A Novel Filtering Based Approach for Epoch Extraction

by

PRAMOD BACHHAV

201311025

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY

in

INFORMATION AND COMMUNICATION TECHNOLOGY

to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

August, 2015

Declaration

I hereby declare that

(i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at DA-IICT and has not been submitted elsewhere for a degree,

(ii) due acknowledgement has been made in the text to all the reference material used.

Pramod Bachhav

Certificate

This is to certify that the thesis work entitled, "A Novel Filtering Based Approach for Epoch Extraction", has been carried out by Pramod Bachhav (201311025) for the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology under my supervision.

Dr. Hemant A. Patil
Thesis Supervisor

Dedicated to my

Parents

And

family

Acknowledgements

A few souls have been explicitly involved in helping me accomplish my research work. I would like to thank my guide ***Dr. Hemant Patil*** to the greatest extent for supervising and standing by me throughout my research journey during M.Tech. I am fortunate that he allowed me to work under him and provided his guidance. His encouragement through his favourite words '*Work hard not hardly*' and inner urge towards writing research papers helped me turn this research into a publishable work. I adore his efforts to setup a well equipped speech lab at DA-IICT which provides an ecstatic research environment. I thank my thesis subject expert ***Prof. M.V. Joshi*** and thesis examiner ***Prof. Asim Banerjee*** for their valuable suggestions during the presentations which helped me to understand the deeper research issues in my work which in turn instilled new directions in my thought process.

I would like to thank *Ms. Tanvina Patel* for her involvement in every technical discussion and her selfless help during technical writing. I thank *Mr. Nirmesh Shah* who helped me through his idea when I was stuck with my work. I am thankful to members of speech research group at DA-IICT *viz.*, Mr. Maulik Madhavi, Mr. Mohammadi Zaki, Mr. Hardik Sailor, Ms. Avni Rajpal, Ms. Purvi Agrawal & Mr. Ankit Nagpal to be part of my research and social life at DA-IICT. I am indebted to DA-IICT for providing research provoking infrastructure in the form of Resource centre (RC) which made me accessible every book and research paper I requested and labs with the latest equipments. I thank all my DAIICTian friends *viz.,* Palak, Kamal, Vijay, Saurabh, Jainisha, Arpit, Soumya, Vineet, Lexmi and many more who helped me despite of their busy schedule to participate in the subjective test conducted. I am grateful to *Dr. A.P. Prathosh* who helped me to make sure that I was proceeding in the right direction.

I must thank Department of Electronics and Information Technology (DeitY) for providing me the funds through the consortium project *viz.*, "Development of prosodically guided phonetic engine for searching speech databases in Indian languages" and "ASR Phase II" project. I thank IEEE SPS for supporting me by providing the travel grant to present our joint research work at ICASSP 2015. I am obliged to pay my gratitude towards my *parents* and maternal *uncle* who undoubtedly supported me for research without any family interruptions. And last but not the least its almighty whom I am finally thankful and indebted to, for arranging all the above helping hands for me.

Contents

Abstract

Accurate estimation of source excitation features is important in many speech analysis-synthesis applications. According to source-filter theory, separating source features from speech signal is basically a deconvolution problem [1]. For voiced speech, the vocal tract is excited with a sequence of impulse-like glottal pulses. The extent of excitation is significant around these pulses. The work presented in this thesis is aimed at estimating the instants of significant excitation of vocal tract system which occur at glottal closure instants (GCIs), also known as epochs [2]. Unlike the conventional methods, we have proposed a method which does not require the modelling of vocal tract system for epoch estimation and thus, does not use the traditional linear prediction residual (LPR).

The proposed epoch extraction method uses lowpass filter on the positively clipped and negated speech signal followed by peak detection. The method assumes the quasiperiodicity of speech signal. The lowpass filtering removes the vocal tract characteristics from the speech signal and a peak detection method is employed to detect the epoch candidates. The method has been evaluated over a phonetically balanced database and compared with the other state-of-the art methods*, viz.,* Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) and the Zero Frequency Resonator (ZFR)-based method. The proposed method gave comparable or better results on clean as well as noisy speech signals.

In addition, using the estimated epoch locations, we proposed an *event-based* approach for pitch estimation. In this work, we have also presented an approach to evaluate the performance of a pitch estimation algorithm in the absence of ground truth. The proposed pitch estimation approach has been compared with other state-of-the-art pitch extraction methods, in the framework of voice conversion.

List of Symbols and Acronyms

List of Acronyms

GCIs	Glottal Closure Instants
LP	Linear Prediction
LPR	Linear Prediction Residual
HE	Hilbert Envelope
STFT	Short-Time Fourier Transform
DP	Dynamic Programming
PSF	Phase Slope Function
ZFR	Zero Frequency Resonator
SEDREAMS	Speech Event Detection using the Residual Excitation and a Mean-Based Signal
ILPR	Integrated Linear Prediction Residual
PI	Plosion Index
DPI	Dynamic Plosion Index
HWILPR	Half-wave rectified and negated integrated linear prediction residual
HTILPR	Hilbert transform of integrated linear prediction residual
EGG	Electroglottograph
DEGG	Differenced Electroglottograph
GOIs	Glottal Opening Instants
SNR	Signal-to-Noise Ratio

IDR	Identification Rate
MR	Miss Rate
FA	False Alarm Rate
IDA	Identification Accuracy
Proposed-WC	Proposed Method Without Positive Clipping
Proposed-NC	Proposed Method Without Clipping
YAAPT	Yet Another Algorithm for Pitch Tracking
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of WeiGHTed Spectrum
PDA	Pitch Detection Algorithm
MCD	Mel Cepstral Distance
MOS	Mean Opinion Score
GLOAT	GLOttal Analysis Toolbox
MCC	Mel Cepstral Coefficients
MLSA	Mel Log Spectrum Approximation Filter
PESQ	Perceptual Evaluation of Speech Quality
VC	Voice Conversion
GMM	Gaussian Mixture Model

List of Symbols

$s[n]$	Speech Signal

ω_n	Cut-off Frequency
τ	Group Delay
E_i	Peaks in the Lowpass Filtered Positive Clipped and Negated Speech Signal
F_o	Pitch or Fundamental Frequency
T	Threshold (which is $1/12^{th}$ of the average peak-valley difference in entire filtered speech signal)
E_p	Preceding Epoch
E_f	Following Epoch
T_i	Refined Epochs After Voiced-Unvoiced Decision
P_i	Pirch at Every Epoch Locartion T_i
P_{avg}	Average Pitch
P1	Proposed Method
P2	YAAPT
P4	PDA
P4	STRAIGHT
X	Time Sequence of Source Spectral Feature Vector
Y	Time Sequence of Target Spectral Feature Vector
ω_m	Weight of the m^{th} Mixture Component
μ_m^Z	Mean Vector
Σ_m^Z	Covariance Matrix

List of Tables

List of Figures

Chapter 1

Introduction

According to the source-filter model of speech production, speech signal can be considered as the output of a time-varying and quasi-stationary vocal tract system excited with a sound source. For voiced speech sounds-like vowels, the source is the quasi-periodic puff of air whereas the noisy airflow source gives rise to unvoiced speech sounds such as the fricative consonants. Impulsive source produces sounds such as plosives [1]. The simplicity of source-filter model makes it convenient for use in many applications. It allows modelling of voiced speech as the convolution of the impulse response of the vocal tract system with the impulses located at epoch locations [3]. Thus, epoch extraction is important for voiced speech production.

The dictionary meaning of an epoch is an instant of time marked by distinctive features called *events*. In the context of speech production, epoch is defined as the 'instant' of significant excitation of the vocal tract system, within a pitch period (i.e., T_0).

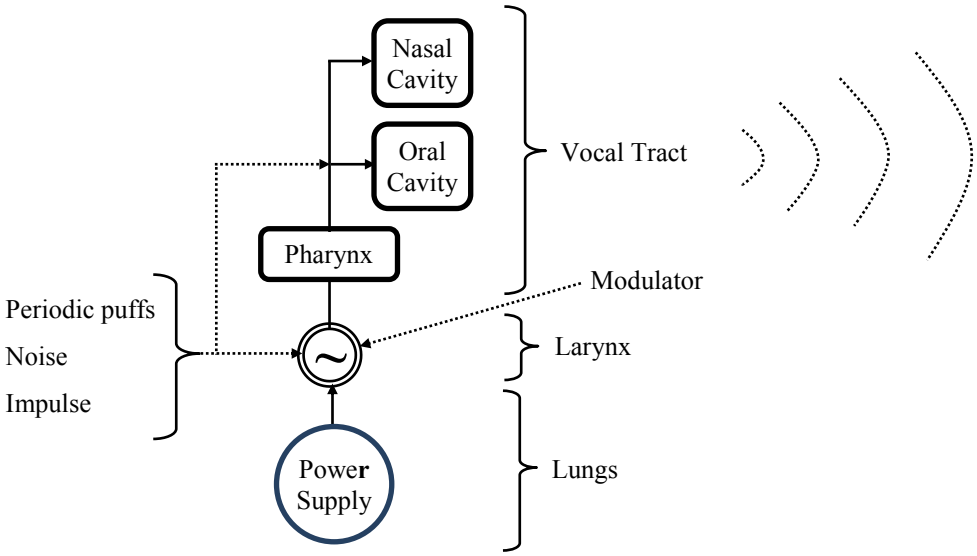

Figure 1.1 Model of human speech production (Adapted from [1]).

Figure 1.1 shows basic speech production system in humans. Lungs act as power supply which provide airflow to the larynx. Larynx acts as a modulator and provides a periodic puff-like or a noisy or an impulsive airflow source to the vocal tract. Vocal tract "colors" the modulated airflow by shaping its spectrum. Larynx is a complicated system which controls the movement of *vocal folds* [1].

Vocal folds are two masses of flesh, ligament and muscle, which stretch between the front and back of the larynx. The *glottis* is the slit-like orifice between the two folds. If airflow velocity at the glottis for periodic source is measured as a function of time, a waveform similar to that shown in Figure 1.2 is observed. This waveform is called as *glottal airflow velocity* or *glottal flow* [1]. One glottal cycle consists of *closed phase* (during which no flow occurs and folds remain closed), *open phase* (the time interval during which glottal airflow builds up to the maximum from zero) and *return phase* (airflow goes to zero from it's maximum during this phase, *i.e.,* the interval from time of maximum airflow to the time of glottal closure). The time duration corresponding to one glottal cycle is called as *pitch period* it's reciprocal is termed as *pitch*. *Pitch* may vary from *60 Hz* to *400 Hz* depending upon factors like gender, stress, speaking rate, etc [1].

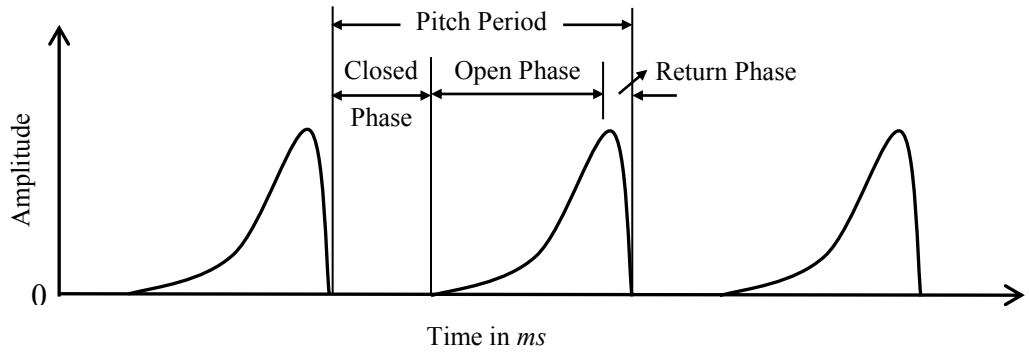

Figure 1.2 Periodic airflow velocity (Adapted from [1]).

Miller proposed the inverse filtering technique and deduced that epoch lies close to the glottal closure instants (GCIs) [2]. During the glottal vibration, due to abrupt closure of vocal folds, the significant impulse-like excitation takes place during the closed phase of the glottal cycle [3]. Therefore, epochs are supposed to lie around GCIs and both the terms are used interchangeably throughout this work.

1.1 Motivation

Estimation of the epoch locations within a glottal pulse influences many speech analysis situations [4]. Some of them are listed below.

