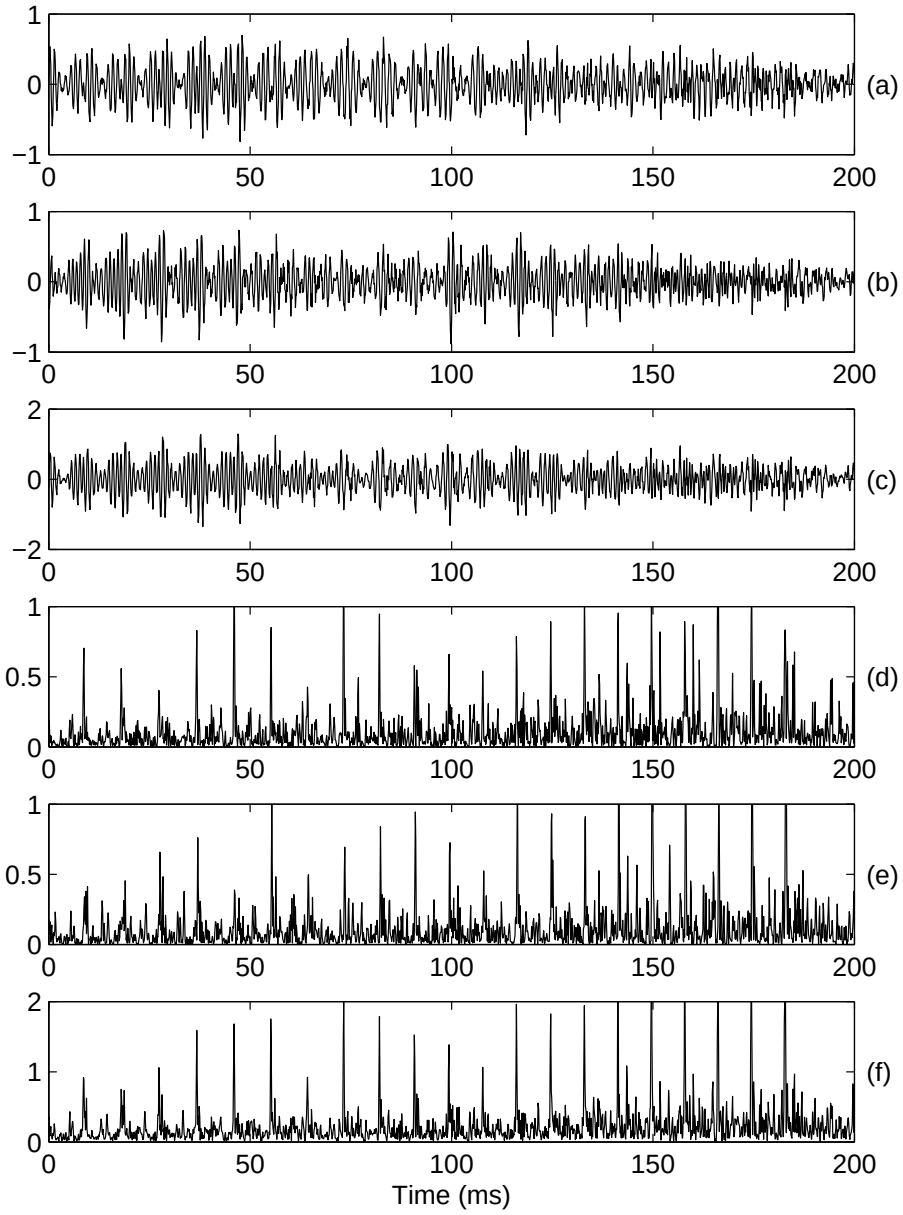


Fig. 6.6: Effectiveness of coherent addition of Hilbert envelopes for emphasizing peaks due to epochs over the peaks due to reflected components. (a) Speech signal collected at *Mic-1*, (b) speech signal collected at *Mic-2*, (c) coherently added speech signals, (d) Hilbert envelope of LP residual of speech signal in (a), (e) Hilbert envelope of LP residual of speech signal in (b), and (f) coherently added Hilbert envelopes.

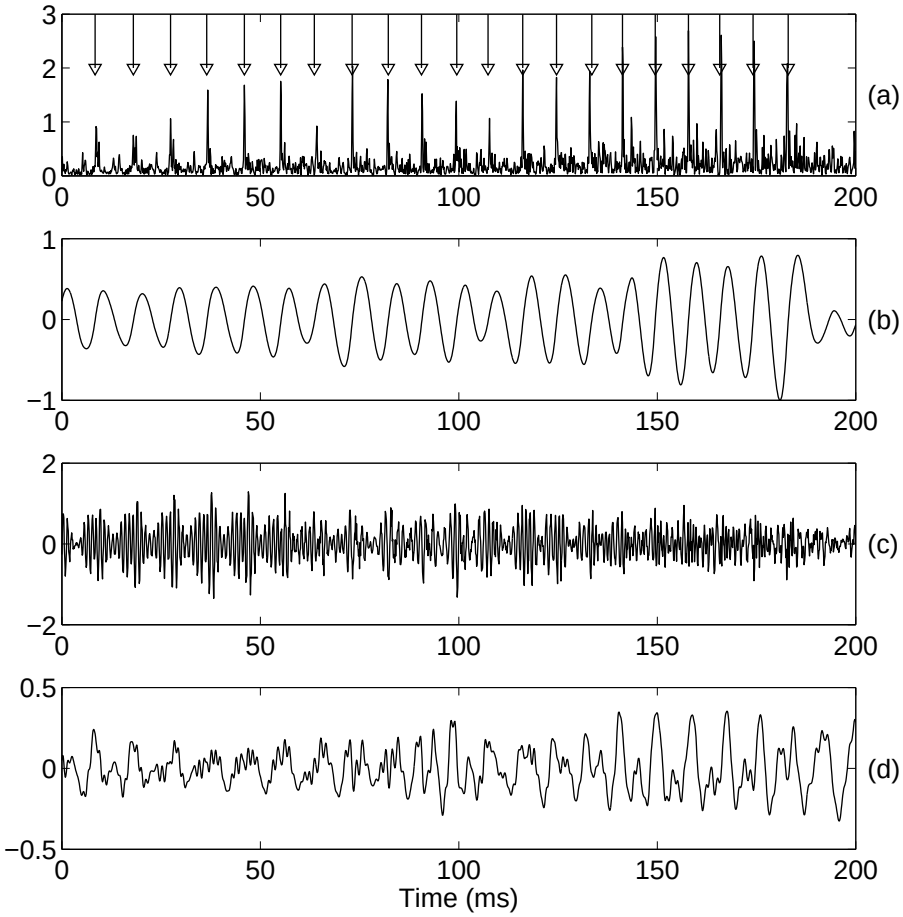


Fig. 6.7: Illustration of zero-frequency filtering on coherently added Hilbert envelope. (a) Coherently added Hilbert envelope and its (b) zero-frequency filtered signal. (c) Coherently added speech signal and its (d) zero-frequency filtered signal. The arrows in (a) indicate the epoch locations obtained from positive zero-crossings of zero-frequency filtered signal.

crossings of the filtered signal are in close agreement with the locations of the peaks of the coherently added Hilbert envelope. Fig. 6.7(d) shows the filtered signal of the coherently added speech signal shown in Fig. 6.7(c). As mentioned earlier the filtered signal of the coherently added speech signals contains spurious zero-crossings, as the effect of reverberation is not minimized.

The pitch frequency is measured as the reciprocal of the time interval between two successive positive zero-crossings of the filtered signal of the coherently added Hilbert envelope. Fig. 6.8(a) and Fig. 6.8(b) show speech signals collected using a pair of microphones, and Fig. 6.8(c) shows the coherently added Hilbert envelope derived from them. The arrow marks in Fig. 6.8(c) indicate the positive zero-crossings of the filtered signal

(Fig. 6.8(d)) of the coherently added Hilbert envelope. The pitch contours for the voiced segments are obtained from the positive zero-crossings of the filtered Hilbert envelope. The unvoiced segments are indicated by zero pitch values. The voicing decision is obtained from the filtered Hilbert envelope shown in Fig. 6.8(d). Notice that the voiced regions can be clearly observed from the filtered signal, which is not evident from the speech signals because of the effects of reverberation and background noise.

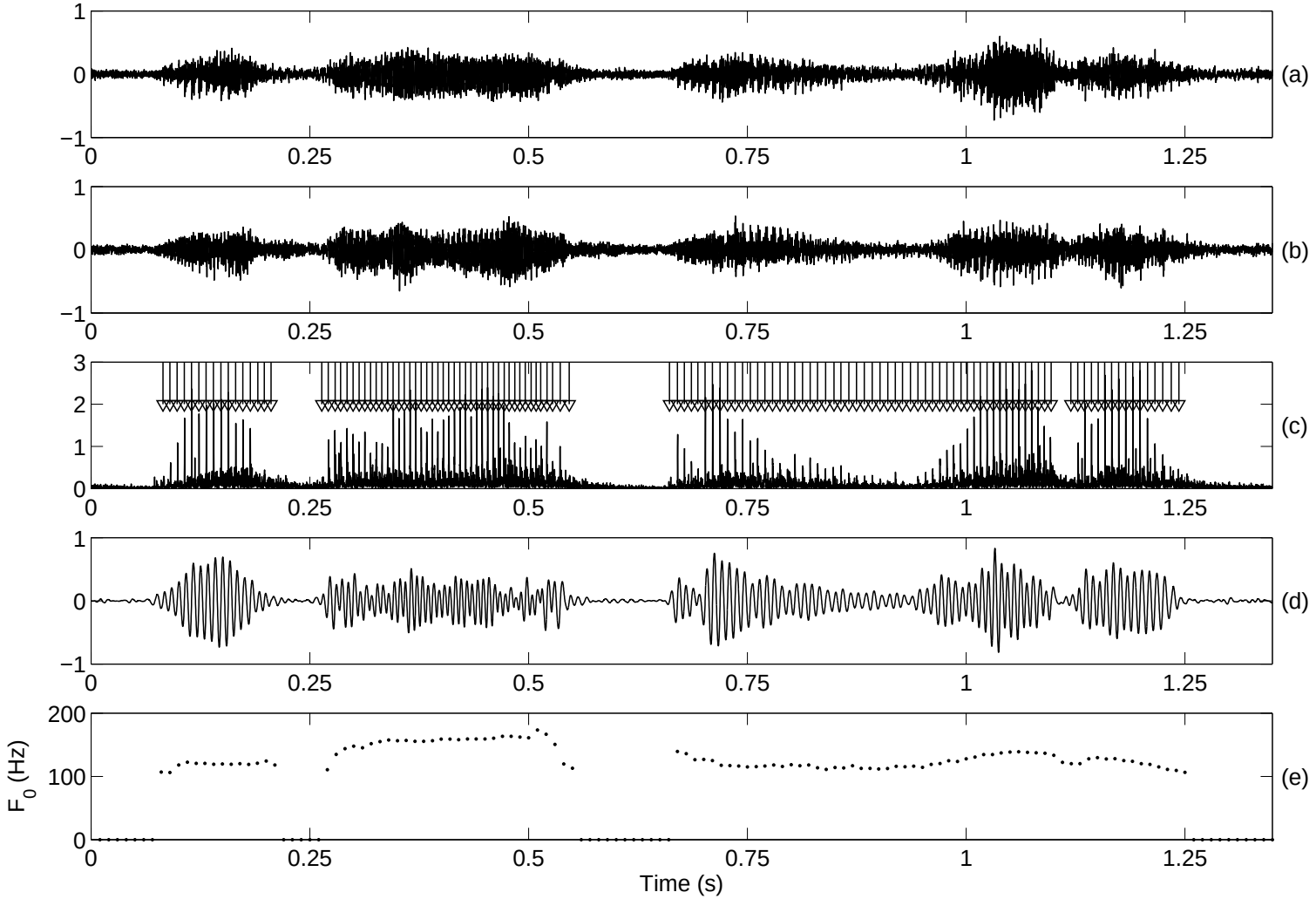


Fig. 6.8: Illustration of pitch estimation from multimicrophone speech signals in reverberant environment. Speech signal collected at (a) microphone-1, (b) microphone-2, (c) coherently added Hilbert envelope, its (d) zero-frequency filtered Hilbert envelope and (e) pitch contour derived from successive positive zero-crossings of filtered Hilbert envelope.

6.2.2 Performance evaluation

Performance of the proposed pitch estimation method using multiple microphones was evaluated on the speech data collected in a laboratory environment. Fifteen randomly selected TIMIT utterances were recorded from three male speakers. The speech data was simultaneously recorded using three microphones - a close speaking microphone and two distant microphones located at 1.8 m and 1.5 m from the speakers. All the recordings are done at a sampling frequency of 8 kHz. Close speaking microphone recordings are used to create the reference pitch values using the pitch estimation method presented in Chapter 5. Distant microphone recordings are used to evaluate the proposed method.