- Glottal-synchronous processing of speech is a field of speech science in which the pseudo-periodicity of voiced speech is exploited. GCIs mark the boundaries of individual cycles of speech signal and thus, their estimation in speech signal helps in pitch tracking [5]. Therefore, epochs can be used as pitch markers for prosody

modification, which finds application in various areas like in text-to-speech (TTS) synthesis, voice conversion and speech rate conversion [2].

- During closed phase of glottal cycle, the vocal tract system gets acoustically decoupled from the trachea. Thus, the speech produced during the closed phase region is the result of free resonances of vocal tract system. Frequency response of the vocal tract system, therefore, can be accurately estimated through the analysis of the speech signal during closed phase. As GCI precede the closed phase region during a glottal cycle, it's accurate estimation is important to study the vocal tract characteristics [2] .
- Epoch locations may be useful for estimating the time-delay between speech signals collected over a pair of spatially distributed microphones [2].

Therefore, it motivates the speech researchers to develop signal processing techniques for their accurate estimation.

1.2 Objective of the Present Work

As discussed earlier, the voiced speech analysis can be viewed as a problem of *deconvolution* in which the aim is to estimate vocal tract characteristics and the excitation source [3]. This work is aimed at estimating the excitation source characteristics such as epoch locations which are useful for many speech analysis problems, as mentioned in previous sub-Section.

1.3 Contributions in the Thesis

The research work reported in this thesis contributes the following :

- A simple and novel filtering based-approach for epoch extraction has been proposed which operates on speech signal directly.
- The proposed algorithm has been extended for pitch estimation.
- A novel idea for comparison of the pitch estimation algorithms in the framework of voice conversion (VC), in the absence of ground truth has been presented. The proposed pitch estimation algorithm is evaluated in the framework of voice conversion (VC).

1.4 Organization of the Thesis

The thesis is organized as follows :

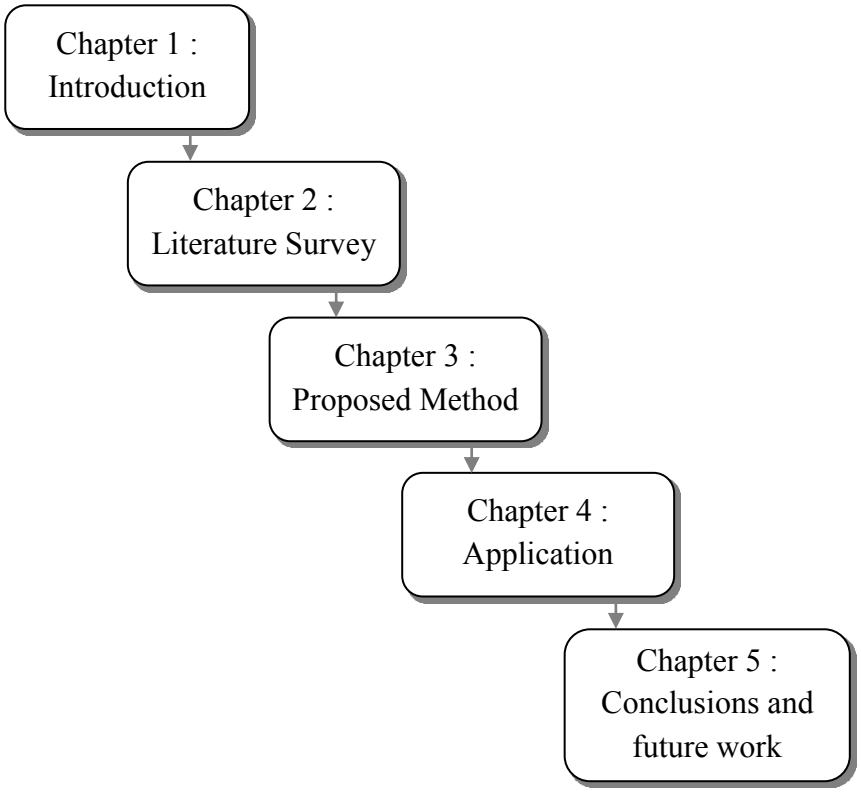

Figure 1.3 Flow of the thesis.

Chapter 2 consists of literature survey which provides a brief overview of the various existing state-of-the-art methods for epoch extraction. The need of signal processing techniques for the problem statement has also been discussed in this chapter.

In **Chapter 3,** we present the proposed simple and novel filtering-based approach for epoch extraction by exploiting the effect of positive clipping of the speech signal. The proposed method assumes the quasiperiodicity of the speech signal. The performance of the proposed approach is evaluated on clean as well as noisy speech signals and compared with other existing state-of-the-art methods.

In **Chapter 4,** we present the results of applying and modifying the proposed epoch extraction algorithm to develop an event-based algorithm for pitch extraction. The results of the developed algorithm are compared with state-of-the-art pitch extraction algorithms in the framework of VC task.

Chapter 5 concludes the contribution of the work presented along with the possible directions of future research.

1.5 Chapter Summary

In this chapter, we have briefly introduced the problem of epoch estimation, explained the motivation for taking up the problem and presented contribution in this thesis. The chapter also outlines the flow of thesis presented in subsequent chapters. Next chapter gives a brief survey of literature on epoch estimation methods.

Chapter 2

Literature Survey

2.1 Introduction

In this chapter, philosophy behind various epoch estimation methods available in literature has been discussed and their comparative study has been presented. The survey of the existing methods provided a platform for the work carried out in this thesis. In addition, recent approaches for epoch estimation, *viz.*, Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS), the Zero Frequency Resonator (ZFR)-based method, epoch extraction using Dynamic Plosive index (DPI) algorithm are discussed in detail.

2.2 Overview of the Existing Epoch Estimation Methods

The epoch extraction algorithm can be divided mainly in two steps, *viz.*,

(a) pre-processing of the speech signals to remove vocal tract characteristics and

(b) selection of appropriate candidates which give the epoch locations [6].

There are various signal processing techniques to estimate epochs are available in the literature [3], [2], [7], [8]. Some methods use linear predictability of the speech signal as the basis for epoch estimation. Because of sudden burst of energy at the GCI compared to its neighbourhood, it becomes difficult to predict the speech signal around GCI. Thus, error signal obtained in Linear Prediction (LP) analysis, called as Linear Prediction Residual (LPR) is supposed to contain information related to the epochs. Hence, the large value of error indicates the epoch location [9]. Ideally, LPR should consists of impulses near GCIs. However, there are samples of random polarity around epochs [4]. Some studies have used the Hilbert envelope (HE) of the error signal for unambiguous detection of epochs [10]- [11]. An alternative to the HE of LPR has been proposed which is based on the global phase characteristics of minimum phase signals. The method uses positive zero-crossings of phase slope function to identify epochs. The *phase slope* function (PSF) is calculated by taking average slope of the unwrapped phase of the short-time Fourier transform (STFT) of LPR as a function of time. However, this method gave rise to the

false alarms [12]. Therefore, the Dynamic Programming Phase Slope Algorithm (DYPSA) uses Dynamic Programming (DP) technique to select GCIs from a set of candidates to reduce false alarms [12]. The methods which employ *block* processing result in ambiguous epoch detections [2]. In particular, the methods which rely on LPR derived by inverse filtering need selection of parameters like order of LP analysis, length of window and are dependent on energy of error signal [2].

Table 2.1 Overview of existing epoch extraction methods.

Author	Approach	Description
T. Ananthapadmanabha and B. Yegnanarayana (1979) [11]	HE of LPR.	Unambiguous detection of epochs through detail study of LP residual.
K. Sreenivasa Rao et al. (2007) [10]	HE of LPR and Group Delay.	The epochs detected from HE of LP residual are refined by finding group delay around epochs.
R. Smits and B. Yegnanarayana (1995) [7]	Average slope of the un-wrapped phase of STFT of LPR (Phase slope function (PSF)).	Exploits global phase characteristics of minimum phase signals. Positive zero-crossings of PSF are used to find epochs.
P. Naylor et al. (2007) [12]	PSF with dynamic programming (DYPSA).	Uses PSF, phase-slope projection technique to find epochs then and employs dynamic programming to reduce false alarms.
K. Murty and B. Yegnanarayana (2008) [2]	Smoothing of speech signal using zero frequency filtering.	Uses ZFR for epoch estimation. Needs prior mean average T_0.
T. Drugman and T. Dutoit (2009) [8]	Smoothing of speech signal and LP.	Uses mean-based signal along with LPR to detect epochs. Needs prior mean average T_0.
A. Prathosh and T. Ananthapadmanabha (2013) [6]	HWILPR or HTILPR	Uses HWILPR or HTILPR with DPI to estimate epochs. Independent of threshold and prior T_0 information. Needs manual setting of first epoch for every utterance.
P. Bachhav et al. (2015) [13]	Negation of positively clipped of speech signal.	Employs simple lowpass filtering of positively clipped and negated speech signal for epoch detection. Empirical thresholding is involved.

The ZFR-based method uses the impulsive nature of excitation for epoch extraction [2]. As discontinuity in the time-domain affects *all* the frequencies, the output of ZFR

should have the information of the discontinuities in the speech signal due to impulse-like excitation. The advantage of choosing zero-frequency (*0-Hz*) is that the vocal tract characteristics does not affect the characteristics of discontinuities in the ZFR output because the vocal tract system has resonances around much higher frequencies than at the zero-frequency [2].

Recently, SEDREAMS method which finds rough locations of epochs from mean-based signal and then refines them through peaks of LPR, has been proposed [8]. However, both ZFR and SEDREAMS need mean average pitch period of a speaker apriori. Very recently, epoch extraction using Dynamic Plosion Index (DPI) from pre-processed signal has been reported in [6]. Half-wave rectified and negated integrated linear prediction residual (HWILPR) or Hilbert transform of integrated LPR (HTILPR) has been used as the pre-processed signal. This method needs to initialize first epoch manually for every utterance. In addition, proper source (i.e., either HWILPR or HTILPR) needs to be identified for every utterance.

2.3 Details of the Epoch Estimation Methods

2.3.1 Epoch Extraction using Electroglottography [4]

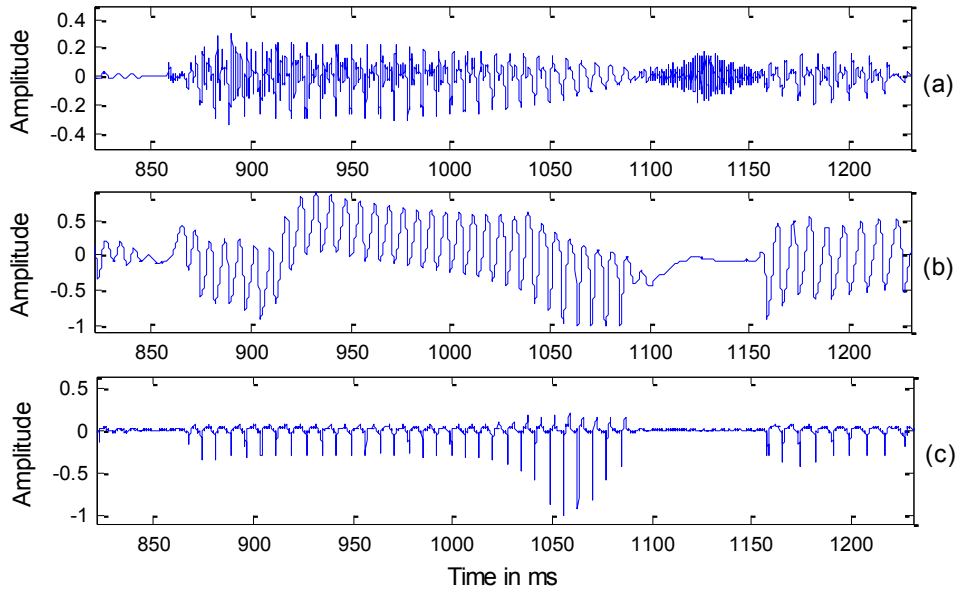

Figure 2.1 (a) A speech segment, (b) EGG, (c) corresponding DEGG.

This is a non-invasive method in which vibrations of vocal folds are recorded. A pair of electrodes is placed across the neck. As area of contact between the vocal folds changes during voicing, the corresponding variation in the impedance to small electrical current

between the electrodes is measured. The measured current is proportional to the area of contact of the folds. The corresponding signal is called as electroglottograph (EGG) which gives the measure of the glottal airflow. The derivative of EGG signal w.r.t. time is called as differenced electroglottograph (DEGG).

As glottal airflow decreases suddenly during glottal closure, the derivative of EGG consists of sharp negative peaks around GCIs. As shown in Figure 2.1, the maximum of negative peaks in the DEGG signal correspond to the GCIs [14]. The recording of EGG signals requires tedious lab setup and thus, cumbersome for the subject. Thus, it motivates the development of techniques which extract epochs directly from the speech signal.

2.3.2 Epoch Extraction using the Zero Frequency Resonator (ZFR)-Based Method [2]

Speech signal is the result of excitation of vocal tract system by one or more of the three types of excitation, namely, glottal vibration, frication and the burst. Glottal vibration is the excitation of interest for voiced speech signal. For voiced speech, there is significant amount of excitation around the instants of glottal closure during each glottal cycle and so excitation is impulsive in nature (large energy in short-time interval). Thus, the excitation can be approximated as sequence of impulses. The ZFR method uses the impulsive nature of excitation for epoch extraction. Output of the ZFR should contain the information of the discontinuities in the speech signal due to impulse-like excitation, as discontinuity in the time-domain affects all the frequencies. As the vocal tract system resonates at much higher frequencies than at the zero frequency, discontinuities in the ZFR output will not get affected by the characteristics of the vocal tract system. This forms the basis for using ZFR filter. An ideal zero-frequency digital resonator is an infinite impulse response (IIR) filter with a pair of poles located on the unit circle [15]. The following steps are involved in locating the epochs from zero frequency filtered (ZFF) signal [2].

1. Differentiate the speech signal $x(n)$ to remove low frequency bias present in the signal.

$$y_1[n] = x[n] - x[n-1]. \tag{2.1}$$

2. Differenced speech signal is passed through an ideal ZFR twice which is equivalent to successive integration four times.

$$y_2[n] = -\sum_{k=1}^{2} a_k\, y_2[n-k] + y_1[n]\,, \tag{2.2}$$

$$y_3[n] = -\sum_{k=1}^{2} a_k\, y_3[n-k] + y_2[n]\,, \tag{2.3}$$

where $a_1 = -2$, and $a_2 = 1$.

3. Trend in $y_3[n]$ is removed by subtracting the average over a window at every sample. The window length is taken equal to one or two average pitch period.

$$y[n] = y_3[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_3\,(n+m), \tag{2.4}$$

$2N+1$ corresponds to the number of samples in the window taken. The resulting signal $y[n]$ is called as the zero frequency filtered signal.

4. The positive zero-crossings in the filtered signal correspond to the epoch locations.

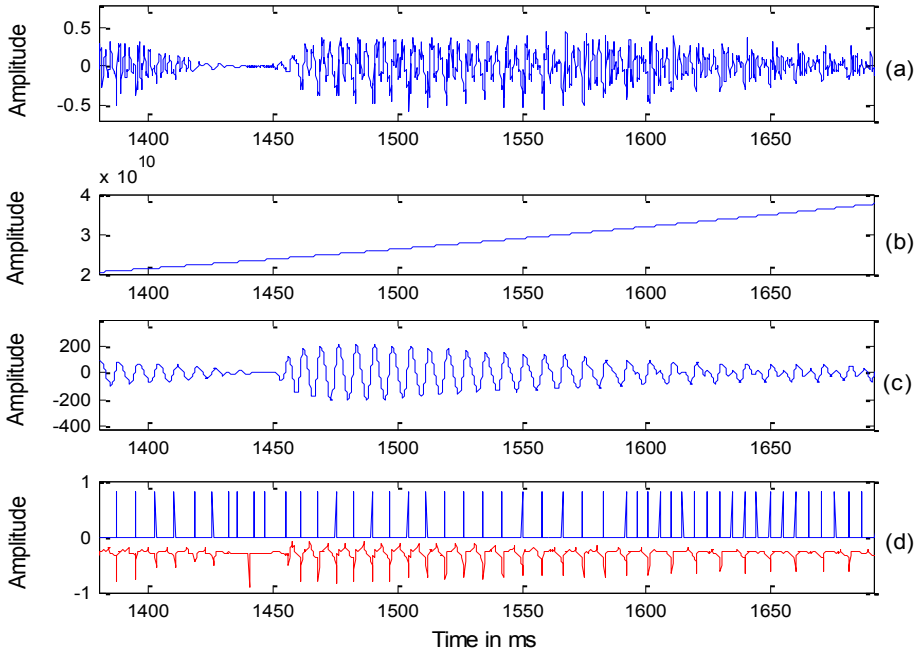

Figure 2.2 Illustration of epoch detection using ZFR. (a) A speech utterance, (b) output of cascade of ZFRs, (c) zero frequency filtered signal after trend removal, (d) top trace - detected epoch locations, bottom trace – DEGG signal as reference.

Figure 2.2 illustrates the epoch estimation using ZFR-based method. The choice of window length has to be proper (one to two pitch periods) to avoid spurious zero crossings [2]. In case of aperiodic signals like expressive voices, the sequence of epochs is affected by the size of window length chosen for trend removal. Therefore, limitations of ZFR-based method are addressed in modified zero-frequency filtering (modZFF) method [16].

2.3.3 Speech Event Detection using the Residual Excitation and a Mean-Based Signal (SEDREAMS) [8]

This method addresses the determination of glottal opening instances (GOIs) along with GCIs from the speech signal. Energy of excitation at GOIs being weaker and more dispersed than at GCIs, their automatic location estimation remains a challenging problem. The procedure is divided in two steps. Initially, the intervals where GCIs are supposed to occur are estimated from the mean-based signal. Then, the precise locations are determined using the LPR. The mean-based signal for a speech signal $s[n]$ can be computed as :

$$y[n] = \frac{1}{2N+1} \sum_{m=-N}^{N} w[m]s[n+m]\,, \tag{2.1}$$

where $w(m)$ is a window of length $2N+1$. The authors have used Blackman window with length *1.75* times the average pitch period (T_0) of the speaker.

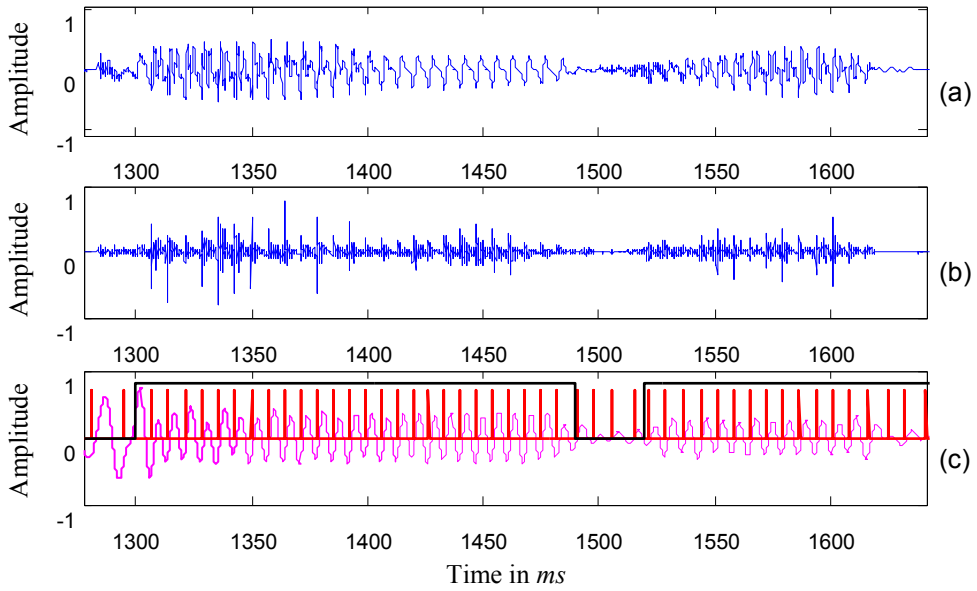

Figure 2.3 Illustration of epoch GCI detection using SEDREAMS. (a) A speech segment, (b) LPR, (c) pink trace : mean-based signal, red trace : detected GCIs, black trace : voicing decision.

It has been observed that a GCI occurs between the minimum and the following positive zero-crossing of the mean-based signal while GOI falls within an interval defined by the maximum and the following negative zero-crossing of the mean-based signal. The above observation provides the rough estimate of the short intervals where GOIs and GCIs are expected to occur. Now, there is a need to refine accurate locations of the GCIs

and GOIs within these estimated intervals. For this purpose, Linear Prediction Residual (LPR) is used. Strongest peak in the LPR corresponds to the significant excitation. Thus, accurate estimation of the GCIs and GOIs is being achieved using mean-based signal and LPR. Figure 2.3 shows the GCI estimation using SEDREAMS.

2.3.4 Epoch Extraction using Dynamic Plosion Index (DPI) [6]

Existing epoch extraction algorithms require either DP techniques or a priori information of the average pitch period. A new non-linear temporal measure named the plosion index has been used for detecting 'transients' in the speech signal. This method uses the integrated linear prediction residual (ILPR) and an extension of PI, called the DPI to make the epoch extraction independent of average pitch period, threshold cost functions and energy contour of a signal. HWILPR is used as the pre-processed signal. In this method, ILPR is obtained by inverse filtering the speech signal, with LP coefficients calculated on the pre-emphasized and Hanning windowed speech segment using the autocorrelation method. The number of predictor coefficients is equal to the sampling frequency in kHz plus four [6]. Since ILPR resembles the derivative of glottal flow, the locations of maximum negative peaks in ILPR are the representatives of the epochs. ILPR is passed through a *5*-point symmetric moving averaging (MA) filter to reduce ambiguous peaks around epoch locations. Since excitation source to the vocal tract system primarily manifests as a large negative peak in the glottal flow derivative (GFD), positive peaks in ILPR does not contain any information about GCIs. Hence, ILPR is negated and then half-wave rectified and referred to as negated half-wave rectified ILPR (HWILPR). The following temporal measures are used to identify the epochs from HWIPLR [6].

Plosion Index : It is a point measure for detecting transients which is useful in burst detection, voicing onset detection, landmark detection, etc. Intuitively, for a signal with a transient (characterized by a significant change in local energy), the ratio of the peak amplitude in the transient to the average of absolute values over an interval of interest excluding the instant of the peak, may be expected to be very high. Using this concept, PI at an instant n_0 for a speech signal $s[n]$ is defined as

$$PI(n_0, m_1, m_2) = \frac{|s(n_0)|}{s_{avg}(n_0, m_1, m_2)}, \tag{2.2}$$

where

$$s_{avg}(n_0, m_1, m_2) = \frac{\sum_{i=n_0-(m_1+m_2)}^{n_0-m_1-1} |s(i)|}{m_2}, \tag{2.3}$$

when m_1 and m_2 are the number of samples corresponding to appropriately chosen intervals preceding n_0. The values of m_1 and m_2 depend upon specific application at hand.

Dynamic Plosive Index : DPI is employed to measure inter-epochal information. DPI is nothing but PI which is computed as a function of m_2 for a given n_0 and m_1 using (2.2) where

$$s_{avg}(n_0, m_1, m_2) = \frac{\sum_{i=n_0+m_1+1}^{n_0+m_1+m_2} |s(i)|}{m_2}, \tag{2.4}$$

when m_1 and m_2 are the number of samples corresponding to appropriately chosen intervals following n_0. m_2 is varied over a range corresponding to an interval of *0* to *15 ms*, assuming that lowest pitch to be extracted is 65 *Hz* which corresponds to a pitch period of approximately *15 ms*. The dimension of DPI vector corresponds to the extent of variation of m_2.