The gross error (GE), the mean absolute error (M), and the standard deviation error (SD),defined in Chapter 5, are used to evaluate the pitch estimates obtained from the proposed method. The reference pitch values are obtained from the speech data recorded using the close speaking microphone. The performance of the proposed method is compared with the performance of Praat's autocorrelation method and the pitch estimation method proposed in Chapter 5, on the speech signals collected at individual microphones. Table 6.3 gives a performance comparison of proposed multimicrophone pitch estimation with the existing methods applied on individual microphone data. AC-Mic1 and AC-Mic2 denote the Praat's autocorrelation method applied on speech signals collected at *Mic-1* and *Mic-2*, respectively. ZFR-Mic1 and ZFR-Mic2 denote the zero-frequency resonator method, proposed in Chapter 5, applied on speech signals collected at *Mic-1* and *Mic-2*, respectively. The performance of Praat's autocorrelation method is better than the performance of the method proposed in Chapter 5 on individual microphone data. This is because of the spurious zero-crossings in the filtered signal due to reflected components. When the epochs are emphasized by coherent addition of Hilbert envelopes, the resulting performance is better than the performance of the existing methods on any one of the individual signals. The smaller GE for the proposed method indicates that the number of frames, for which the estimated fundamental frequency lie within 20% of the reference values, is large. As a result, the computation of mean absolute error for the proposed method includes low SNR frames. Hence the mean absolute error and standard deviation are slightly higher for the proposed method compared to those for the AC method.

Table 6.3: Performance of pitch estimation algorithms on reverberant speech data. AC-Mic1 and AC-Mic2 denote Praat's autocorrelation method applied on speech signals collected at *Mic-1* and *Mic-2*, respectively. ZFR-Mic1 and ZFR-Mic2 denote the zero-frequency resonator method, proposed in Chapter 5, applied on speech signals collected at *Mic-1* and *Mic-2*, respectively.

Method	GE (%)	M (Hz)	SD (Hz)
AC-Mic1	26.50	3.62	5.20
AC-Mic2	22.71	3.37	4.80
ZFS-Mic1	34.38	9.46	7.37
ZFS-Mic2	28.04	8.09	6.77
Coherent Addition	16.43	4.07	4.82

6.3 Multipitch extraction

The signal collected by a microphone in a multispeaker environment is a mixture of speech signals from several speakers. Pitch extraction from multispeaker speech signals is a challenging task, as the pitch periods from all the speakers overlap, making it difficult to even observe the individual pitch periods. Fig. 6.9(a) and Fig. 6.9(b) show the speech signals collected by a pair of microphones when two persons (one male and one female) are speaking simultaneously. It is difficult to observe the pitch periods corresponding to the speakers from any of the two signals. Even the autocorrelation of a frame (30 ms) of the two-speaker signal is not likely to yield two unambiguous peaks corresponding to the pitch periods of both the speakers. In this work, the multispeaker signals from the pair of microphones are processed together to emphasize the epoch information of the individual speakers.

6.3.1 Emphasizing epochs of individual speakers

The characteristics of excitation around the epochs, and the robustness of the relative spacing of the epochs in the speech signals collected at a pair of microphones can be exploited for identifying the epoch locations corresponding to a given speaker. Let $g_1[n]$ and $g_2[n]$ be the preprocessed Hilbert envelopes of the LP residuals of the speech signals

collected at *Mic-1* and *Mic-2*, respectively, as given in (6.1). By aligning the Hilbert envelopes $g_1[n]$ and $g_2[n]$ after compensating for the estimated time-delay ($\hat{\tau}_1$) corresponding to *Spkr-1*, the epochs corresponding to that speaker will be in coherence, whereas the epochs corresponding to *Spkr-2* will be incoherent. By considering the minimum of the Hilbert envelopes $g_1[n]$ and $g_2[n - \hat{\tau}_1]$, only the Hilbert envelopes around the epochs corresponding to *Spkr-1* are retained. Note that this operation of retaining minimum ensures that the Hilbert envelope peaks at the epochs of the other speakers are suppressed. The resulting signal is referred as the Hilbert envelope specific to *Spkr-1*. In a similar manner, the signal that retains the Hilbert envelope around the epochs corresponding to *Spkr-2* can be derived. Let

$$h_{sj}[n] = \min(g_1[n], g_2[n - \hat{\tau}_j]), \qquad j = 1, 2, \tag{6.5}$$

where $h_{s1}[n]$ and $h_{s2}[n]$ are the signals in which the Hilbert envelopes around the epochs corresponding to *Spkr-1* and *Spkr-2*, respectively, are retained.

Fig. 6.9 illustrates the extraction of speaker-specific Hilbert envelopes from two-speaker speech signals collected using a pair of microphones. Fig. 6.9(c) and Fig. 6.9(d) show the Hilbert envelopes of the LP residuals of the two-speaker speech signals shown in Fig. 6.9(a) and Fig. 6.9(b), respectively. The Hilbert envelopes consist of the impulse-like excitations due to the epochs of both the speakers. It is difficult to separate the peaks due to epochs of the individual speakers from any one of them. But the speaker-specific Hilbert envelopes (Fig. 6.9(e) and Fig. 6.9(f)), obtained by computing the minimum of the delay-compensated Hilbert envelopes, clearly show epochs due to individual speakers.

The proposed method of obtaining speaker-specific Hilbert envelopes also aids in identifying the regions specific to the individual speakers, and the overlapped regions. Fig. 6.10(a) and Fig. 6.10(b) show speech signals collected from two male speakers using a pair of microphones. The regions of speech signal specific to the individual speakers are marked by dashed lines in Fig. 6.10(a), and are labelled on the top. The labels '1' and '2' indicate the regions specific to *Spkr-1* only and *Spkr-2* only, respectively. The label '1&2' indicates the region where there is an overlap of voiced speech from both the speakers. The regions are marked and labelled manually by listening to the two-speaker data. Neither the speech signals (Fig. 6.10(a) and Fig. 6.10(b)) nor the Hilbert envelopes

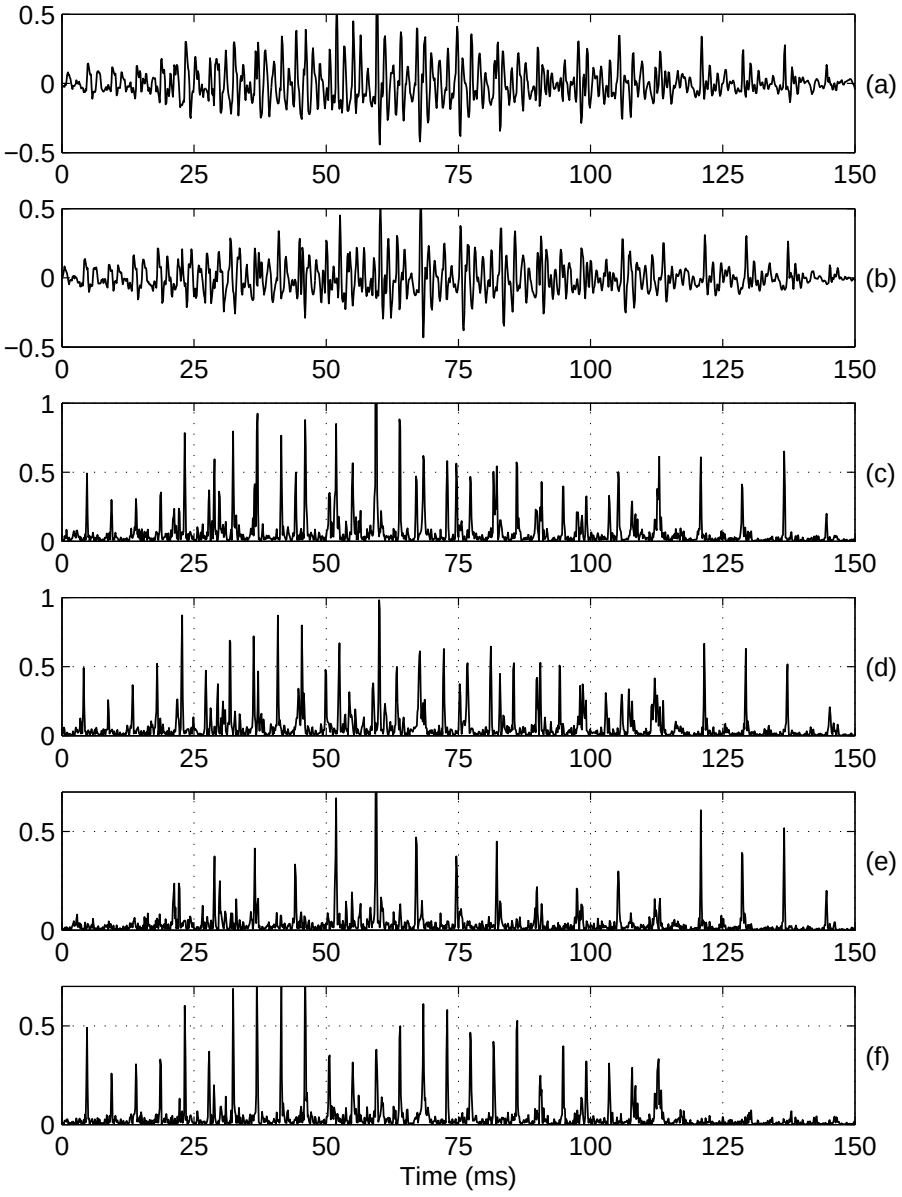


Fig. 6.9: Illustration of extracting speaker-specific Hilbert envelopes from two-speaker data collected using a pair of microphones. Speech signal collected at (a) *Mic-1* and (b) *Mic-2*. Hilbert envelope of LP residual of (c) *Mic-1* signal, and (d) *Mic-2* signal. Speaker-specific Hilbert envelopes of (e) *Spkr-1* and (f) *Spkr-2*. The time-delays of arrival due to *Spkr-1* and *Spkr-2* are 0.5 ms and -0.625 ms, respectively.

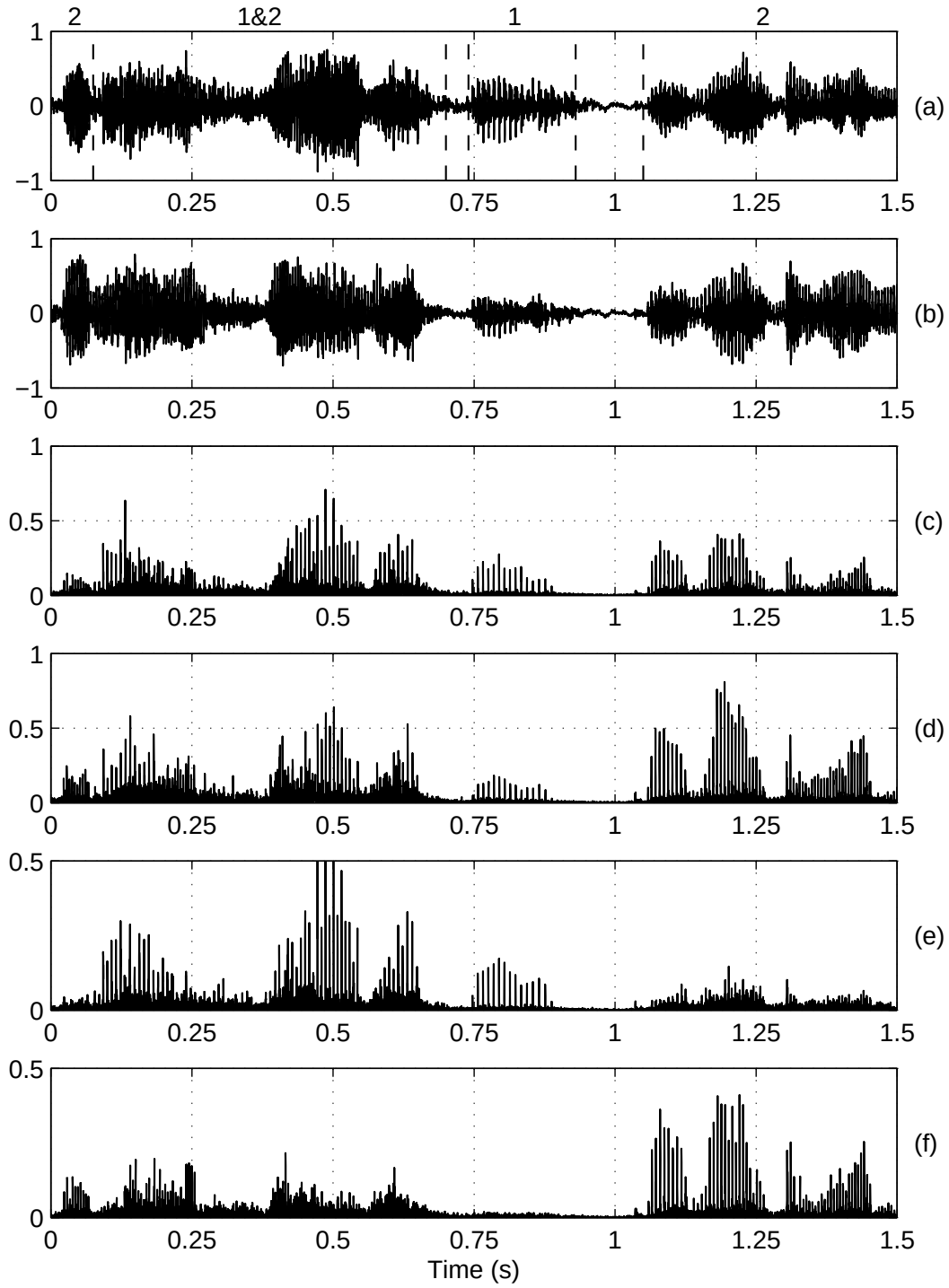


Fig. 6.10: Illustration of extracting speaker-specific regions from multispeaker speech signals. Speech signal collected at (a) *Mic-1* and (b) *Mic-2*. Hilbert envelope of LP residual of (c) *Mic-1* signal, and (d) *Mic-2* signal. Speaker-specific Hilbert envelopes of (e) *Spkr-1* and (f) *Spkr-2*.

of their LP residuals (Fig. 6.10(c) and Fig. 6.10(d)) give any clue for identifying the regions of individual speaker activity. However, the speaker-specific Hilbert envelopes shown in Fig. 6.10(e) and Fig. 6.10(f) clearly separate the regions corresponding to individual speakers. For example, the region from 0.75 s to 0.9 s is due to *Spkr-1* only, which is reflected as impulse-like sequence in Fig. 6.10(e), and is almost zero in Fig. 6.10(f). Likewise, the region from 1.1 s to 1.5 s is due to *Spkr-2* only, which is reflected clearly in Fig. 6.10(f). In the overlapped region from 0.1 s to 0.7 s, both Fig. 6.10(e) and Fig. 6.10(f) show activity, and contains the epochs corresponding to the respective speakers as illustrated in Fig. 6.9. Hence, the speaker-specific Hilbert envelopes can be used to detect the regions of individual speaker activity, and to estimate their individual pitch.

6.3.2 Multipitch extraction using zero-frequency resonator

The speaker-specific Hilbert envelopes predominantly contain impulse-like excitations at the epoch locations of the respective speakers. The pitch period of a given speaker can be estimated by measuring the interval between two successive peaks in the speaker-specific Hilbert envelope of that speaker. This requires detecting peaks from speaker-specific Hilbert envelope, that contains large variation among the peak amplitudes. In order to avoid the difficult task of peak detection, the zero-frequency filtering approach proposed in Chapter 3 is employed to detect the impulse-like excitations in the speaker-specific Hilbert envelope. The positive zero-crossings of the filtered Hilbert envelope closely match with the peaks in the speaker-specific Hilbert envelope as shown in Fig. 6.11. Fig. 6.11(a) and Fig. 6.11(b) show the speaker-specific Hilbert envelope of *Spkr-1* and its filtered signal, respectively. Even the low amplitude peaks around 90 ms to 120 ms are correctly detected, while the spurious peaks around 50 ms to 60 ms are rightly ignored. Similar observations can be made from Fig. 6.11(c) and Fig. 6.11(d), which show speaker-specific Hilbert envelope of *Spkr-2* and its filtered signal. Hence the zero-crossings of the filtered signal of the speaker-specific Hilbert envelopes are used to estimate the pitch of the individual speakers.

Performance of the proposed multipitch detection method is illustrated in Fig. 6.12 for

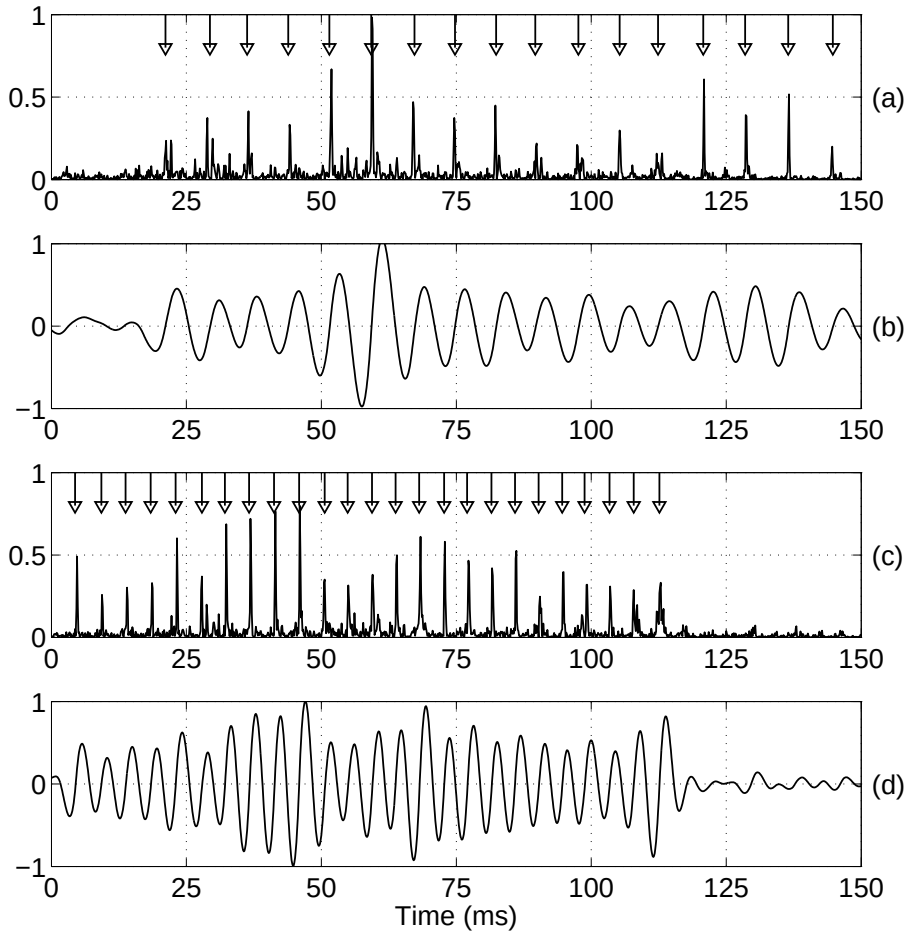


Fig. 6.11: Illustration of epoch extraction from speaker-specific Hilbert envelope using zero-frequency resonator. (a) Speaker-specific Hilbert envelope of *Spkr-1* and its (b) zero-frequency filtered signal. (c) Speaker-specific Hilbert envelope of *Spkr-2* and its (d) zero-frequency filtered signal. The arrow marks in (a) and (c) indicate the epoch locations detected from zero-crossings of filtered signals in (b) and (d), respectively.

artificially mixed signals whose source signals are known. Fig. 6.12(a) and Fig. 6.12(b) show the speaker-specific Hilbert envelope of *Spkr-1* (male) and its filtered signal, respectively. The solid line in Fig. 6.12(c) shows the pitch contour derived from the positive zero-crossings of the filtered signal in Fig. 6.12(b). The dotted line indicates the reference pitch contour derived from the source signal. The pitch contour derived from the multi-speaker signal using the proposed method closely follows the reference pitch contour. A few gross errors occur because of the errors in detecting the speaker-specific regions from the Hilbert envelopes. Fig. 6.12(d), Fig. 6.12(e) and Fig. 6.12(f) show the speaker-specific Hilbert envelope of *Spkr-2*, its filtered signal, and the pitch contour derived from the fil-

tered signal, respectively. The deviation of the estimated pitch contour from the reference may be attributed to spurious peaks occurring due to relative dominance of the speakers. In this case, an informal listening of the mixed speech signals suggests that *Spkr-1* is relatively more dominant than *Spkr-2*. In general, it was observed that the proposed multipitch estimation provides better estimates for the dominant speaker compared to the background speaker.

6.4 Summary

In this chapter, we proposed methods to process multimicrophone data for pitch extraction in reverberant environment and multispeaker environment. A method is discussed for estimating the time-delays (due to multiple speakers) from speech signals collected over a pair of spatially separated microphones. The method uses the knowledge of the excitation source, unlike the commonly used spectrum-based methods. The Hilbert envelope of the LP residual signal derived from speech is used to represent the excitation source information. The time-delays are estimated from the crosscorrrelation function of short segments (50 ms) of Hilbert envelopes. Since the time-delays can be estimated accurately even from short segments of Hilbert envelope, it may be possible to develop an algorithm to track a moving speaker.

Using the knowledge of time-delay, the Hilbert envelopes of LP residuals of individual microphone signals are coherently added. The coherently added Hilbert envelope emphasizes the peaks due to epochs while deemphasizing the spurious peaks due to reverberant components. In order to avoid difficult task of peak detection of coherently added Hilbert envelope, we proposed to use the zero-frequency filtering (proposed in Chapter 3) on coherently added Hilbert envelope to detect impulse-like excitations. The time intervals between successive positive zero-crossings of the filtered signal of coherently added Hilbert envelope are used to measure the pitch in reverberant environment.

Multipitch extraction is achieved by exploiting the differences in time-delays corresponding to different speakers. When the Hilbert envelopes are compensated for the time-delay due to one of the speakers, then the epochs of that speaker will be coherent, while

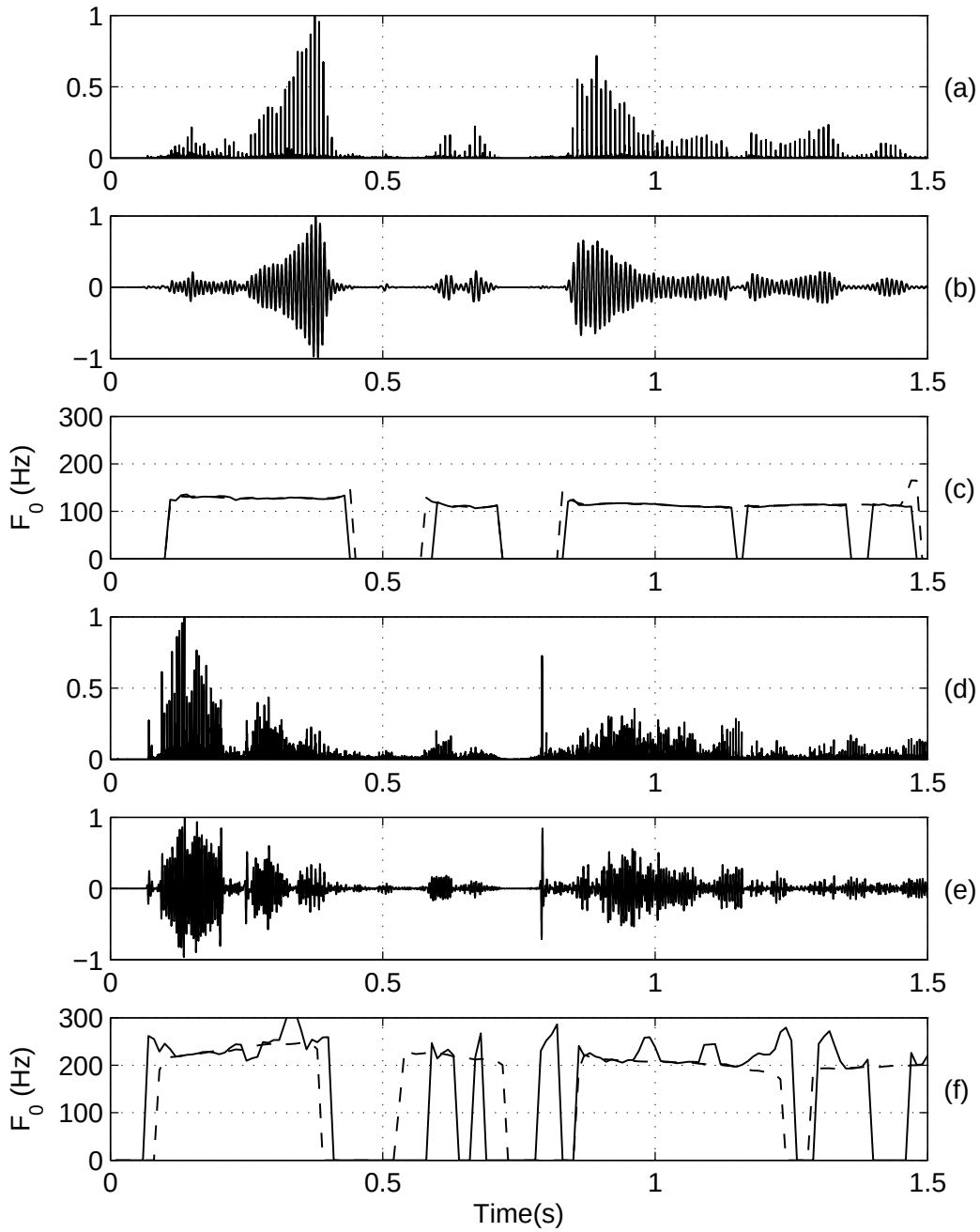


Fig. 6.12: Illustration of performance of proposed method of multipitch extraction. (a) Speaker-specific Hilbert envelope of *Spkr-1*, (b) its filtered signal, and (c) pitch contour derived from zero-crossings of the filtered signal. (d) Speaker-specific Hilbert envelope of *Spkr-2*, (e) its filtered signal, and (f) pitch contour derived from zero-crossings of the filtered signal. The dashed lines in (c) and (f) indicate the reference pitch contours derived from source signals.

the epochs due to the other speaker will be incoherent. Hence, the minimum of the delay compensated Hilbert envelopes emphasizes epochs due to one speaker, while deemphasizing epochs due to the other speaker. The speaker-specific Hilbert envelopes, thus derived, are used for multipitch extraction.

Chapter 7

Analysis of Manner of Articulation of Stop Consonants

Stop consonants are a class of speech sounds whose characteristic feature is an interval during which the airflow is completely blocked within the oral cavity. With the nasal cavity closed by velum, the air pressure built up behind the oral closure is released more or less impulsively as the vocal-tract moves towards a configuration appropriate for the following vowel. Depending on the place of closure in the oral cavity, different linguistic contrasts of the stop consonants can be produced. A stop consonant is said to be voiced (in a phonetic sense) if there is an audible laryngeal pulsation during the closure phase, and unvoiced if it is absent. Another phonetic feature traditionally attributed to the stop consonants is aspiration. If a stop consonant has noisy, plosive release, it is said to be aspirated; if not, it is unaspirated.