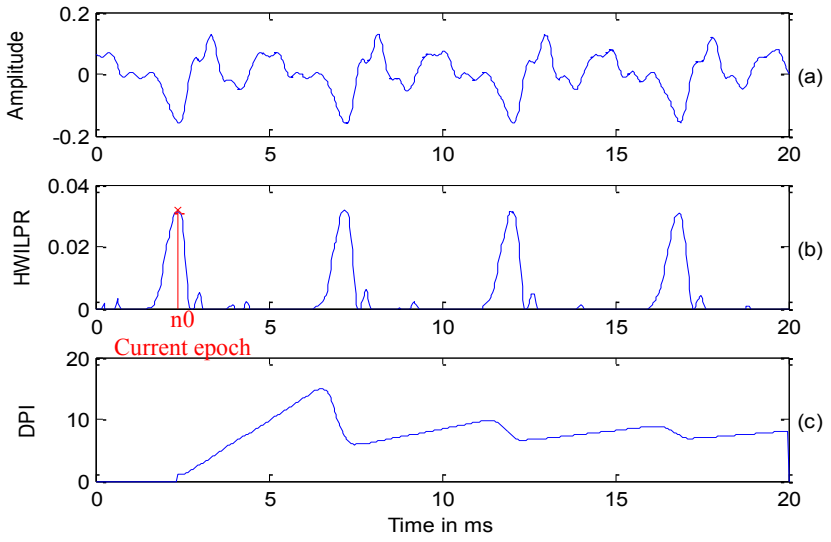

Figure 2.4 Determination of next epoch given current epoch using DPI. (a) A voiced segment, (b) HWILPR and (c) DPI with n_0 taken as reference.

Figure 2.4 shows DPI computed for HWILPR of a voiced segment of *20 ms* duration. Here, n_0 represents the current epoch which needs to be known initially. DPI increases as m_2 increases beyond n_0, reaches a peak and then decreases when m_2 begins to include the next cycle. A local minimum is observed around the peak in HWILPR which is close to the next epoch.

Epoch extraction

The problem is to find the immediate next epoch given current epoch. The initialization of first epoch may be done arbitrarily. Given the current epoch, the next epoch is detected as follows.

- DPI of HWILPR is computed with the current epoch as n_0.
- The peaks and valleys in the DPI are computed. As shown in Figure 2.4, each peak-valley pair corresponds to a cycle of HWILPR.
- The peak-valley pair with the largest difference corresponds to the immediate next cycle. The time instant corresponding to such a valley is noted and HWILPR is searched for its peak within $\pm 2\ ms$ of this instant. The instant of peak found in previous step corresponds to next epoch. This new epoch now acts as current epoch and above procedure is repeated over the entire speech signal.

Therefore, it is called as DPI algorithm.

2.4 Chapter Summary

In this chapter, we have presented the overview of the work done in area of epoch extraction along with their advantages and the limitations. In addition, three recent epoch estimation methods have been explained in detail. The chapter provides a deep insight about how the problem of GCI detection has evolved. This study helps to build a perspective to look at the problem from different directions. In the next chapter, we present the proposed epoch estimation algorithm in detail.

Chapter 3

A Novel Filtering-Based Approach for Epoch Extraction

3.1 Introduction

This chapter discusses the proposed epoch extraction algorithm which uses simple lowpass filtering of positive clipped and inverted speech signal, as the pre-processing for detection of epochs. Lowpass filtering with an appropriate cut-off frequency removes the effect of vocal tract system characteristics (as the formants lie in relatively higher frequency regions). The algorithm has been evaluated using the entire CMU-ARCTIC database [17] consisting of the EGG signals, which is a publicly available corpus. Noise robustness of the proposed algorithm is evaluated in the presence of additive white noise with various signal-to-noise ratio (SNR) levels. Experimental results show that lowpass filtering makes the proposed algorithm noise robust. The method gives comparable or better results when compared with the two state-of-the-art methods, *viz*., ZFR and SEDREAMS (which require apriori knowledge of pitch period). In addition, the proposed method shows an improvement in identification accuracy.

3.2 Details of the Proposed Method

The proposed algorithm for epoch extraction is divided into two steps. Initially, speech signal is pre-processed to remove the effect of vocal tract system characteristics and then peak detection method is employed to detect epochs. The details of the proposed algorithm are as follows :

3.2.1 Pre-processing

The voiced speech of a typical adult male has a fundamental frequency or *pitch* (i.e., F_o) in the range of *85-180 Hz* and that of a typical adult female in the range of *165-255 Hz* [18]. If a speech signal is passed through a lowpass filter (with cut-off frequency (ω_n) 250 *Hz* for female voice and *180 Hz* for male voice), the filtered signal should contain pitch information. Thus, filtered signal can be approximated as a sinusoidal signal with local period close to a pitch period. This observation forms the basis of the proposed method. Before lowpass filtering, the speech signal was positive clipped. From Figure

3.1, we observe that, positive clipping just introduces DC component and does not affect frequency distribution of the speech signal under consideration. In particular, from Fourier series representation, we verify that the DC coefficient of a sinusoid is *0* whereas that of clipped sinusoid is $1/\pi$ and fundamental frequency of original sinusoidal is retained in the clipped sinusoidal signal.

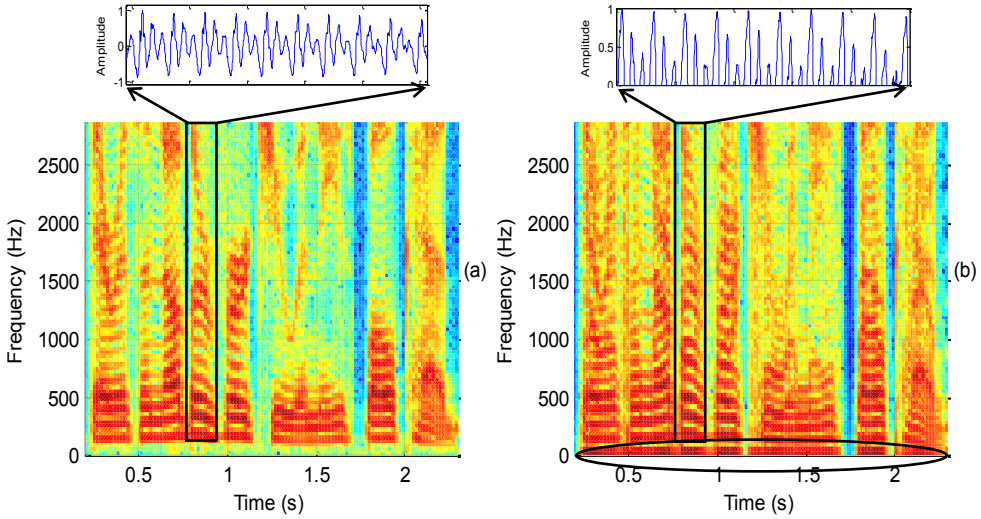

Figure 3.1 (a) Narrowband spectrogram of a speech signal, (b) narrowband spectrogram of positive clipped speech signal. Dotted circle indicates spectral energy density around *0-Hz*.

Thus, proposed algorithm pre-processes the speech signal as follows :

- Positive portion of the speech signal $x(n)$ is clipped and then it is inverted.

$$x_c(n) = \begin{cases} -x(n) & if\ x(n) < 0, \\ 0 & otherwise. \end{cases}$$

- Negated positive clipped signal $x_c(n)$ is passed through the 3^{rd} order Butterworth IIR lowpass filter with cut-off frequency, ω_n *Hz*. $y(n)$ is the lowpass filtered signal.

The choice of the Butterworth filter is based on the fact that it exhibits flat frequency response in passband whereas order is set empirically. Lowpass filtering introduces delay in the output signal. The frequency content of $y(n)$ will be roughly in the range of 0-ω_n *Hz*. Thus, it is assumed that the *group* of frequencies present in $y(n)$ suffer from constant delay equal to the group delay of filter at ω_n. In addition, Figure 3.2 depicts that in frequency range 0-ω_n *Hz*, the filter almost exhibits *linear* phase response.

- Thus, calculate the group delay $\tau(\omega_n)$ of filter at cut-off frequency ω_n and then,
- adjust the delay introduced by filtering, i.e.,

$$y(n) = y\big(n + \tau(\omega_n)\big).$$

Figure 3.3 shows a speech segment and output of lowpass filter when positive clipped and then negated speech segment passed through it. It can be observed that negative peaks in DEGG coincide with peaks of delay adjusted lowpass filtered output (solid trace). Therefore, we have experimentally observed that, after adjusting delay introduced by lowpass filter, peaks in the filtered signal correspond closely to the GCIs. After this, epochs are detected by peak-detection algorithm explained in next sub-Section.

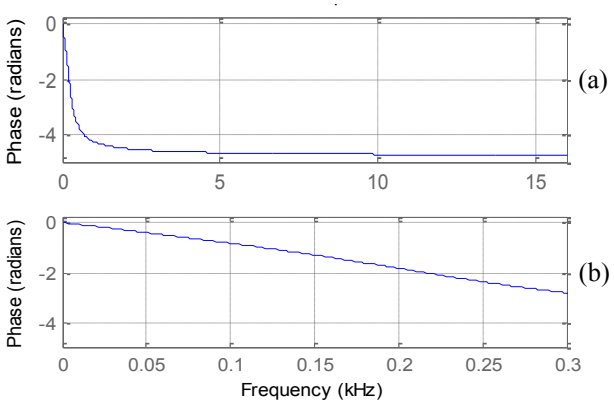

Figure 3.2 (a) Phase response of filter for range (a) *0-16 kHz* ,(b) *0-300 Hz*.

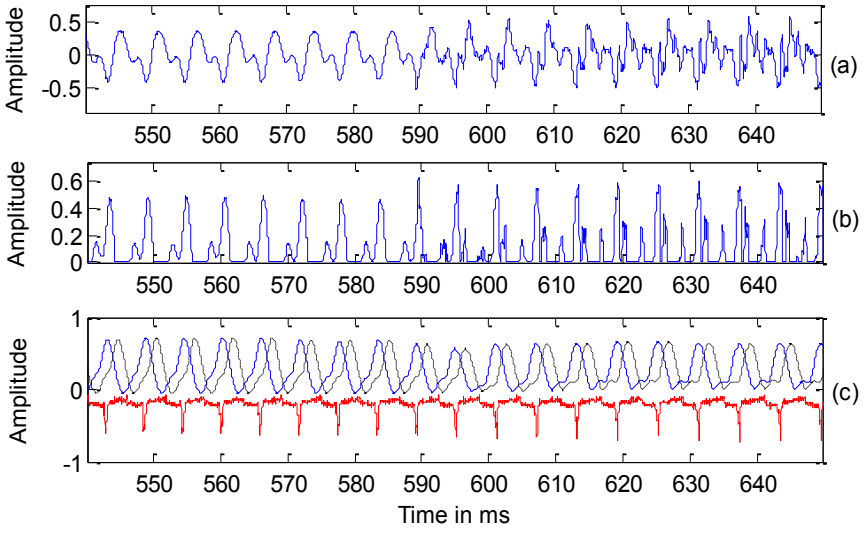

Figure 3.3 (a) Original speech segment, (b) positive clipped and negated speech segment, (c) dashed trace: lowpass filtered output, solid trace: lowpass filtered output after adjusting delay, bottom trace: DEGG signal, negative peaks of which are taken as reference epochs (i.e., the ground truth).

3.2.2 Selection of GCIs by Peak Detection Method

GCIs are located from the peaks in filtered speech as follows:

- The peaks (E_i) in lowpass filtered speech signal with significant peak-valley difference, are selected which correspond to epoch locations.

- The peaks with small peak-valley difference were removed by applying appropriate threshold *T* (threshold is selected experimentally to be *1/12*th of the average peak-valley difference in entire filtered speech signal).

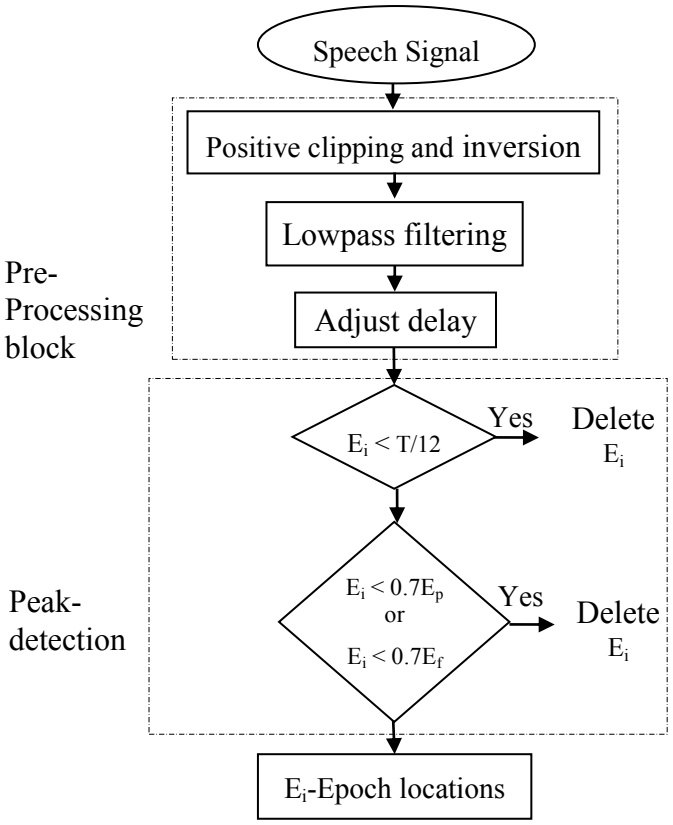

Figure 3.4 Illustration of the proposed algorithm for GCI estimation (Adapted from [13].