Stop consonants of consonant-vowel type form an important subset of alphabet in most Indian languages. Table 7.1 lists the stop consonants with the vowel ending /a/ for Indian languages. The stop consonants have clear manners of articulation besides the distinct places of articulation. The acoustic-phonetic description of the consonants in each of these syllables is precise, and is expressed in terms of voiced (V), unvoiced (uV), aspirated (A) and unaspirated (uA) categories. That is, for a given place of articulation there exists a four way contrast, among the stop consonants, depending on the manner

of articulation. While the information about the place of articulation is characterized by dynamics of the vocal-tract system, the manner of articulation is primarily dictated by the excitation source. One of the challenging tasks in speech analysis is to determine the acoustic correlates of the articulatory events during the production of the stop consonants, that are difficult to extract even from clearly articulated speech signals.

The objective of this work is to study the role of source features in the analysis of manner of articulation of stop consonants. Note that some features of the excitation source may be too short as in the case of burst, or too random as in the case of aspiration. As a result, mere spectral description of these characteristic regions may not provide unique and clear features. Moreover, the effects of block processing may limit the visibility of these features in the spectrogram. Hence, we propose to use the excitation source information derived from the speech signal for analyzing the stop consonants.

This chapter is organized as follows: Section 7.1 highlights the importance of excitation source information for analysis of manner of articulation of stop consonants. In Section 7.2, we describe the excitation source features chosen to study the stop consonants. In Section 7.3, we illustrate the potential of the proposed excitation source features in distinguishing voiced, unvoiced, aspirated and unaspirated stop sounds. Finally, Section 7.4 summarizes the results presented in this chapter.

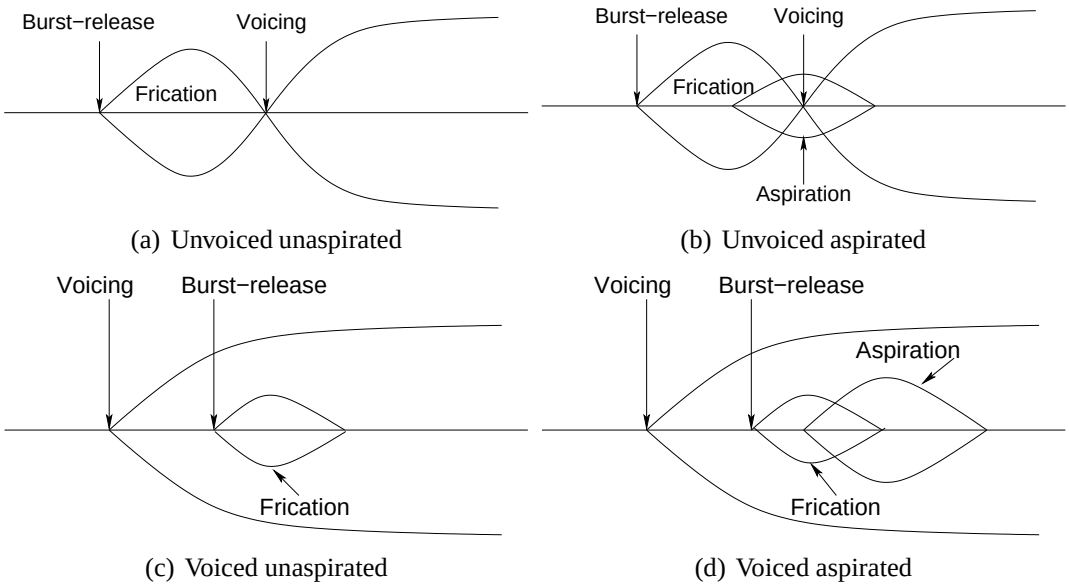
Table 7.1: Stop consonants in Indian languages.

	uVuA	uVA	VuA	VA
Velar	/ka/	/k^{h}a/	/ga/	/g^{h}a/
Post-alveolar	/ʈa/	/ʈ^{h}a/	/ɖa/	/ɖ^{h}a/
Dental	/ta/	/t^{h}a/	/da/	/d^{h}a/
Labial	/pa/	/p^{h}a/	/ba/	/b^{h}a/

7.1 Significance of glottal activity in stop consonant analysis

During speech production, the articulators in the vocal-tract are briefly coupled in a functional manner to produce the acoustic characteristics of speech sounds. For example, the production of bilabial unvoiced stop consonant /p/ requires the following set of actions. The lips are closed by the joint activity of the jaw and the lips. The velum is elevated to seal off the entrance into the nasal cavity. The glottis is widened and the longitudinal tension of the vocal folds is often increased to prevent glottal vibrations. All these articulatory actions contribute to the period of silence in the acoustic signal, and the increase in oral air pressure that is associated with an unvoiced stop consonant. At some point during the phonation of consonant-vowel /pa/, the speaker closes down the glottis, and, given a suitable balance of airflow and muscular tension, the vocal folds begin to vibrate. This shift in the mode of glottal activity occurs more or less abruptly with a change in supralaryngeal articulation, from the closure phase of the stop (/p/) to the progressively more open oral configuration of the following vowel (/a/). The acoustic consequences of this combination of glottal and oral activities depend very much on their relative timing. For example, during the phonation of bilabial voiced stop /ba/, the vocal fold vibration starts during the closure itself, yielding a low-frequency band during the closure leading to a full formant pattern after the release of the stop. The vocal-fold vibration during the oral closure is commonly referred to as *voice-bar*. On the other hand, during the phonation of bilabial unvoiced aspirated stop /p^{h}a/, the onset of vocal fold vibration is delayed until some time after the release. There is an interval between the closure release and vocal fold vibration, when the relatively unimpeded air rushing through the glottis provides the turbulent excitation commonly called *aspiration*. This aspiration phase is characterized by considerable attenuation of the first formant, an effect that can be attributed to the presence of the trachea below the open glottis. The attenuation of the first formant and the accompanying band limited noise may extend into the transition region and beyond. Finally, the intensity of the burst, that is, the plosive excitation of the oral cavity upon release of the stop, may be affected by the glottal closure. Thus, it is reasonable to

suppose that all these acoustic features can be related to the glottal activity. More importantly, the timing of glottal activity with respect to the oral activity is an important clue for discriminating different manners of articulation for a given place of articulation. Fig. 7.1 illustrates the relative placement of the important events in the production of various stop consonants.



(a) Unvoiced unaspirated (b) Unvoiced aspirated

(c) Voiced unaspirated (d) Voiced aspirated

Fig. 7.1: Schematic representation of the important events in the stop consonants

7.1.1 Voice onset time

The instant of onset of vocal fold vibration relative to the release of closure (burst) is the commonly used feature to analyze the manner of articulation in production of stop consonants. The interval between the time of burst release to the time of onset of vocal fold vibration is defined as voice onset time (VOT) [107]. It is important to note that VOT is merely one of the large set of interrelated acoustic consequences of variation in relative time of oral and glottal activities. Abramson and Lisker used the measure of VOT mainly because the onset of glottal pulses seen as vertical striations in wide-band spectrograms is a clear sign that glottis has shifted from the fully open state to the vibratory state [142].

Accurate determination of VOT from acoustic signals is important both theoretically and clinically. From a clinical perspective, the VOT constitutes an important clue for as-

sessment of speech production of hearing impaired speakers [143]. From a theoretical perspective, the VOT of stop consonants often serves as a significant acoustic correlate to discriminate voiced from unvoiced, and aspirated from unaspirated stop consonants. The unvoiced unaspirated stop consonants typically have low and positive VOTs, meaning that the voicing of the following vowel begins near the instant of closure release. The unvoiced aspirated stop consonants followed by a vowel have slightly higher VOTs than their unaspirated counterparts, as the burst is followed by the aspiration noise. The duration of the VOT in such cases is a practical measure of aspiration. The longer the VOT, the stronger is the aspiration. On the other hand, voiced stop consonants have a negative VOT, meaning that the vocal folds start vibrating before the stop is released.

Since the voicing onset is a characteristic of glottal activity, the VOT can be accurately determined from the EGG signal along with speech waveform. Sometimes even with EGG waveform it may be difficult to mark the voicing onset due to 'subcritical' vocal fold vibration in breathy voicing conditions [144, 145]. Moreover, EGG signals are not commonly available in practice. Therefore it is necessary to derive the voicing onset information from the acoustic speech signal itself.

Most commonly used methods for measuring the onset of voicing are based on the onset of periodicity in the acoustic waveform, possibly supplemented by spectrographic analysis [108], especially the onset of visible energy in the first formant [110] or higher formants [111]. One of the issues in using spectrographic information for determining onset of voicing is that there are obvious differences between the latency of voicing onset at different frequencies. For example, vertical striations due to voicing typically appear later in higher formants compared to the first formant. Moreover, in the case of aspirated stops the attenuation of the first formant during aspiration may extend into the following vowel making it difficult to accurately locate the voicing onset. Hence, choosing a unique landmark in the spectrogram as voicing onset is not a trivial task. Comparative study of accuracy and variability of five acoustic (F_0, F_1, F_2, F_3 and periodicity) measures of the voicing onset showed that measurements based on waveform provide the best results [144].

The ideal acoustic measurement of the voicing onset is one that is both accurate and

relatively consistent. The main problem with the above mentioned acoustic measures is that the desired information of the glottal activity is in a very low frequency region (within the F_1), where the energy of the acoustic signal is low compared to the amplitude at other frequencies. In spectrographic analysis, the effects of block processing sometimes limit the visibility of formant features. The presence of noise and voicing in the aspirated (breathy) regions [146], and the low amplitude of the voice-bar in voiced stops make the direct measurements from the waveform difficult. Thus the analysis of stop sounds, especially the voiced aspirated stops, to extract information about the voicing onset remains a challenge.

Since the glottal activity is primarily due to the excitation source of the vocal tract system, it is likely that if the analysis is focused on the excitation component in the speech signal, it may provide new insights into the phonation characteristics present in the signal. It is also desirable to avoid spectral analysis, as it may invariably use block processing, resulting in blurring the details of voicing onset information. In the following section, we propose nonspectral features of the speech signal to study the role of excitation source in the production of stop consonants.

7.2 Excitation-based nonspectral analysis of stop consonants

The primary and most important mode of excitation is due to the activity at the glottis. In normal voiced excitation (called modal voicing), there will be vibrations of the vocal folds resulting in glottal opening, followed normally by an abrupt closure of the vocal folds, and then a closing phase of the glottis, before the glottis is opened again for the next cycle due to build up of pressure from the lungs. Other aspects of glottal activity include vibration with large opening for production of breathy voice, a complete opening for the production of unvoiced sounds, a partial closure of the vocal folds for production of creaky voice, and finally a complete closure of the vocal folds such as for glottal stops. Fig.7.2 illustrates the continuum of phonation types as proposed by Gardon and Ladefoged [146]. We focus on

extraction of the excitation due to glottal activity, and try to derive the acoustic correlates of stop consonants from this excitation information.

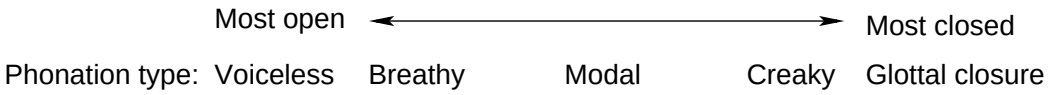


Fig. 7.2: Phonation types [146]

The source of excitation to the vocal-tract system during the production of stop consonants can occur in any of the following two modes: (a) The regions where the vocal-tract is excited by vocal fold vibration. (b) The regions where the vocal-tract is excited by unimpeded airflow rushing through open glottis (burst and aspiration). The regions of vocal-fold vibration can be extracted from zero-frequency filtered signal proposed in Chapter 3. Using the zero-frequency resonator, the information about vocal fold vibration can be extracted irrespective of the vocal-tract dynamics, i.e., the place of articulation of the stop consonant and the nature of the following vowel. Hence the excitation information as reflected in the filtered signal can be used to detect the regions of vocal fold vibration and the precise instant of voicing onset.

During release of the closure and the aspiration regions, the vocal-tract system is excited by a rush of air through open glottis with irregular vocal fold vibrations. This region reflects in the spectrogram as band-limited noise. Linear prediction analysis is employed to analyze the noisy component of the excitation source. The linear prediction residual, derived by inverse filtering the speech signal, is used as an approximation to the excitation component in the speech signal. The choice of the LP order, the frame size and the frame rate used for the LP analysis are not critical. In this work, a 10^{th} order LP analysis is performed on frames of width 20 ms shifted by 10 ms. The ratio of the energy of the LP residual and the speech signal for every block of the frame size and for every sample shift is computed. The normalized error for each sample shift is computed as

$$\eta[n] = \frac{\sum_{m=n-N/2}^{n+N/2} e^2[n+m]}{\sum_{m=n-N/2}^{n+N/2} x^2[n+m]}, \tag{7.1}$$

where $N + 1$ is the total number of samples in each frame. The resulting plot is called the normalized error as function of the sample index, and it is used to distinguish the excitation information due to noisy voiced segments and clean voiced segments. Note that the spectral information in LPCs is ignored here, by considering only the residual. Though block processing is used to derive the LP residual, the effects of blocking are insignificant for the analysis of the acoustic correlates of the stop consonants under consideration.