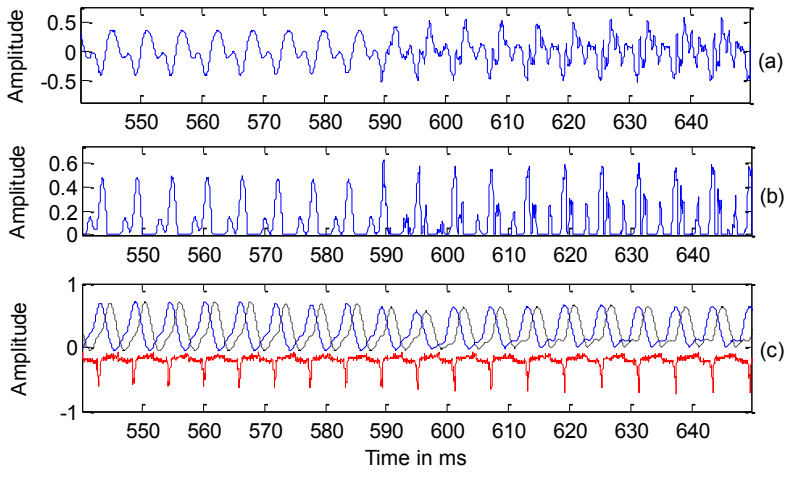

Figure 3.5 (a) Original speech segment, (b) positive clipped and negated speech segment (c) dashed trace: lowpass filtered output, solid trace: lowpass filtered output after adjusting delay, bottom trace: DEGG signal, negative peaks of which are taken as reference epochs i.e., the ground truth [13].

- Spurious small peaks seem to appear between two major peaks in some filtered speech signals. Hence, if the peak is *0.7* times less than either preceding (E_p) or following epoch (E_f), then it is considered spurious and is removed.

Figure 3.4 illustrates flowchart of the proposed algorithm for epoch estimation. Figure 3.5 (b) shows the lowpass filtered speech consisting of few peaks with very small peak-valley difference. These peaks give rise to the false alarms. Figure 3.5(c) shows the detected epochs with reduced false alarms.

3.3 Experimental Results

3.3.1 Experimental Setup

CMU-ARCTIC database was used for evaluation of proposed method [17]. The database consists of *3377* phonetically balanced utterances of *3* speakers: SLT (US female-*1132*), JMK (Canadian male-*1114*), and BDL (US male-*1131*), digitized at *32* kHz along with EGG signals [17]. After adjusting delay of *0.7 ms*, DEGG is used as the *ground truth* [2]. The maximum negative peaks in DEGG are taken as reference epoch locations. The DEGG was estimated by choosing the forward difference over the other differentiation approaches to obtain the good time resolution. The proposed method has been evaluated on voiced segments only. Voiced-unvoiced (V-UV) decision can be made by using threshold of *1/6* times peak-to-peak value of DEGG signal [19]. Here, V-UV decision is made by applying threshold of $1/9^{th}$ of maximum negative value of DEGG signal. This is the worst case for threshold to capture low energy voiced regions [6]. The performance measures for evaluation of the epoch detection algorithms are as follows [5] :

- *Identification Rate* (IDR) : the percentage of glottal cycles for which exactly *1* epoch is detected;
- *Miss Rate* (MR) : the percentage of glottal cycles for which *no* epoch is detected;
- *False Alarm Rate* (FA) : the percentage of glottal cycles for which more than *1* epoch is detected;

The glottal cycles, for which exactly *1* epoch gets detected, the timing error between the detected epoch and the reference epoch is found.

- *Identification Accuracy* (IDA) : the standard deviation of the timing error vector. Small value of IDA corresponds to higher identification accuracy.

- *Accuracy to ± 0.25 ms* : the percentage of detections for which the timing error is less than ± *0.25 ms*.

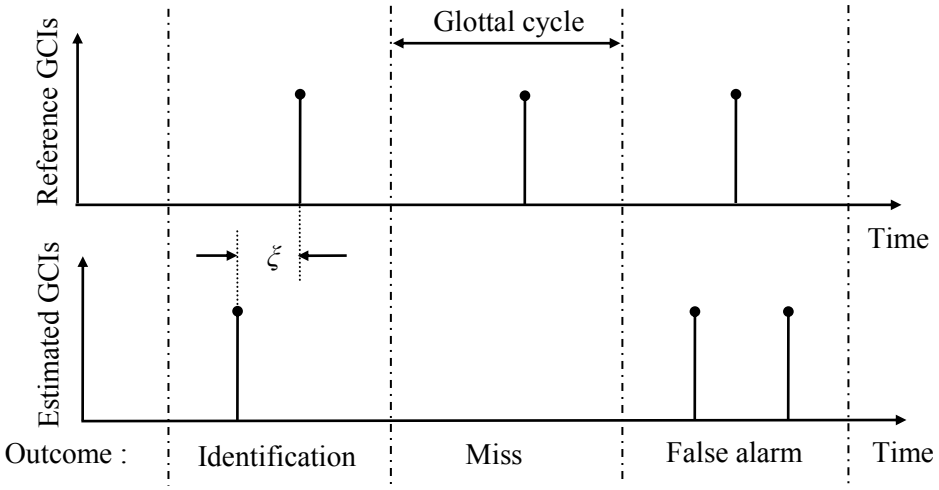

Figure 3.6 Illustration of GCI characteristics with the three glottal cycles showing examples of each possible outcome from estimation of GCI. Identification accuracy is characterized by ξ. (Adapted from [5]).

3.3.2 Performance on Clean Speech

Table 3.1 Comparison of results over CMU-ARCTIC database.

Speaker	Method	IDR (%)	MR (%)	FA (%)	IDA (ms)	Acc. to ± 0.25 ms (%)
BDL	Proposed*	98.43	0.98	**0.59**	**0.34**	63.27
	Proposed-WC	91.89	0.28	7.83	0.35	60.83
	Proposed-NC	84.44	0.76	14.8	0.75	34.96
	ZFR	96.62	**0.09**	3.29	0.38	74.76
	SEDREAMS	**98.44**	0.41	1.15	0.42	**81.81**
JMK	Proposed*	97.54	2.18	**0.28**	**0.56**	47.62
	Proposed-WC	98.14	1.34	0.52	0.59	45.06
	Proposed-NC	96.70	1.26	2.04	0.56	36.24
	ZFR	**99.04**	**0.09**	0.87	0.66	34.72
	SEDREAMS	98.97	0.61	0.42	0.65	**65.59**
SLT	Proposed*	98.93	0.44	0.64	**0.28**	69.12
	Proposed-WC	96.66	0.19	3.15	0.38	60.51
	Proposed-NC	95.77	0.67	3.56	0.66	25.56
	ZFR	99.16	**0.03**	0.81	0.35	**78.67**
	SEDREAMS	**99.45**	0.07	**0.48**	0.32	72.99

* Proposed method with positive clipping

Table 3.1 shows the performance of ZFR, SEDREAMS and the proposed method on CMU-ARCTIC database. We compare our experimental results with the ZFR and SEDREAMS as both methods perform better than HE-based, group delay-based methods

and DYPSA [2], [5]. We investigate importance of *positive* clipping by evaluating performance of the proposed method without positive clipping (Proposed-WC) and with negative clipping (Proposed-NC). IDR and FA for the proposed method are almost comparable with that of ZFR and SEDREAMS. The selection of proper threshold may help to improve MR. The proposed method gives better IDA over these two recently proposed methods. The advantage of proposed method over other two is its simplicity. It does not require prior mean pitch period information as required in existing ZFR and SEDREAMS methods.

3.3.3 Performance in Signal Degradation or Noisy Conditions

Table 3.2 Comparison of epoch extraction techniques for additive white noise on CMU-ARCTIC database at various SNR levels.

	Proposed				Proposed-WC			
SNR(dB)	-10	-5	0	5	-10	-5	0	5
IDR (%)	98.29	98.30	98.30	98.30	95.50	95.50	95.50	95.52
MR (%)	1.17	1.17	1.18.	1.19	0.59	0.59	0.60	0.60
FA (%)	**0.54**	**0.53**	**0.52**	**0.51**	3.90	3.91	3.90	3.88
IDA (%)	**0.39**	**0.39**	**0.39**	**0.39**	0.44	0.44	0.44	0.44
Acc.to ±*0.25* ms (%)	59.98	59.98	60	60	55.46	55.46	55.47	55.47

	ZFR				SEDREAMS			
SNR(dB)	-10	-5	-10	-5	-10	-5	-10	-5
IDR (%)	98.27	98.26	98.27	98.26	**98.97**	**98.93**	**98.91**	**98.94**
MR (%)	**0.07**	**0.07**	**0.07**	**0.07**	0.35	0.37	0.38	0.37
FA (%)	1.66	1.67	1.66	1.67	0.68	0.70	0.70	0.69
IDA (%)	0.48	0.48	0.48	0.48	0.50	0.49	0.48	0.47
Acc.to ±*0.25* ms (%)	62.17	62.21	62.17	62.21	66.36	69.15	71.16	72.78

The performance of the method has been tested on noisy speech signals and compared with the other two methods. For this purpose, white Gaussian noise and babble noise were added to the speech signals from CMU-ARCTIC database at different Signal-to-Noise Ratio (SNR) levels. The performance measures were averaged over all *3* speakers available in the database. The SNR is varied from *-10* dB to *5* dB in steps of *5* dB. The white noise samples were taken from NOISEX-92 database [20]. Table 3.2 and Table 3.3 display comparison of the proposed method over ZFR and SEDREAMS for white and babble noise at various SNR levels, respectively. It is observed that the proposed method performs well in terms of FA and IDA and gives comparable results in terms of IDR over other two methods under degraded or noisy conditions.

Table 3.3 Comparison of epoch extraction techniques for babble noise on CMU-ARCTIC database at various SNR levels.

	Proposed				Proposed -WC			
SNR(dB)	-10	-5	0	5	-10	-5	0	5
IDR (%)	97.28	97.92	98.17	98.25	90.36	92.13	92.86	93.08
MR (%)	1.68	1.34	1.22	1.20	2.65	1.77	1.41	1.30
FA (%)	1.04	**0.74**	**0.60**	**0.55**	6.99	6.10	5.74	5.62
IDA (%)	**0.44**	**0.41**	**0.40**	**0.39**	0.85	0.83	0.82	0.84
Acc.to ±*0.25* ms (%)	58.02	58.65	59.79	59.88	43.32	46.27	46.11	46.12

	ZFR				SEDREAMS			
SNR(dB)	-10	-5	0	5	-10	-5	0	5
IDR (%)	98.23	98.34	98.32	98.30	**98.96**	**99.00**	**99.04**	**99.06**
MR (%)	**0.09**	**0.08**	**0.07**	**0.07**	0.23	0.25	0.26	0.26
FA (%)	1.68	1.59	1.60	1.63	**0.81**	**0.74**	0.70	0.68
IDA (%)	0.47	0.46	0.48	0.48	0.52	0.50	0.48	0.47
Acc.to ±*0.25* ms (%)	56.95	60.20	61.71	62.10	65.36	68.36	71	72.50

3.3.4 Analysis of Proposed Method in Voiced Fricatives and Low Voicing Regions

The proposed method detects epochs in low voicing regions as well as voiced fricatives. In Figure 3.7, a segment of speech with a vowel followed by a voiced fricative followed by a vowel, is shown. It is observed that, epochs in the voiced fricative are detected properly. Thus, it is clear that algorithm detects epochs in all voiced sounds. Figure 3.8 depicts the performance in low voicing regions. It means that the proposed method does not depend on energy of vowel and epochs in low voicing regions can be detected.