The filtered signal $y[n]$ and the normalized error $\eta[n]$ are used to represent the excitation information derived from the speech signal. This information, together with the speech signal and its wideband spectrogram are plotted for voiced aspirated syllable /g^{h}a/ in Fig. 7.3, to study the acoustic correlates, especially voicing onset, burst release and aspiration. The waveform and the spectrogram are given only for reference. Their features are not used to derive the acoustic correlates of the stop consonants. From the filtered output in Fig. 7.3(c), it is easy to determine the onset of glottal activity (marked as V) and the ending of the glottal activity. In the initial voicing region (voice-bar) the filtered output is relatively high compared to the amplitude of the signal in that region. The value of the normalized error in this region is low. At the release of the burst (marked as B) of the stop sound there is a significant increase in normalized error shown in Fig. 7.3(d). The burst duration (region from B to A) cannot be seen either in the waveform or in the filtered output. During aspiration (region from A to M) the filtered output is large indicating significant glottal activity, and is irregular (from 0.15 s to 0.22 s) due to noisy plosive release. But the filtered output alone is not sufficient to distinguish the glottal activity during aspiration and the following modal voicing (region after M), as it appears nearly periodic in the transition region. The evidence from the normalized error can also be exploited to distinguish aspiration region from the modal voicing region. The normalized error is significantly high in the aspirated region compared to that in the modal voicing region. Hence the filtered output together with the normalized error can be used to analyze different manners of articulation of stop consonants.

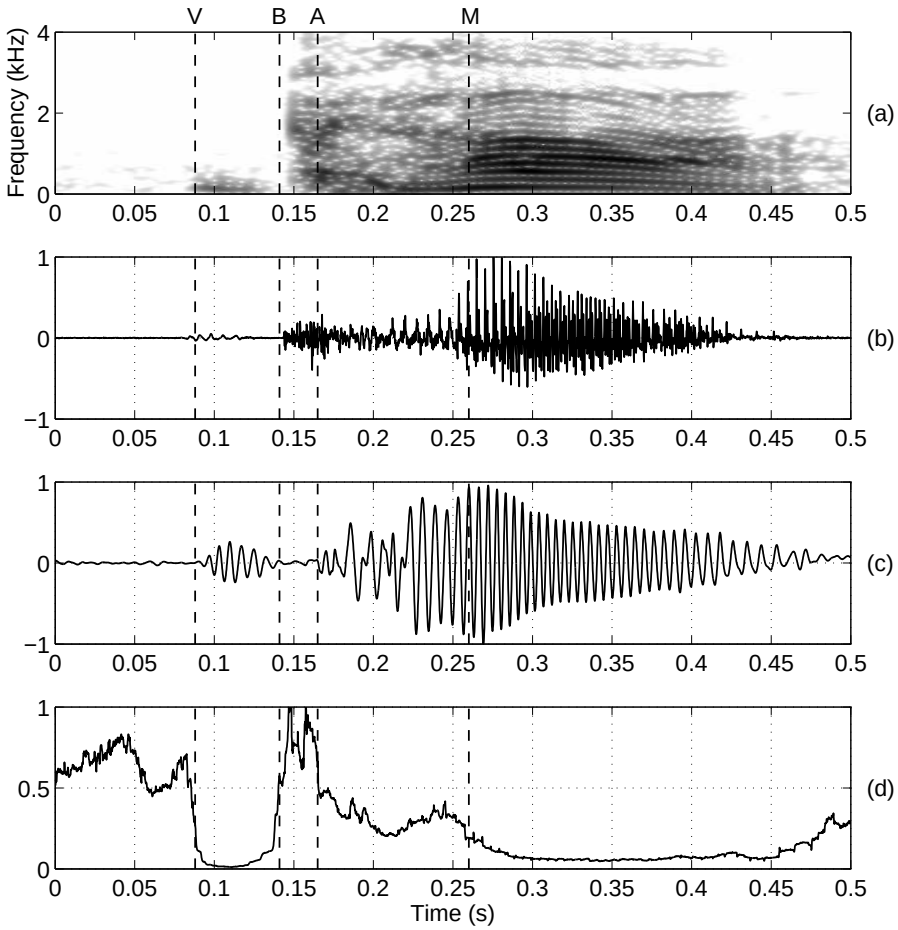


Fig. 7.3: Illustration of excitation source features for voiced aspirated stop consonant /g^{h}a/. (a) Wideband spectrogram, (b) speech signal, (c) filtered output, and (d) normalized error. The wideband spectrogram and waveform are used only for reference.

7.3 Analysis of manner of articulation for stop consonants

Isolated utterances of the CV units listed in Table 7.1 are used in this study. The utterances are produced by three male speakers. Each utterance is repeated 5 times. The speech signal is sampled at 8 kHz. The data is collected for five different vowel endings (/a/,/i/,/u/,/e/, and /o/) for the 16 stop consonants. All the data was collected in a laboratory environment using a close-speaking microphone. Thus the data can be considered as clearly articulated clean data.

Fig. 7.4 shows the waveform, filtered output and the normalized error plots for the four velar stops /ka/, /k^{h}a/, /ga/ and /g^{h}a/. The plots of the filtered outputs in each case

clearly show the regions of glottal activity. The following observations can be made to distinguish the four categories:

(a) Unvoiced unaspirated: There is sudden increase in the normalized error at the release of the burst. The normalized error is large in the short burst region relative to the modal voicing region.

(b) Unvoiced aspirated: There is sudden increase in the normalized error at the release of the burst. The large $\eta[n]$ is extended over the aspirated region due to the presence of breathy noise. The $\eta[n]$ is low in the modal voicing region. The filtered output is *somewhat less periodic* in the aspirated region.

(c) Voiced unaspirated: There is relatively large output in the filtered signal due to initial voicing compared to the relatively small amplitude in the waveform. There is increase in the $\eta[n]$ during the short burst region.

(d) Voiced aspirated: The filtered output is large during the initial voicing region, and then in the aspirated and modal voicing regions. There is a dip in the filtered output at the burst release. But the $\eta[n]$ has an abrupt raise at the burst release, followed by large $\eta[n]$ in the aspirated region due to breathy noise.

Similar features are observed in stop consonants produced at other places of articulation. The acoustic correlates derived from post-alveolar, dental and bilabial stop consonants are shown in Fig. 7.5, Fig. 7.6 and Fig. 7.7, respectively. The burst release instant (marked as B) is determined as the instant where there is a large increase in $\eta[n]$. The starting instant of the glottal activity (marked as V) is derived from the filtered output. In all the cases the burst release instant (B) and voicing onset (V) can easily be identified. The interval between these two instants is used as VOT in this study.

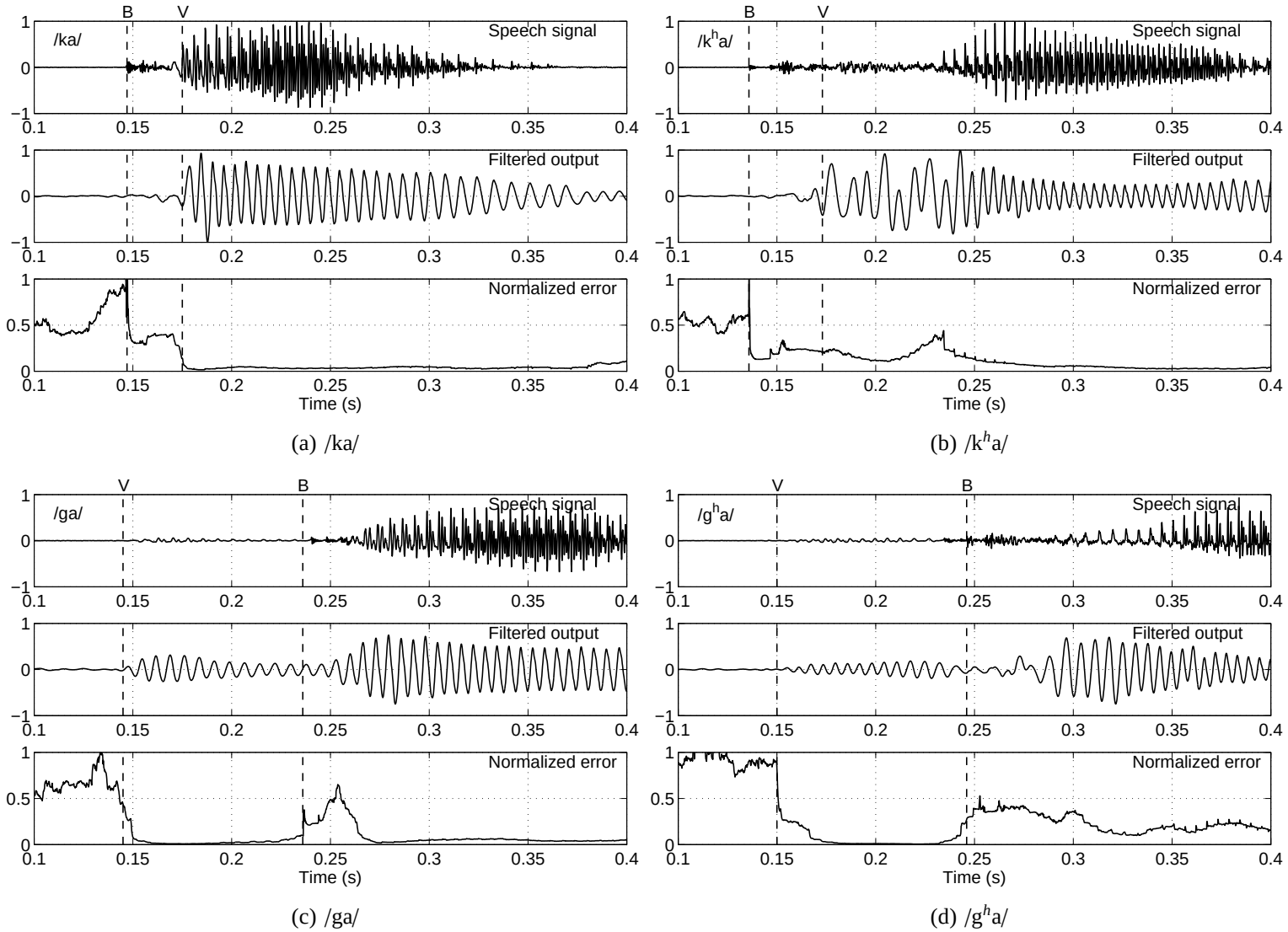


Fig. 7.4: The speech signal, filtered output, and the normalized error for four different velar stop sound units

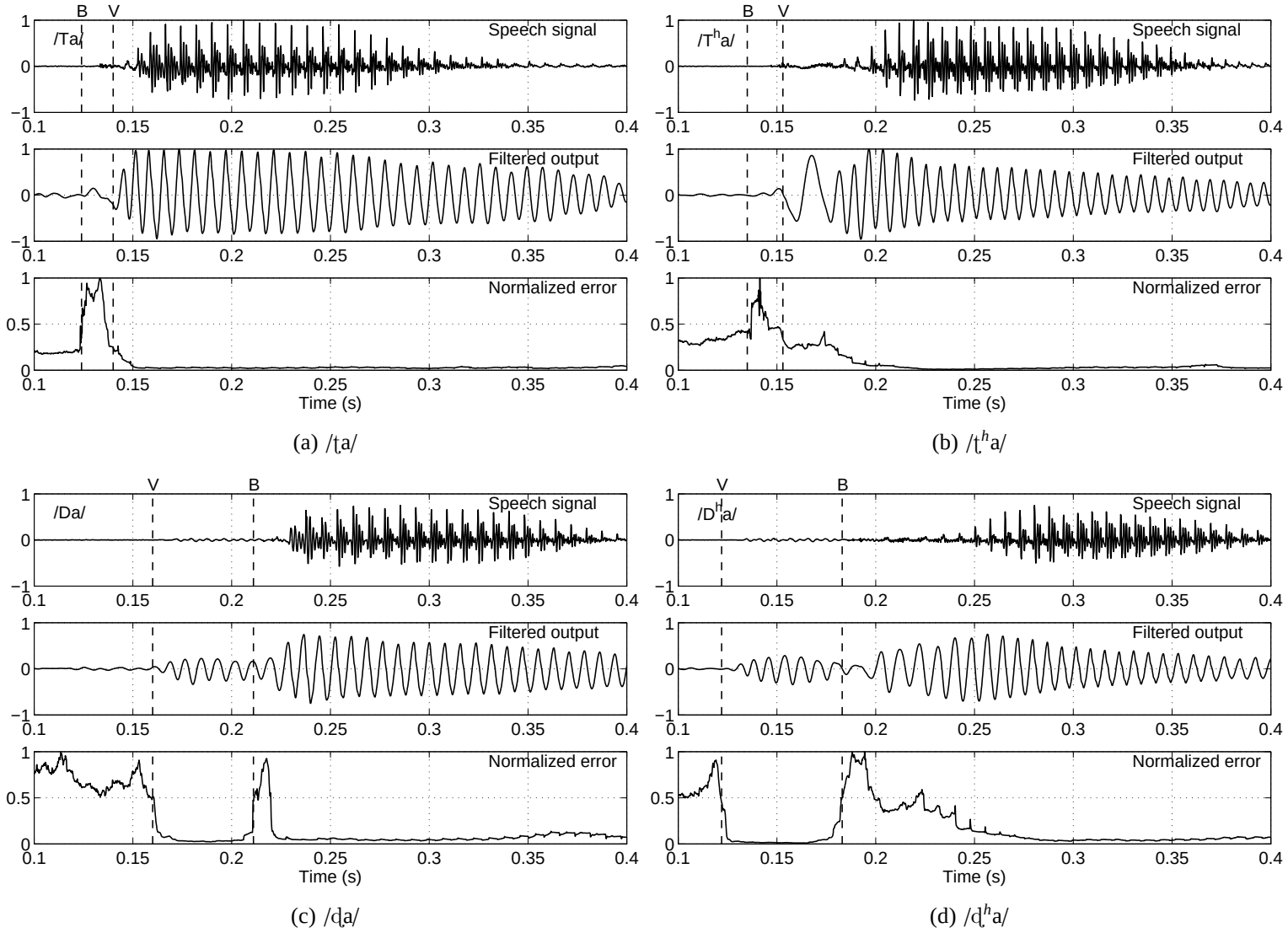


Fig. 7.5: The speech signal, filtered output, and the normalized error for four different post-alveolar stop sound units

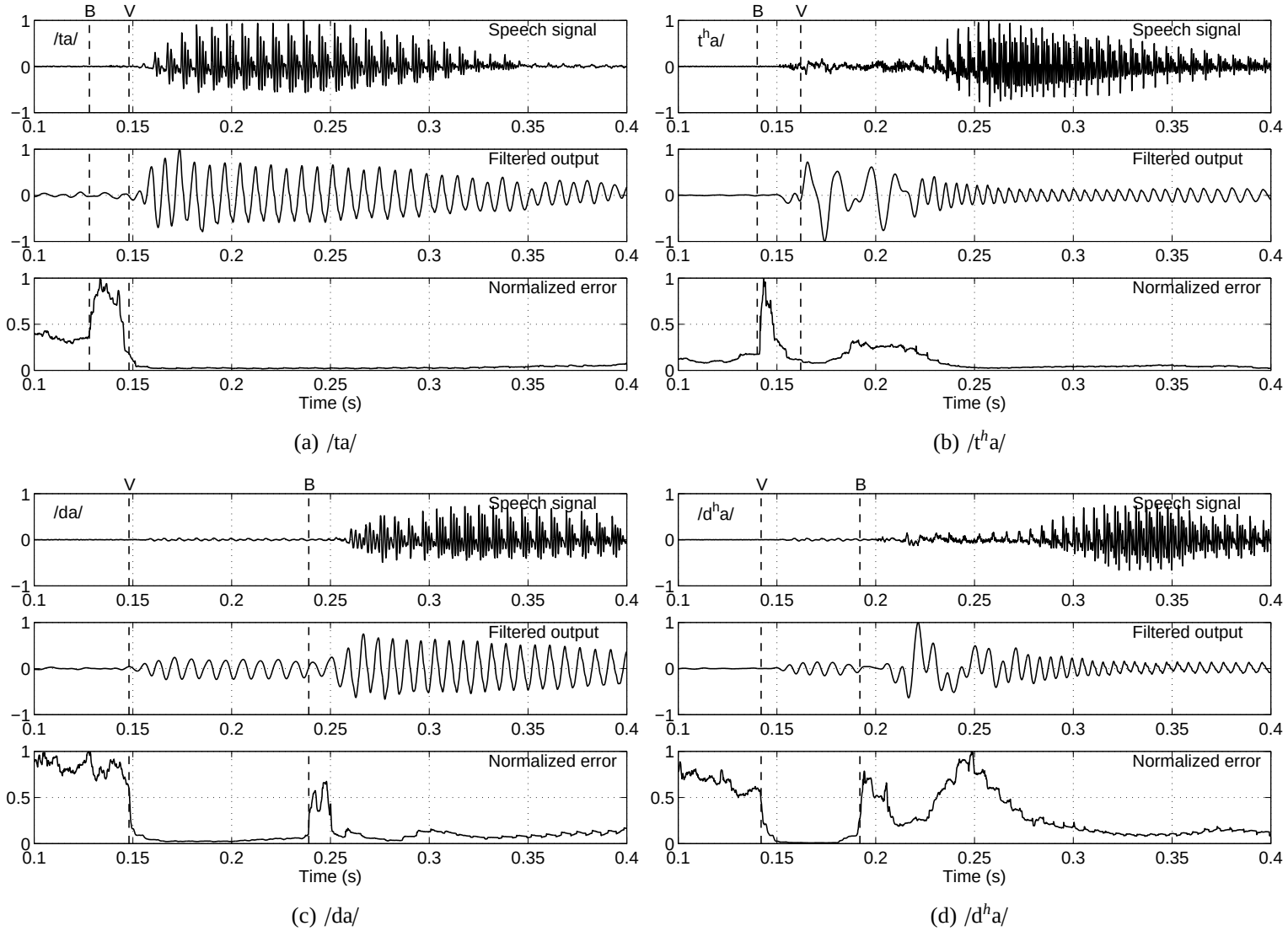


Fig. 7.6: The speech signal, filtered output, and the normalized error for four different dental stop sound units

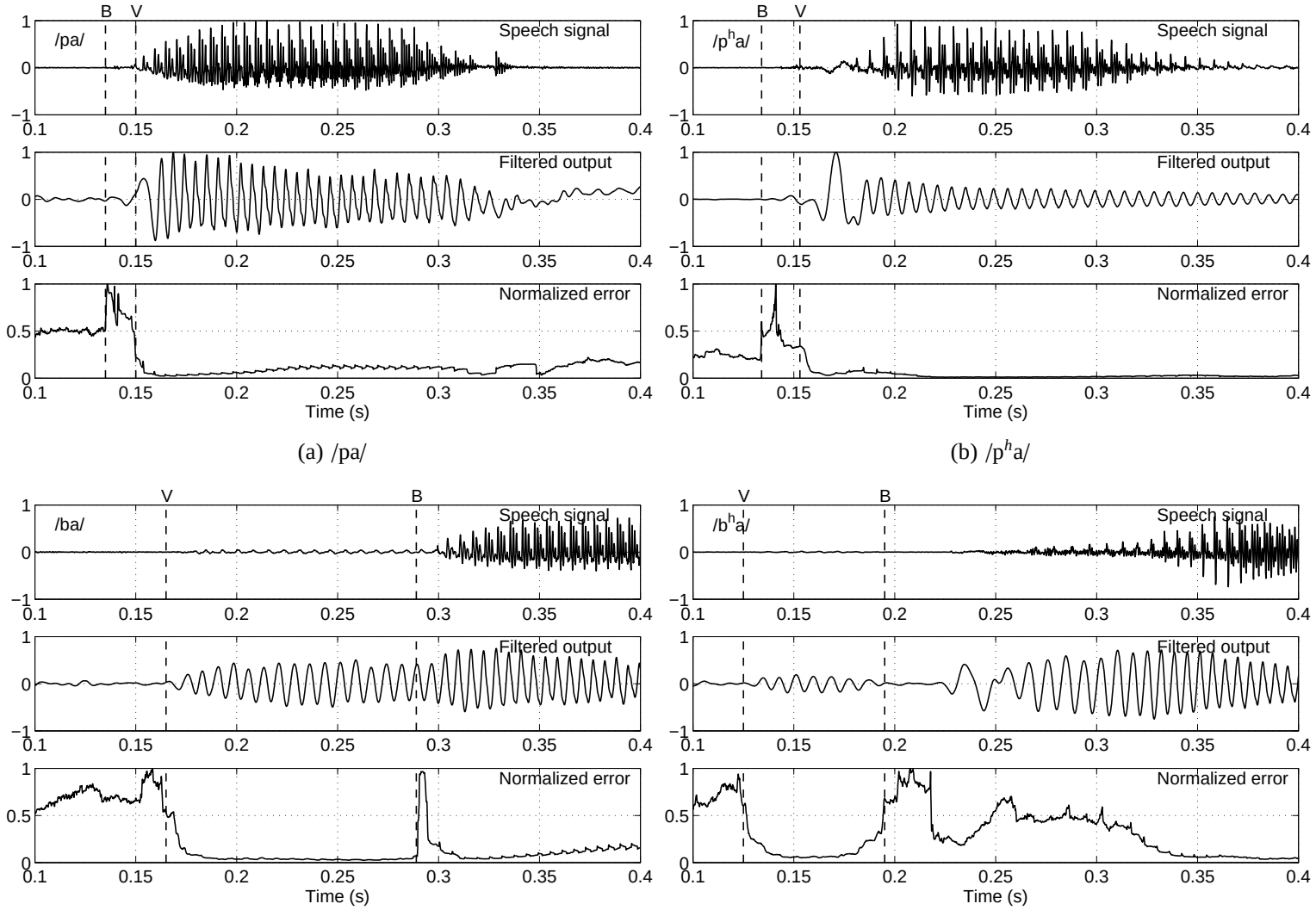


Fig. 7.7: The speech signal, filtered output, and the normalized error for four different bilabial stop sound units

All the above observations are valid for stop consonants with different vowel endings. Table 7.2 shows the average durations of VOT for different categories of CV units ending with the vowel /a/. All the VOTs are obtained manually from the filtered output and the normalized error as shown by the markers in Fig. 7.4. For unvoiced stops, the burst release (B) takes place before the onset (V) of the glottal activity. The interval between these two is the VOT, and is also the burst duration in this case. The VOT is generally larger for velar stops compared to the other three categories. The relatively smaller volume of the cavity behind the point of constriction in velar stops causes a greater pressure, which will take longer time to fall and allow an adequate transglottal pressure for the initiation of the vocal folds vibration [147]. The extent of articulatory contact area in dental and velar stops is more, resulting in a slower release because of the Bernoulli effect pulling the articulators together [147]. As the articulators come apart more slowly, there is a longer time before an appropriate transglottal pressure is produced. As a result, the durations of VOT for /ka/ and /ta/ are longer than those for /ʈa/ and /pa/. The VOT durations for aspirated stop consonants are consistently longer than their unaspirated counterparts, as the aspiration region follows the closure release in case of aspirated stops. The precise duration of aspiration is difficult to measure from the plot of the normalized error, as the effect of aspiration extends into the following vowel.

In the voiced stop consonants, the voicing onset (V) due to the glottal activity occurs before the closure release (B), resulting in a negative VOT duration. The VOT durations for different voiced stop consonants are given in Table 7.2. Notice that there is no clear distinction between the unaspirated stop consonants and their aspirated counterparts. In fact, Abramson had pointed that, VOT may distinguish the voiced aspirated stop consonants from unvoiced stop consonants, in the context of Indian languages, but certainly not from the voiced unaspirated stops [142].

In the case of voiced stop consonants, the burst duration is different from VOT. In our study, it is observed that the burst durations provide some evidence regarding the presence of aspiration in voiced stop consonants. The burst durations for different voiced stop consonants are given in Table 7.3. The burst durations for the voiced aspirated sounds are consistently longer compared to their unaspirated counterparts. In some cases, the voicing

during the closure (voice-bar) continues to the following the vowel, and the burst release due to the closure release comes during the voicing itself. The beginning and end of the burst can be seen clearly in normalized error, while the simultaneous glottal activity can be observed from the filtered signal, especially for voiced unaspirated stops. For voiced aspirated stops, the burst duration is sometimes difficult to identify as the breathiness during aspiration and the burst duration may overlap. The normalized error may remain large throughout the aspiration region, even though there is closure release during that period.

Table 7.2: The average (across three speakers) durations of VOT in stop consonants (in ms)

Unvoiced	ka	k^{h}a	ʈa	ʈ^{h}a	ta	t^{h}a	pa	p^{h}a
Duration	32	36	16	20	23	29	12	15
Voiced	ga	g^{h}a	ɖa	ɖ^{h}a	da	d^{h}a	ba	b^{h}a
Duration	-82	-65	-81	-86	-74	-73	-60	-52

Table 7.3: The duration of burst in voiced stop consonants (in ms)

Voiced	ga	g^{h}a	ɖa	ɖ^{h}a	da	d^{h}a	ba	b^{h}a
Duration	19	23	9	16	12	17	7	13

7.4 Summary

In this work, we have attempted to make a case for nonspectral methods for analysis of stop consonants. The methods are intended to focus on excitation characteristics during the production of stop consonants. We have proposed the use of zero-frequency filtered signal to extract the region of glottal activity, and the normalized error from LP residual to determine the noise regions of excitation during burst release and during aspiration. The onset of voicing can be detected from the filtered signal and the instant of burst release can be detected from the normalized error. Voicing onset time for all the stop consonants are measured using these two features together.

Chapter 8

Summary and Conclusions

8.1 Summary of the work

During the production of voiced speech, impulse-like excitation is delivered to the vocal-tract system at the instant of glottal closure. The instant of glottal closure referred to as epoch, and the rate of glottal closure referred to as strength of excitation (at the epoch) form important features of the excitation source. In this thesis, we proposed a novel method to extract the epoch locations and their strengths. The proposed approach does not depend on the characteristics of the vocal-tract system. Most of the existing methods for epoch extraction rely on modeling the vocal-tract system as a linear filter, and then inverse filtering the speech signal to extract the excitation source. On the contrary, the proposed approaches exploit the impulse-like nature of excitation in the sequence of glottal cycles to extract the epoch locations and to estimate their strengths.

The impulse-like excitation to the vocal-tract system causes a discontinuity in the speech signal whose effect spreads uniformly across the frequency domain. The time-instants of these discontinuities may not be evident from the speech signal because of the time-varying response of the vocal-tract system. In this work, we attempted to confine the analysis to a narrow band of frequencies, to highlight the effect due to the discontinuity, by filtering the speech signal through a resonator with a narrow bandwidth. We demonstrated that the instantaneous frequency computed around a carefully chosen center frequency

gives locations of the discontinuities. In this approach, the choice of center frequency critically depends on the vocal-tract configuration.

The discontinuity due to the impulse-like excitation is reflected uniformly in the frequency domain, including at the zero-frequency. The influence of the vocal-tract system is relatively less at the zero-frequency, as the vocal-tract system has resonances at much higher frequencies. Hence, we use a zero-frequency resonator to extract the epoch locations and their strengths. The method involves passing the speech signal through a cascade of two ideal resonators located at the zero-frequency. The filtered signal is derived from the output of the resonators by subtracting the local mean computed over an interval corresponding to the average pitch period. The sharper zero-crossings in the filtered signal are shown to coincide with the instants of significant excitation within each glottal cycle.