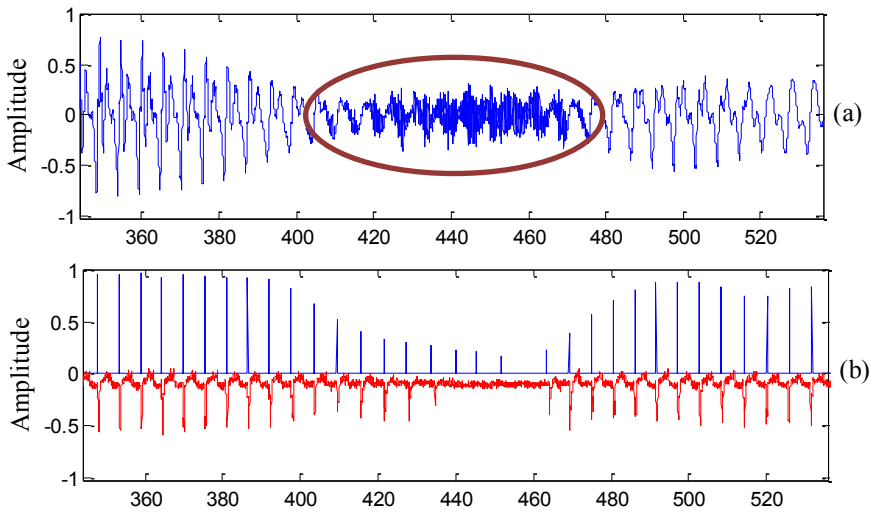

Figure 3.7 Illustration of proposed method for voiced fricative. (a) A speech segment with a voiced fricative /z/ in circular portion, (b) top trace- estimated epoch locations, bottom trace- DEGG signal.

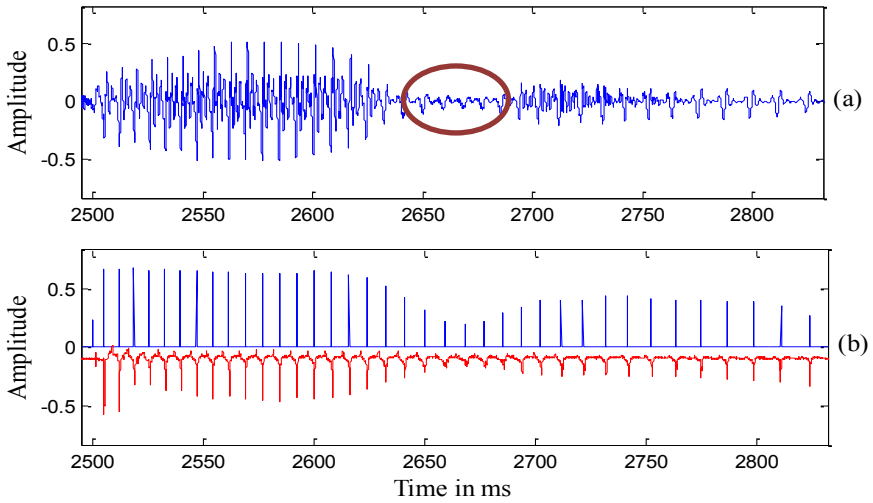

Figure 3.8 Illustration of proposed method for low voicing region. (a) A speech segment with low voicing region in circular portion, (b) top trace- estimated epoch locations, bottom trace - DEGG signal.

3.4 Chapter Summary

A simple method which uses lowpass filtering for epoch extraction has been proposed in this chapter. Performance of the method has been reported comparable or better with other existing techniques. Though the MR is high, the proposed method shows significant improvement in the FA and IDA. In addition, IDR is found to be comparable to the other methods. Lowpass filtering makes the proposed method immune to severe degraded conditions. In the next chapter, we apply the proposed algorithm for an application of pitch estimation.

Chapter 4

Application of Proposed Epoch Extraction Method

4.1 Introduction

In this work, we propose a novel pitch extraction algorithm which initially estimates the glottal closure instants (GCIs) as discussed in [13] and then computes the corresponding fundamental frequency or pitch (F_o). Hence, proposed method eliminates the assumption that F_o is constant over a segment of short duration (i.e., *20-30 ms*) and is able to track F_o even for low and short duration voiced regions. We use our recently proposed novel filtering-based approach for GCI estimation. As the proposed method directly operates on the entire speech signal, it does not need adhoc setting of window length and thus, it is free from the problem of *spectral leakage*. As accurate estimation of pitch has an impact on quality of voice, we compare the proposed method with three state-of-the-art methods, *viz.*, Yet Another Algorithm for Pitch Tracking (YAAPT) [21], Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) [22] and Pitch Detection Algorithm (PDA) [23], in the framework of voice conversion task. The *quality* and *naturalness* of converted voices has been evaluated using various objective measures (such as Mel Cepstral Distance (MCD)) and subjective measures (such as Mean Opinion Score (MOS) and ABX test). These experimental results show that proposed method performs comparable or better than the others.

4.2 Motivation for F_0 extraction

For voiced speech signal, rate of vibration of the vocal folds signifies fundamental frequency (F_0). Over the wide range of applications, F_0 and pitch are considered to be the same [24]. Pitch extraction algorithms mainly use time-domain, frequency-domain or statistical properties of speech signal. The time-domain approaches are further classified as either *block-based* or *event-based*. In block-based approach, a single value of pitch is estimated for a frame of voiced speech signal (typically for *20-30 ms* duration). Hence, these methods fail to capture variations in pitch within a speech frame. In the event-based approach, GCIs are located and then pitch is estimated from time difference between successive instances.

For voiced speech, source of excitation, namely, glottal flow is not perfectly periodic. In addition, as it is modulated by the time-varying vocal tract, speech signal is quasi-periodic in nature. Therefore, it is better to estimate the pitch value for every glottal cycle (rather than for a frame) which would give a better estimate of instantaneous pitch [25]. In this work, we propose an event-based algorithm for pitch extraction which is an application of novel filtering-based approach for glottal closure instant (GCI) estimation detailed [13]. In the absence of ground truth or reference estimates, *i.e.,* EGG, it is difficult to analyze the performance of a pitch extraction algorithm. This is because recording of EGG signals needs special apparatus and the setup is not user friendly [4]. In addition, periodicity in EGG does not always correspond to voicing and vice-versa. Thus, EGG is not perfect ground truth to get reference pitch estimates [26], [27]. Therefore, we choose an application which requires F_0 contour as an input parameter. As effective estimate of pitch contour is necessary for better transform of voice from one speaker to another, i.e., voice conversion (VC) [28], [29], we intend to compare the effectiveness of the proposed pitch extraction method in the framework of VC application.

VC refers to modifying the speaker's identity from the source speaker to the target speaker keeping the linguistic information unchanged [30]. It has many potential applications. For example, speaker-specific text-to-speech (TTS) systems, speech-to-speech translation, voice repair, voice pathology, *etc.* [31]. For effective conversion of speaker's identity, one needs to modify both system-based parameters (such as spectral parameters) and source-based parameters (i.e., excitation parameters) accurately [28]. Excitation parameters like F_0 estimated from the speech signal captures speaker's identity [32]. In particular, variations in pitch add to the speech prosody which in turn is useful parameter for VC. Therefore, an accurate estimation of pitch plays an important role in VC. In this work, various subjective and objective measures have been used for evaluating the performance of VC systems developed using YAAPT, STRAIGHT and PDA with the proposed method.

This chapter is organized as follows: sub-Section 4.3 explains the proposed algorithm for pitch estimation. The significance of F_o estimation in speech quality is discussed in sub-Section 4.4. VC method is explained in sub-Section 4.5. Experimental evaluation and results are presented in sub-Section 4.5 whereas sub-Section 4.6 concludes and summarizes the work presented in this chapter.

4.3 Proposed Algorithm for F_o Estimation

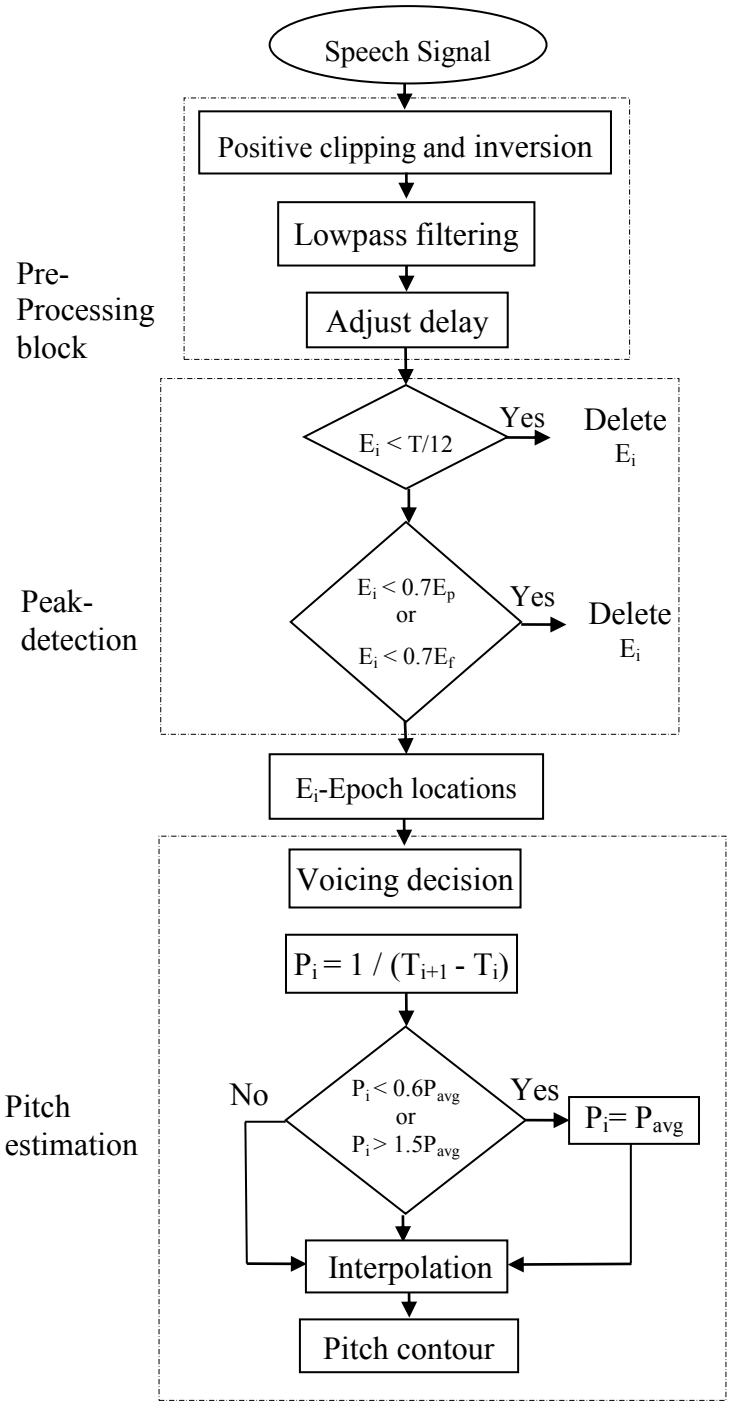

Figure 4.1 Proposed algorithm for F_o estimation (Adapted from [13]).

The proposed algorithm for F_o estimation is divided into two steps. First GCIs are estimated using the filtering-based approach discussed in [13]. Secondly, corresponding F_o contour is computed. The details of the approach are as follows :

4.3.1 Pre-processing

The speech signal is pre-processed as discussed in sub-Section 3.2.1.

4.3.2 Selection of GCIs by Peak Detection Method

The selection of GCIs after pre-processing of speech signal has been detailed in sub-Section 3.2.2.

4.3.3 Estimation of Pitch (F_0) Contour from GCIs

The steps in computation of pitch contour after estimation of GCIs are as follows :

1. The epochs in unvoiced region are removed by using voiced-unvoiced decision algorithm from GLOttal Analysis Toolbox (GLOAT) [33]. The refined epochs correspond to epoch locations in voiced regions (T_i).
2. The pitch frequency (P_i) at every epoch location is computed by taking reciprocal of time interval between successive epochs and average pitch frequency (P_{avg}) is calculated.