The contribution of the vocal-tract system around the zero-frequency is significantly less compared to the contribution due to the impulse-like excitation. Hence, the narrow-band nature of the zero-frequency resonator is exploited for estimating the strength of excitation at the epoch from the speech signal. It is observed that the slopes of the filtered signal around epoch locations closely follow the amplitudes of the negative peaks in the differenced EGG signal. The strength of excitation is significant in the regions of vocal fold vibration where the vocal-tract system is excited by impulse-like excitation. In the unvoiced regions, the filtered signal is close to zero due to the absence of impulse-like excitation. A method is proposed using the energy of the zero-frequency resonator to identify the regions of vocal fold vibration from speech signal.

Using the epoch locations as anchor points within each glottal cycle, a method to estimate the instantaneous fundamental frequency of voiced speech segments is proposed. The fundamental frequency is estimated as the reciprocal of the interval between two successive epoch locations, derived from filtered speech signal. Since the proposed method is based on the point property of epoch and does not involve any block processing, it provides cycle-to-cycle variations in pitch during voicing. Hence we call the resulting estimate as instantaneous fundamental frequency as opposed to ‘mean pitch’ derived from conventional block processing applications. Errors due to spurious zero-crossings in the weak voiced regions are corrected using the filtered signal of Hilbert envelope of the

speech signal.

Since the proposed method of pitch estimation exploits the impulse-like excitation characteristic, the method does not work if there are additional impulses due to echoes or reverberation or overlapping speech from competing speakers. The zero-frequency filtered signal of a reverberant speech signal contains several spurious zero-crossings due to discontinuities introduced by the reflected components. In this work, a method is proposed for pitch estimation in reverberant environment from speech signals collected using a pair of spatially separated microphones. The spatial separation of microphones results in a fixed time-delay of arrival of speech signals at the pair of microphones. A method based on excitation source is discussed for time-delay estimation. The crosscorrelation of segments of Hilbert envelopes of the LP residuals from the two microphone signals is used for time-delay estimation. The delay compensated Hilbert envelopes are coherently added to emphasize the regions around the epochs while reducing the effect of reverberation. The pitch is estimated from the zero-crossings of the filtered signal of the coherently added Hilbert envelope.

In this work, a method is proposed for multipitch extraction from speech signals collected using a pair of microphones. One important point to be noted in the multispeaker environment is that, as the speakers are spatially distributed, unique time-delay is associated with each speaker. In this work, the differences in the time-delays for different speakers are exploited to emphasize the epochs due to individual speakers. It is observed that the minimum of the delay compensated Hilbert envelopes emphasizes the epochs due to a given speaker, and deemphasizes the epochs due to the other speaker. The individual pitch tracks are estimated from the zero-crossings of the filtered signal of the speaker-specific Hilbert envelopes.

Finally, we made an attempt to study the usefulness of excitation source information to analyze the manner of articulation of stop consonants. Two measures of excitation source investigated in this study are the filtered speech signal and the normalized error. The filtered speech signal is used to characterize the excitation information during vocal-fold vibration. The normalized error derived from LP analysis is used to highlight the regions of noisy excitation caused by a rush of air through open glottis during closure

release and aspiration. It is observed that these two features jointly highlight important events, like onset of voicing and instant of closure release, in the stop consonants. Using the two excitation source features, the voice onset time and the burst durations of stop consonants in Indian languages were measured.

8.2 Major contributions of the work

The important contribution of the research work reported in this thesis is the extraction and processing of excitation source information of speech. Extraction of excitation source information requires suppressing the vocal-tract response from the speech signal, which is not a trivial task. In this work, we proposed methods to extract the epoch locations and their strengths by confining the analysis to a narrow band around the zero-frequency where the effect of vocal-tract system is significantly low compared to the impulse-like excitation. The major contributions of this thesis are:

- Studies on suitability of instantaneous frequency for epoch extraction
- Epoch extraction from speech signals using zero-frequency resonator
- Estimation of the strengths of excitation at epochs from speech signals.
- Glottal activity detection from speech signals
- Estimation of instantaneous fundamental frequency of voiced speech
- Estimation of time-delay from multimicrophone data using excitation source information
- Pitch extraction in reverberant environment from multimicrophone data
- Multipitch extraction from multimicrophone data
- Analysis of manner of articulation of stop consonants using excitation source information

8.3 Directions for future work

- The proposed zero-frequency based method for epoch extraction may not work well on speech data collected over telephone channels, and high pass filtered speech signals, where the low frequency components are deliberately attenuated. In these cases, the epochs may be extracted by confining the analysis to higher frequencies than at zero-frequency. Instantaneous frequency of the speech signal filtered around a carefully chosen center frequency is shown to indicate approximate locations of the epochs. The choice of the center frequency critically depends on the time-varying response of the vocal-tract system. Methods have to be explored to adaptively choose the center frequency of the resonator from the speech segment under consideration. It was observed that there exist multiple center frequencies which can be used for epoch extraction. The evidences from different center frequencies can be combined to obtain a robust and accurate epoch locations.

- The strengths of the excitation can be used to estimate the shimmer which is known to be a speaker-specific characteristic.

- The fundamental frequency estimation proposed in this work provides the period to period variations in pitch during the production of voiced speech. These finer variations can be used to estimate the jitter which is a speaker-specific characteristic.

- The glottal activity detection approach presented in this work, uses the energy of the filtered signal. The performance of the method may be improved by using the knowledge of the intervals between successive zero-crossings. Regularity of the zero-crossings observed in the voiced regions will be absent in the noisy regions.

- The proposed methods for processing multimicrophone data uses the speech signals collected over two microphones only. The performance of these methods may be improved by collecting speech signal using more number of microphones, and then selecting the set of microphones closer to individual speakers for processing.

- In the proposed approach for stop consonant analysis, the voicing onset time and burst durations are measured manually by observing the filtered signal and the nor-

malized error. Methods have to be developed to detect the onset of voicing and instant of burst release automatically using the proposed excitation source features.

- The proposed excitation source features may be useful in the analysis of consonant-vowel units. The proposed method of glottal activity may be useful in accurate detection of vowel-onset point, the instant at which transition from consonant to vowel occurs, in a consonant-vowel unit.

References

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

[2] 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. 27, no. 4, pp. 309–319, Aug. 1979.

[3] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

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

[5] D. Veeneman and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Signal Processing*, vol. 33, pp. 369–377, Apr. 1985.

[6] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp. 313–327, Jul. 1998.

[7] C. Hamon, E. Moulines, and F. Charpentier, "A diphone synthesis system based on time domain prosodic modifications of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, May 1989, pp. 238–241.

[8] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 3, pp. 972–980, May 2006.

[9] B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswamy, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 6, pp. 1110–1118, Nov. 2005.

[10] B. Yegnanarayana and P. S. Murty, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[11] A. Neocleous and P. A. Naylor, "Voice source parameters for speaker verification," in *Proc. Eur. Signal Process. Conf.*, 1998, pp. 697–700.

[12] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 7, pp. 569–586, Sept. 1999.

[13] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–56, Jan. 2006.

[14] E. R. M. Abberton, D. M. Howard, and A. J. Fourcin, "Laryngographic assessment of normal voice: A tutorial," *Clinical Linguistics and Phonetics*, vol. 3, no. 3, pp. 263–296, 1989.

[15] J. van den Berg, "Myoelastic-aerodynamic theory of voice production," *Journal of Speech and Hearing*, vol. 1, pp. 227–244, 1958.

[16] K. N. Stevens, "Physics of laryngeal behavior and larynx models," *Phonetica*, vol. 34, pp. 264–279, 1977.

[17] R. C. Scherer, D. G. Druker, and I. R. Titze, *Vocal Physiology: Voice Production Mechanisms and Functions*. New York: Raven Press Ltd., 1988.

[18] F. L. E. Lecluse, "Elecroglottography," Dissertation, Univ. of Rotterdam, 1977.

[19] W. Hess and H. Indefrey, "Accurate time-domain pitch determination of speech signals by means of a laryngograph," *Speech Communication*, vol. 6, pp. 55–68, 1987.

[20] M. Huckvale. (2000) Speech filing system: Tools for speech research. [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/

[21] 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.

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

[23] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 637–655, 1971.

[24] T. V. Ananthapadmanabha and G. Fant, "Calculations of true glottal volume-velocity and its components," *Speech Communication*, vol. 1, pp. 167–184, 1982.

[25] Y. M. Cheng and O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 1805–1815, Dec. 1989.

[26] J. G. McKenna, "Automatic glottal closed-phase location and analyis by Kalman filtering," in *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Aug. 2001.

[27] 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.

[28] C R Jankowski Jr, T. F. Quatieri, and D. A. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Detroit, MI, USA, May 1995, pp. 325–328.

[29] J. L. Navarro-Mesa, E. Lleida-Solano, and A. Moreno-Bilbao, "A new method for epoch detection based on the cohen's class of time frequency representations," *IEEE Signal Process. Lett.*, vol. 8, no. 8, pp. 225–227, Aug. 2001.

[30] V. N. Tuan and C. d'Alessandro, "Robust glottal closure detection using the wavelet transform," in *Proc. European Conf. Speech Processing, Technology*, Budapest, Sep. 1999, pp. 2805–2808.

[31] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 6, pp. 562–570, Dec 1975.

[32] 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*, Detroit, USA, May 1995, pp. 776–779.

[33] P. S. Murty and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant excitation from speech signals," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 6, pp. 609–619, Nov. 1999.

[34] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 2, pp. 456–466, Mar. 2006.

[35] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 11, Orlando, FL, May 2002, pp. 349–352.

[36] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[37] P. Alku, J. Vintturi, and E. Vilkman, "On the linearity of the relationship between the sound pressure level and the negative peak amplitude of the differentiated glottal flow in vowel production," *Speech Communication*, vol. 28, pp. 269–281, 1999.

[38] J. Gauffin and J. Sundberg, "Spectral correlates of glottal voice source waveform characteristics," *Journal of Speech and Hearing Research*, vol. 32, pp. 556–565, 1989.

[39] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Communication*, vol. 18, pp. 131–138, 1996.

[40] ——, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia Phoniatr Logop*, vol. 48, pp. 240–254, 1996.

[41] P. Alku, T. Bakstrom, and E. Vikman, "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, Aug. 2002.

[42] P. Alku, M. Airas, E. Bjorkner, and J. Sundberg, "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity," *J. Acoust. Soc. Amer.*, vol. 120, no. 2, pp. 1052–1062, Aug. 2006.

[43] T. Backstrom, P. Alku, and E. Vilkman, "Time-domain parameterization of the closing phase of glottal airflow wavefrom from voices over a large intensity range," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 3, pp. 186–192, Mar. 2002.

[44] E. Holmberg, R. Hillman, and J. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice," *J. Acoust. Soc. Amer.*, vol. 84, pp. 511–529, 1988.

[45] P. Alku, "Glottal wave analysis with pitch synchronous adaptive inverse filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.

[46] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, 1991.

[47] L. Mary and B. Yegnanarayana, "Prosodic features for speaker verification," in *Proc. Interspeech - 2006*, Pittsburgh, PA, USA, Sep. 2006, pp. 917–920.

[48] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46, no. 3-4, pp. 455–472.

[49] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Accepted for publication in Speech Communication*.

[50] D. Vergyri, A. Stolcke, V. R. R. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 2003, pp. 208–211.

[51] A. Waibel, *Prosody and Speech Recognition*. San Mateo, California: Morgan Kaufmann, 1988.

[52] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. Int. Conf. Spoken Language Processing*, Denver, Sep. 2002, pp. 2037–2040.

[53] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154.

[54] G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Computational Linguistics*, vol. 27, no. 1, pp. 31–57, Mar. 2001.

[55] W. J. Hess, *Pitch Determination of Speech Signals*. Berlin: Spinger-Verlag, 1983.

[56] D. J. Hermes, "Pitch analysis," in *Visual Representations of Speech Signals*, M. Cooke, S. Beet, and M. Crawford, Eds. New York: Wiley, 1993, pp. 3–25.

[57] W. J. Hess, "Pitch and voicing determination," in *Advances in speech signal processing*, S. Furui and M. M. Sondhi, Eds. New York, USA: Marcel Dekker, 1992, pp. 3–48.

[58] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Prentice Hall, 1999.

[59] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, no. 1, pp. 24–33, Feb. 1977.

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

[61] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electrocoust.*, vol. AU-16, pp. 262–266, Jun. 1968.

[62] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. 20, pp. 367–377, Dec. 1972.

[63] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 1, pp. 2–8, Feb. 1976.

[64] P. Boersma, "Accurate short-term analysis of fundamental frequency and the hormincs-to-noise ratio of a sampled sound," *Proc. Institute of Phonetic Sciences*, vol. 17, pp. 97–110, 1993.

[65] B. S. Atal, "Automatic speaker recognition based on pitch contours," PhD Thesis, Polytechnic institute of Brooklyn, 1968.

[66] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, no. 5, pp. 353–362, Oct. 1974.

[67] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. Signal Processing*, vol. 39, no. 1, pp. 40–48, Jan. 1991.

[68] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442–448, Aug. 1969.

[69] N. J. Miller, "Pitch detection by data reduction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 72–79, Feb. 1975.

[70] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 917–924, Mar. 1992.