The miss rates and false alarms introduced by epoch detection algorithm lead to *pitch halving* and *doubling* respectively. Intuitively, these values can be approximated to P_{avg} so as to get better pitch contour.

3. The *pitch halving* because of missed epochs is removed by setting pitch values less than *0.6* times P_{avg} equal to P_{avg}.

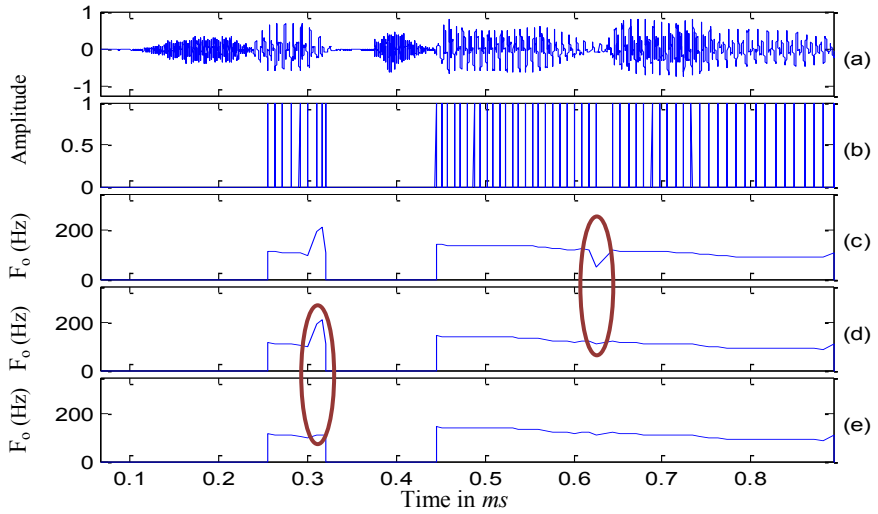

Figure 4.2 (a) A segment of speech signal, (b) detected epoch locations, (c) *Fo* contour obtained in step 2, (d) *Fo* contour obtained in step 3, (e) *Fo* contour obtained after step 4 and step 5.

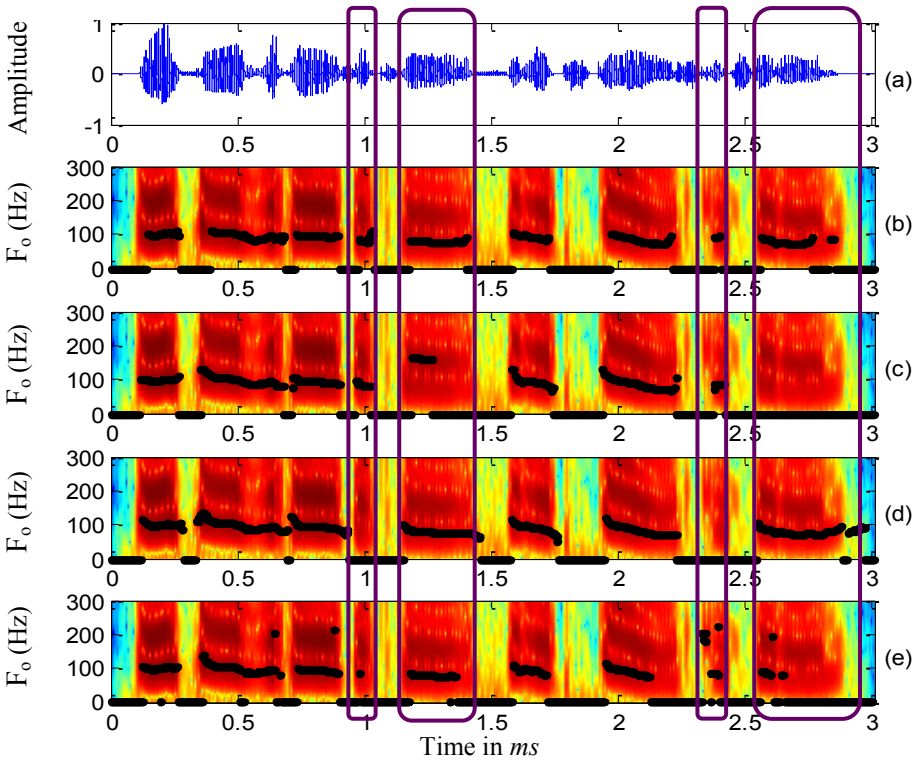

Figure 4.3 (a) A speech utterance, 'Author of the danger trail, Philip steels, etc', from a male speaker [17], pitch contour extracted using (b) proposed method, (c) YAAPT, (d) STRAIGHT, (e) PDA. Solid boxes highlight the comparison of different regions of pitch contours.

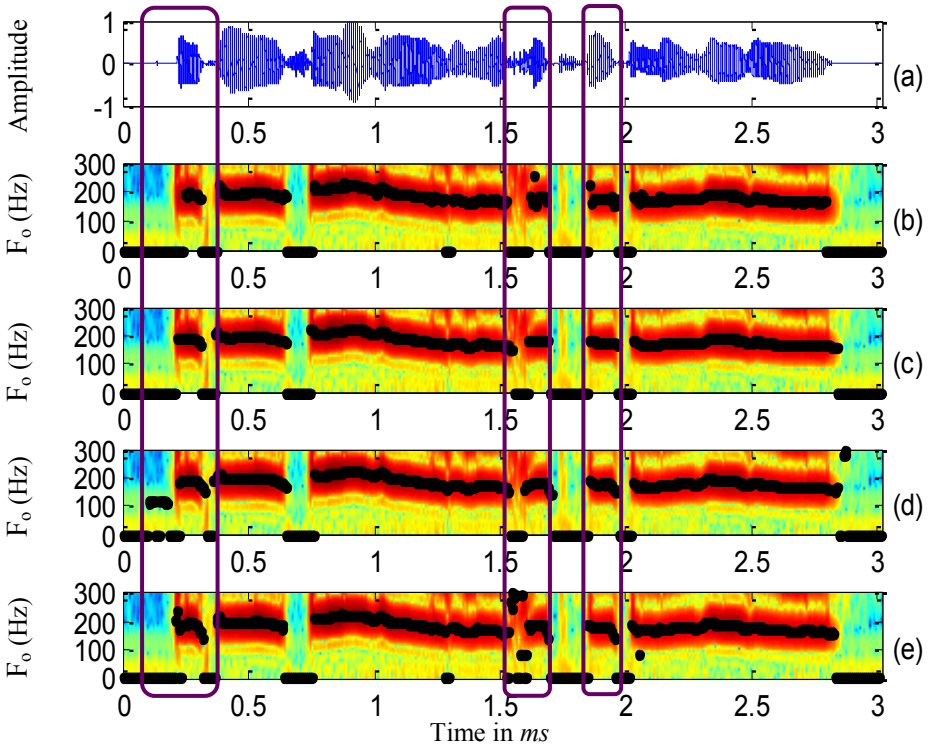

Figure 4.4 (a) A speech utterance, 'Author of the danger trail, Philip steels, etc', from a female speaker [17], pitch contour extracted using (b) proposed method, (c) YAAPT, (d) STRAIGHT, (e) PDA. Solid boxes highlight the comparison of different regions of pitch contours.

4. The false alarms in estimated epochs lead to *pitch doubling*. This can be avoided by setting pitch values which are more than *1.5* times the P_{avg} to P_{avg}.
5. The pitch (F_0) contour is obtained by linearly interpolating the values of pitch at adjacent GCIs [25]. Interpolation helps to estimate pitch value at regular intervals of time.

Figure 4.2 shows the pitch extracted for a segment of speech signal using the proposed algorithm. Figure 4.3 and Figure 4.4 illustrate the comparison of pitch contours extracted using four different algorithms for a male and female speech segment, respectively. It is evident that proposed method is able to capture pitch information more precisely, especially, for short voiced regions as compared to others as shown by solid boxes.

4.4 Effectiveness of F_0 on Speech Quality

In order to analyze the effectiveness of F_o contour on voice quality, we use an analysis-synthesis system where spectral parameters (*viz.*, Mel cepstral coefficients (MCCs)) and source features (*viz.*, F_0) are calculated during analysis phase whereas in synthesis phase, excitation signal is generated based on F_0 value. For the voiced frames, we generate an impulse train where the impulses are separated by length of pitch period whereas white Gaussian noise is generated for unvoiced frames. In this work, excitation is generated using F_o contour extracted using proposed method (*P1*), YAAPT (*P2*), PDA (*P3*) and STRAIGHT tool (*P4*). The excitation signal is then fed to the time-varying filter which represents the human vocal tract system. In this work, Mel-log spectrum approximation (MLSA) filter is used for synthesis task. *100* speech utterances of a male speaker (*viz.*, BDL) taken from CMU-ARCTIC database have been synthesized using four different pitch extraction methods (i.e., P1-P4) [17]. Objective evaluation of speech quality is done using Perceptual Evaluation of Speech Quality (PESQ) as recommended in ITU-T P.862 [34]. PESQ score is calculated between synthesized speech and corresponding natural speech which is shown in Table 4.1

Table 4.1 PESQ score for speech quality.

System	P1	P2	P3	P4
PESQ score	3.12	3.15	2.67	**3.34**

From Table 4.1, it is evident that quality of voice synthesized using proposed algorithm is comparable *w.r.t.* quality of *P2, P4* voices. In addition, the proposed method significantly outperforms *P3* in terms of PESQ score.

4.5 Voice Conversion (VC)

To convert speaker's identity, one has to convert both spectral as well as excitation features of speech signal. In this work, state-of-the-art Gaussian Mixture Model (GMM)-based VC is used.

4.5.1 Spectral Parameters

GMM-based VC technique finds *mapping function* between source and target spectral features [35], [30], [36]. Spectral features of parallel utterances from a source speaker and a target speaker is first aligned using dynamic time warping (DTW) algorithm. Let $X = [X_1, X_2, \dots, X_r, \dots, X_T]$ and $Y = [Y_1, Y_2, \dots, Y_r, \dots, Y_T]$ are the time sequence of source and target spectral feature vectors, respectively with $X_r = [x_r; \Delta x_r] \in \mathbb{R}$ and $Y_r = [y_r; \Delta y_r] \in \mathbb{R}$. Joint column vector $Z_r = [x_r; y_r]_{2d\times 1}$ is modeled by a GMM,

$$P\left(Z_t \mid \lambda^{(z)}\right) = \sum_{m=1}^{M} \omega_m N\left(Z_t \mid \mu_m^z, \Sigma_m^z\right), \tag{5.1}$$

where ω_m is weight of the m^{th} mixture component. The mean vector μ_m^Z and covariance matrix Σ_m^Z of the m^{th} mixture components are given by,

$$\mu_m^Z = \begin{bmatrix} \mu_m^X \\ \mu_m^Y \end{bmatrix}, \ \Sigma_m^Z = \begin{bmatrix} \Sigma_m^{XX} & \Sigma_m^{XY} \\ \Sigma_m^{YX} & \Sigma_m^{YY} \end{bmatrix}. \tag{5.2}$$

Here, spectral conversion is done using maximum likelihood estimation (MLE). Likelihood function is given by,

$$P\left(Y \mid X, \lambda\right) = \sum_{all\ m} p\left(m \mid X, \lambda\right) P\left(Y \mid X, m, \lambda\right), \tag{5.3}$$

where $m = \left\{m_{i_1}, m_{i_2}, ..., m_{i_T}\right\}$ is a mixture sequence. At frame t, $p\left(m_i \mid X_t, \lambda\right)$ and $p\left(Y_t \mid X_t, m_i, \lambda\right)$ are given by [5],

$$p(m_i \mid X_t, \lambda) = \frac{\omega_i N\left(X_t \mid \mu_i^x, \Sigma_i^{xx}\right)}{\sum_{j=1}^{M} \omega_j N\left(X_t \mid \mu_j^x, \Sigma_j^{xx}\right)} \ , \tag{5.4}$$

$$p(Y_t \mid X_t, m_i, \lambda) = N\left(Y_t \mid E_t(m_i), D(m_i)\right) \ , \tag{5.5}$$

where

$$E_t(m_i) = \mu_i^Y + \Sigma_i^{YX}\left(\Sigma_i^{XX}\right)^{-1}(X_t - \mu_i^X) \ , \tag{5.6}$$

$$D(m_i) = \Sigma_i^{YY} - \Sigma_i^{YX}\left(\Sigma_i^{XX}\right)^{-1}\Sigma_i^{XY} \ . \tag{5.7}$$

First, optimum mixture sequence m is determined. Then logarithm of log-likelihood function is given by [36],

$$\log p(Y \mid X, m, \lambda) = -\frac{1}{2} Y^T D_m^{-1} Y + Y^T D_m^{-1} E_m + K \ , \tag{5.8}$$

where

$$E_m = \left[E_1(m_{i_1}), E_2(m_{i2}), ..., E_T(m_{i_T})\right] \ , \tag{5.9}$$

$$D_m^{-1} = diag\left[D(m_{i_1})^{-1}, D(m_{i2})^{-1}, ..., D(m_{i_T})^{-1}\right]. \tag{5.10}$$

The constant K is independent of Y. The relationship between static spectral features y and a sequence of the static and dynamic feature vectors Y, can be obtained using transformation matrix W [37]. , that maximizes the eq. (5.8) is given by [36],

$$\left({}_m^{-1} W\right)^{-1} W^T D_m^{-1} E_m \ . \tag{5.11}$$

4.5.2 Excitation Source (F_o) Parameter

The Mean-Variance (MV) scaling approach is used for transformation of excitation parameters (*viz*., F_0) from source speaker to target speaker which is given by [36],

$$\log(f_{0conv}) = \mu_{tgt} + \frac{\sigma_{tgt}}{\sigma_{src}}(\log(f_{0src}) - \mu_{src}) \ , \tag{5.12}$$

where μ_{src} and σ_{src} is the mean and variance of F_o in logarithm for the source speaker, respectively. Similarly, μ_{tgt} and σ_{tgt} is the mean and variance of F_o in logarithm for target

speaker, respectively. MLSA filter then synthesizes target speech using converted spectral and excitation source features [38].