[71] M. S. Obaidat, C. Lee, B. Sadoun, and D. Nelson, "Estimation of pitch period of speech signal using a new dyadic wavelet transformation," *Journal of Information Sciences*, vol. 119, pp. 21–39, 1999.

[72] S. Kadambe and G. F. Boudreaux-Barlets, "A comparison of wavelet functions for pitch detection of speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Toronto, Canada, May 1991, pp. 449–452.

[73] M. S. Obaidat, A. Bradzik, and B. Sadoun, "A performance evaluation study of four wavelet algorithms for pitch period estimation of speech signals," *Journal of Information Sciences*, vol. 112, pp. 213–221, 1998.

[74] S. Kadambe, "The application of time frequency and time-scale representations in speech analysis," PhD Thesis, Dept. of Elect. Eng., Univ. of Rhode Island, 1991.

[75] P. K. Ghosh, A. Ortega, and S. Narayan, "Pitch period estimation using multipulse model and wavelet transformation," in *Proc. Interspeech 2007*, Antwerp, Belgium, Aug. 2007, pp. 2761–2764.

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

[77] D. J. Hermes, "Measurement of pitch by subhormonic summation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, Jan. 1988.

[78] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1995, pp. 756–759.

[79] H. Kawahara, H. Katayose, A. d. Cheveigne, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f_0 and periodicity," in *Proc. European Conf. Speech Processing, Technology*, Budapest, Hungary, Sep. 1999, pp. 2781–2784.

[80] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, and K. Shikano, "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 907–910.

[81] T. Nakatani and T. Irino, "Robust fundamental frequency estimation against background noise and spectral distortion," in *Proc. Int. Conf. Spoken Language Processing*, Denver, Sep. 2002, pp. 1733–1736.

[82] ——, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3690–3700, Dec. 2004.

[83] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 5, pp. 418–423, Oct. 1976.

[84] T. W. Parks and J. D. Wise, "Maximum likelihood pitch estimation," in *Decision and Control including the 16th Symposium on Adaptive Processes and A Special Symposium on Fuzzy Set Theory and Applications, 1977 IEEE Conference on*, pp. 1092–1095.

[85] R. McAulay, "Maximum likelihood pitch estimation using state-variable techniques," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1978, pp. 12–14.

[86] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum aposteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 1, pp. 76–87, Jan. 2004.

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

[88] G. Carter, "Variance bounds for passively locating an acoustic source with a symmetric line array," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 922–926, 1977.

[89] J. H. DiBiase, H. F. Silverman, and M. S. Branstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Branstein and D. B. Ward, Eds. New York, NY, USA: Springer, 2001, ch. 8, pp. 157–180.

[90] W. Hahn, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inform. Theory*, vol. 19, no. 5, pp. 608–614, Sep. 1973.

[91] S. Haykin, *Adaptive Filter Theory*, 2nd ed. New Jersey: Prentice Hall, 1991.

[92] 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, no. 1, pp. 45–50, Jan. 1997.

[93] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of timedelay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[94] A. Stephenne and B. Champagne, "Cepstral prefiltering for time-delay estimation in reverberant environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Detroit, MI, USA, May 1995, pp. 3055–3058.

[95] ——, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, no. 3, pp. 253–266, 1997.

[96] O. M. M. Mitchell, C. A. Ross, and G. H. Yates, "Signal processing for a cocktail party effect," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 656–660, 1971.

[97] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, no. 4, pp. 911–918, Oct. 1976.

[98] C. K. Lee and D. G. Childers, "Cochannel speech separation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 274–280, Jan. 1988.

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

[100] 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.

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

[102] 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, no. 1, pp. 56–69, Jan. 1990.

[103] S. I. Amari and A. Cichocki, "Adaptive blind signal processing - neural network approaches," *Proc. IEEE*, vol. 86.

[104] S. Choi and A. Cichocki, "Blind separation of nonstationary and temporally correlated sources from noisy mixtures," in *NNSP'*, Sydney, Australia, Dec 2000, pp. 405–414.

[105] S. Choi, H. Hong, H. Glotin, and F. Berthommier, "Multichannel signal separation for coktail party speech recognition: A dynamic recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Beining, China, May 2000, pp. 83–87.

[106] 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.

[107] A. S. Ambramson and L. Lisker, "Voice onset time in stop consonants: acoustic analysis and synthesis," in *Proc. 5th International congress on phonetic sciences*, Liege, 1965, p. A51.

[108] A. S. Abramson, "Laryngeal timing in Karen obstruents," in *Producing speech:Contemporary issues, for Katherine Safford Harris*, Bell-Berti and L. J. Raphael, Eds. New York: AIP Press, 1995, pp. 155–165.

[109] L. L. Koenig, "Distributional characteristics of vot in children's voiceless aspirated stops and interpretation of developmental trends," *Journal of Speech, Language, Hearing*, vol. 44, pp. 1058–1068, 2001.

[110] G. E. Peterson and I. Lehiste, "Duration of syllabic nuclei in English," *J. Acoust. Soc. Amer.*, vol. 32, pp. 693–703, 1960.

[111] D. H. Klatt, "Voicing onset time, frication, aspiration in word initial consonant clusters," *Journal of Speech, Hearing*, vol. 18, pp. 686–706, 1975.

[112] L. Lisker and A. Abramson, "A cross language study of voicing in initial stops: Acoustic measurements," *Word*, vol. 20, pp. 384–422, 1967.

[113] R. D. Kent and C. Read, *The acoustic analysis of speech*, 2nd ed. San Diego: Singular, 2002.

[114] P. Lieberman and S. Blumstein, *Speech Physiology, Speech Perception, and Acoustic Phonetics*. New York: Cambridge University Press, 1988.

[115] L. Cohen, *Time-Frequency Analysis: Theory and Applications*. New York: Prentice-Hall Signal Processing Series, 1995.

[116] B. Boushash, "Estimating and interpreting the instantaneous frequency of a signal – part 1: fundamentals," *Proc. IEEE*, vol. 80, no. 4, pp. 520–538, Apr. 1992.

[117] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Singapore: Pearson education, 2004.

[118] Christophe d'Alessandro. Voqual: Voice material. [Online]. Available: http://archives.limsi.fr/VOQUAL/voicematerial.html

[119] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group-delay function," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762–765, Oct. 2007.

[120] M. Brookes. Voicebox: A speech processing toolbox for MATLAB.2006. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[121] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, 2004, pp. 223–224.

[122] CMU-ARCTIC speech synthesis databases. [Online]. Available: http://festvox.org/cmu_arctic/index.html

[123] Noisex-92. [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[124] R. J. Baken and R. F. Orlikoff, *Clinical measurement of speech and voice*, 2nd ed. San Diego: Singular Thomson Learning, 2000.

[125] P. Ladefoged and N. P. Mackkiney, "Loudness, sound pressure and subglottal pressure in speech," *J. Acoust. Soc. Amer.*, vol. 35, no. 4, pp. 454–460, Apr. 1963.

[126] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. European Conf. Speech Processing, Technology*, Greece, Sept. 1997, pp. 1895–1898.

[127] Alain de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[128] R. Goldberg and L. Riek, *A Practical Handbook of Speech Coders*. CRC Press, 2000.

[129] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 5.0.10). [Online]. Available: http://www.fon.hum.uva.nl/praat/

[130] A. de Cheveigne. YIN, a fundamental frequency estimator for speech and music. [Online]. Available: http://www.auditory.org/postings/2002/26.html

[131] F. Plante, G. F. Meyer, and W. A. Aubsworth, "A pitch extraction reference database," in *Proc. European Conf. on speech comm. (Eurospeech)*, Madrid, Spain, Sep. 1995, pp. 827–840.

[132] G. F. Meyer. Keele pitch database. [Online]. Available: http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html

[133] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F_0 contours for computer and intonation teaching," in *Proc. European Conf. on speech comm. (Eurospeech)*, Berlin, Germany, Sep. 1993, pp. 1003–1006.

[134] P. Bagshaw. Evaluating pitch determination algorithms. [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/fda/

[135] H. Kutturuff, *Room Acoustics*, 4th ed. Taylor & Francis, 2000.

[136] M. S. Brandstein and E. D B Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. Berlin: Springer-Verlag, 2001.

[137] M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor, "Multichannel dypsa for estimation of glottal closure instants in reverberant speech," in *Proc. EUSIPCO*, Poznan, Poland, Sep. 2007.

[138] M. R. P. Thomas, N. D. Gaubitch, J. Gudnason, and P. A. Naylor, "A practical multichannel dereverberation algorithm using multichannel dypsa and spatiotemporal averaging," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct. 2007, pp. 50–53.

[139] F. Flego and M. omologo, "Multi-microphone periodicity function for robust f_0 estimation in real noisy and reverberant environments," in *Proc. ICSLP*, Pittsburg, Sep. 2006, pp. 2146–2149.

[140] F. Flego, C. Zieger, and M. Omologo, "Adaptive weighting of microphone arrays for distant-talking f_0 anv voiced/unvoiced estimation," in *Proc. Interspeech - 2007*, Antwerp, Belgium, Aug. 2007, pp. 2961–2964.

[141] R. K. Swamy, K. S. R. Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 51–54, Jul. 2007.

[142] A. S. Abramson, "Laryngeal timing in consonant distinctions," *Phonetica*, vol. 34, no. 4, pp. 295–303, 1977.

[143] R. B. Monsen, "Normal and reduced phonological space: The study of vowels by a deaf adolescent," *Journal of Phonetics*, vol. 4, pp. 189–198, 1976.

[144] A. L. Francis, V. Ciocca, and J. M. C. Yu, "Accuracy and variability of acoustic measures of voicing onset," *J. Acoust. Soc. Amer.*, vol. 113, no. 2, pp. 1025–1032, 2003.

[145] R. Wayland and A. Jongman, "Acoustic correlates of breathy and clear vowels: the case of Khmer," *Journal of Phonetics*, vol. 31, no. 2, pp. 181–201, 2003.

[146] M. Gardon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[147] T. Cho and P. Ladefoged, "Variations universals in VOT: evidence from 18 languages," *Journal of Phonetics*, vol. 27, no. 2, pp. 207–229, Apr. 1999.

List of Publications

Refereed Journals

1. B. Yegnanarayana, R. K. Swamy and K. S. R. Murty, "Determining mixing parameters from multispeaker data using speech-specific information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1196-1207, Aug. 2009.

2. K. S. R. Murty, B. Yegnanarayana and Anand Joseph M. "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469-472, June 2009.

3. B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 614-624, May 2009.

4. K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

5. R. K. Swamy, K. S. R. Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Processing Letters*, vol. 14, no. 7, pp. 481-484, July 2007.

6. K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–56, Jan. 2006.

Conferences

1. K. S. R. Murty, S. Khurana, Y. U. Itankar, M. R. Kesheorey and B. Yegnanarayana, "Efficient representation of throat microphone speech," in *Proc. Interspeech - 2008,* Brisbane, Australia, Sep. 22-26, 2008, pp. 2610–2613.

2. B. Yegnanarayana, K. S. R. Murty and S. Rajendran, "Analysis of stop consonants in Indian languages using excitation source information in speech signal," in *Proc. ISCA-ITRW Workshop on Speech Analysis and Processing for Knowledge Discovery,* Aalborg university, Denmark, Jun. 4-6, 2008.

3. K. S. R. Murty, B. Yegnanarayana, "Event-based interpretation of HMM state sequences for speech analysis," in *Proc. Managing Complexity in Distributed World, MCDES-2008*, Bangalore, India, May 2008.

4. S. Guruprasad, B. Yegnanarayana, K. S. R. Murty, "Detection of instants of glottal closure using characteristics of excitation source" in *Interspeech - 2007*, Antwerp, Belgium, Aug. 2007, pp. 554–557.

5. K. S. R. Murty, B. Yegnanarayana, and S. Guruprasad, "Voice activity detection in degraded speech using excitation source information," in *Proc. Interspeech - 2007*, Antwerp, Belgium, Aug. 2007, pp. 2941–2944.

6. K. S. R. Murty, S. M. Prasanna, and B. Yegnanarayana, "Neural network models for extracting complementary speaker-specific information from residual phase," in *Proc. International Conference on Intelligent Sensing and Information Processing*, Chennai, India, Jan. 2005, pp. 421–425.

7. K. S. R. Murty, S. R. M. Prasanna, and B. Yegnanarayana, "Speaker-specific information from residual phase," in *Proc. International Conference on Signal Processing and Communications*, Bangalore, India, Dec. 2004, pp. 516–519.

8. L. Mary, K. S. R. Murty, S. R. M. Prasanna, and B. Yegnanarayana, "Features for speaker and language identification," in *Proc. ODYSSEY-The Speaker and Language Recognition Workshop*, Toledo, Spain, June 2004, pp. 323–328.

Curriculum Vitae

Name: Sri Rama Murty Kodukula

Date of Birth: 22 August, 1981

Educational Qualifications:

- [1998 – 2002] Bachelor of Technology (B.Tech.)
- [2002 – 2009] Doctor of Philosophy (Ph.D.)

Permanent Address

s/o K. Lakshmana Rao

13-13-30, Chintavari Street

Anakapalli - 531 001

Vizag District, Andhrapradesh

ph: +91 08924 226531

email:ksrmurty@gmail.com

Doctoral Committee

Chairperson:

- Prof. T. A. Gonsalves

Guides:

- Prof. B. Yegnanarayana
- Dr. C. Chandra Sekhar

Members:

- Prof. D. Janakiram
- Prof. V. Kamakoti
- Prof. C. Sujatha
- Dr. K. Sridharan