4.6 Experimental Results

4.6.1 Experimental Setup

In this work, *12* different VC systems have been built among two different males (*viz*., BDL (*m1*) and RMS (*m2*)) and two different females (*viz.*, CLB (*f1*) and SLT (*f2*)) from CMU-ARCTIC database [17], using FESTVOX Transform tool [39]. *100* parallel speech utterances were taken for training each system. *25-D* Mel cepstral coefficients (MCC) (including 0^{th} coefficient) per frame (with 25 *ms* frame duration and 5 *ms* frame shift) have been used. 6*4* Gaussian mixture components were taken for GMM training. *12* systems were built using each of the *4* pitch extraction algorithms. Hence, total *48* systems were built out of which *16* systems were taken for subjective evaluation and *48* systems were taken for objective evaluations.

Table 4.2 MOS scores for developed VC framework.

System	P1	P2	P3	P4
m1-m2	3.05	**3.38**	2.06	3.11
f1-m2	**2.93**	2.76	2.63	2.58
f1-f2	3.43	3.16	3.4	**3.67**
m1-f2	2.58	**3.28**	2.15	2.87
Average	3.00	**3.15**	2.56	3.06

Table 4.3 Results of ABX test.

System	ABX (% preference)					
	P1	P2	P1	P3	P1	P4
m1-m2	38.75	**61.25**	**96.25**	3.75	**50**	**50**
f1-m2	**51.25**	48.75	**55**	45	46.25	53.75
f1-f2	**52.5**	47.5	**91.3**	8.7	**50**	**50**
m1-f2	45	55	**78.7**	21.3	45	55
Average	46.88	**53.13**	**80.31**	19.69	47.81	**52.19**

4.6.2 Subjective Evaluation

4.6.2.1 MOS Test

MOS test is conducted to evaluate quality of the speech signal depending on the scores given by subjects in experiments on the listening-quality scale (*1*=Bad, *2*=Poor, *3*=Fair,

4=Good, *5*=Excellent) [40]. Five synthesized utterances from each VC system were taken for mean opinion score (MOS) from *20* listeners. Each randomly played sample from different systems was rated by subjects for speech quality. Table 4.2 illustrates that *P1* performs better than *P3* and gives comparable results with *P2* and *P4*.

4.6.2.2 ABX Test

In ABX test, given sounds A, B and X, subject has to choose one among A and B which sounds similar to X. ABX test has been conducted between 1) *P1* and *P2*, 2) *P1* and *P3* and 3) *P1* and *P4*. In this test, target natural utterance and the utterances converted using two different approaches were played. Subjects were asked to choose the utterance which was more closer to the target speech in context of speaker's identity. Table 4.3 shows the result of ABX test. From, Table 4.3 it is clear that people preferred P1 over P3 while the preference was almost same over P2 and P4.

4.6.3 Objective Evaluation

Table 4.4 MCD score for developed VC.

System	P1 (dB)	P2 (dB)	P3 (dB)	P4 (dB)
f2-f1	9.71	9.67	9.78	**9.56**
m1-f1	10.21	10.17	10.27	**10.10**
m2-f1	**10.01**	10.10	10.32	10.02
f1-f2	10.29	10.34	**9.95**	10.34
m1-f2	10.95	10.91	**10.52**	10.96
m2-f2	**10.47**	10.56	10.97	10.48
f1-m1	7.63	7.63	7.75	**7.60**
f2-m1	7.42	**7.33**	7.56	7.41
m2-m1	7.53	**7.50**	7.60	7.54
f1-m2	8.56	8.60	**8.46**	8.55
f2-m2	8.55	8.53	**8.43**	**8.43**
m1-m2	9.03	9.08	8.95	8.95
Average	9.20	9.20	9.21	**9.16**

The Mel-cepstral distance (MCD) is widely used objective measure in VC and speech synthesis. In order to measure the spectral distance between target and voice converted speech (generated using F_0 contours from four different algorithms), we calculated MCD. MCD between the converted spectrum and target spectrum is given by [41], [36],

$$sd^{mel}[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}\left(c_d - \hat{c}_a\right)}, \quad , \tag{13}$$

where c_d and $\hat{c}_a$ are the d^{th} MCC of the original and converted target, respectively. D is the order of MCC. From Table 4.4, it is evident that the proposed pitch extraction method gives comparable results with existing methods in the context of MCD.

4.7 Chapter Summary

In this work, we proposed a novel pitch (F_o) estimation algorithm which uses simple lowpass filtering and peak detection for GCI estimation. The proposed method is able to track F_o even for low and short duration voiced regions in the speech signal. In the absence of the ground truth, we compare performance of the proposed F_o estimation method with existing state-of-the-art F_o estimation methods, in the framework of VC task. We get better or comparable results in terms of voice quality measures. The proposed method is an *event-based* approach which directly works on the time-domain speech signal and therefore, free from block processing settings such as tuning of window length and artifacts like spectral smearing. Next chapter summarizes and concludes the work presented in this thesis.

Chapter 5

Summary and Conclusions

5.1 Summary of the work presented in this thesis

In this work, we have proposed a novel method which passes the positive clipped and inverted speech signal through a lowpass filter for epoch extraction. The proposed method assumes quasi-periodicity of the speech signals. As the resonances of the vocal tract system lie in the high frequency regions, lowpass filtering serves to remove characteristics of the vocal tract effect from the speech signal. Thus, the proposed method does not depend on LPR or pre-processing of LPR in which initially vocal tract system characteristics are estimated from LP analysis and then excitation is extracted using inverse filtering.

We extended the proposed epoch extraction algorithm and developed an event-based approach for pitch estimation. The proposed pitch extraction algorithm was able to track pitch even for low and short duration voiced segments. In addition, when the proposed approach was compared with the existing state-of-the-art pitch extraction methods in the framework of voice conversion, we got better or comparable results in terms of voice quality measures. This in turn signified that proposed algorithms were able to capture excitation source parameters like GCIs and pitch effectively. In addition, we presented a novel idea for comparison of the pitch estimation algorithms in the framework of VC, when the ground truth was not available.

5.2 Limitations of the Present Work

- Empirical selection of thresholds.
- Analysis with 3rd order IIR Butterworth lowpass filter only.
- High MR.

5.3 Future Research Directions

The work presented in this thesis opens variety of problems to deal as future research. Few of the directions for future work are as follows :

- To study the significance of positive clipping of speech signal in the problem of epoch extraction.
- To employ some adaptive thresholding technique so as to reduce MR. Reduction in MR ultimately will help to improve other performance measures.
- The peak-valley difference of lowpass filtered output may be extended to characterize strength of excitation (SOE) in a glottal cycle which we would like to investigate in future.
- Analysis of the performance of the proposed epoch extraction method compared to other existing methods for voiced fricatives.
- Analysis of expressive voices like laughter, shout.
- Effect on performance of the proposed method with the choice of filter, filter order, cut-off frequency.
- To investigate the performance of proposed *pitch* (F_o) extraction approach in noisy conditions and it's evaluation in terms of performance measures.

References

[1] T. F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, 2nd ed., Delhi. Prentice Hall, 2002.

[2] K. S. R. Murty and B. Yegnanarayana, "Epoch Extraction From Speech Signals," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 16, no. 2, pp. 1602-1613, November 2008.

[3] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. 23, no. 6, pp. 562-570, 1975.

[4] S. V. Gangashetty and B. Yegnanarayana, "Epoch-based analysis of speech signals," *Sadhana,* vol. 36, no. 5, pp. 651-697, 2011.

[5] T. Drugman, M. Thomas, J. Gudnason, P. Naylor and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 20, no. 3, pp. 994-1006, 2012.

[6] A. P. Prathosh, T. V. Ananthapadmanabha and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 21, no. 12, pp. 2471-2480, 2013.

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

[8] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals.," in *INTERSPEECH*, Brighton, UK, pp. 2891-2894, October 2009.

[9] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America,* vol. 50, no. 28, pp. 637-655, 1971.

[10] K. Sreenivasa Rao, S. Prasanna and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group delay function," *IEEE Signal Processing Letters,* vol. 14, no. 10, pp. 762-765, 2007.

[11] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. 27, no. 4, pp. 309-319, 1979.

[12] P. A. Naylor, A. Kounoudes, J. Gudnason and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 15, pp. 34-43, 2007.

[13] P. Bachhav, H. Patil and T. Patel, "A novel filtering based approach for epoch extraction," in *International Conference Acoustic Speech and Signal Processing (ICASSP)*, Brisbane, Australia, pp. 4784-4788, April, 2015.

[14] W. Hess and H. Indefrey, "Accurate time-domain pitch determination of speech signals by means of a laryngograph," *Speech Communication,* vol. 6, no. 1, pp. 55-68, 1987.

[15] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, 2nd ed., Englewood Cliffs: Prentice-hall, 1989.

[16] V. K. Mittal, "Analysis of nonverbal sounds," Ph.D. dissertation, IIIT Hyderabad, November 2014.

[17] "CMU-ARCTIC Speech Synthesis Databases," Available [Online] : http://festvox.org/cmu_arctic/index.html {Last Accessed: April 20, 2015}.

[18] I. R. Titze and D. W. Martin, "Principles of voice production," *The Journal of the Acoustical Society of America,* vol. 104, no. 3, pp. 1148-1148, 1998.

[19] D. G. Childers and A. Chieteuk, "Modeling the glottal volume-velocity waveform for three voice types," *The Journal of the Acoustical Society of America,* vol. 97, no. 1, pp. 505-519, 1995.

[20] “NOISEX-92,” Available [Online] : http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html {Last Accessed: September 25, 2014}.

[21] S. A. Zahorian and H. Hongbing, “A spectral/temporal method for robust fundamental frequency tracking,” *The Journal of the Acoustical Society of America,* vol. 123, no. 6, pp. 4559-4571, 2008.

[22] H. Kawahara, A. De Cheveigné and I. Masuda-Katsuse, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication,* vol. 27, no. 3, pp. 187-207, 1999.

[23] Y. Medan, E. Yair and D. Chazan, “Super resolution pitch determination of speech signals,” *IEEE Trans. on Signal Processing,* vol. 39, no. 1, pp. 40-48, 1991.

[24] A. Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America,* vol. 111, no. 4, pp. 1917-1930, 2002.

[25] B. Yegnanarayana and K. Murty, “Event-based instantaneous fundamental frequency estimation from speech signals,” *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 17, no. 4, pp. 614-624, 2009.

[26] H. Kawahara, M. Morise, R. Nisimura and T. Irino, “Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Canada, pp. 6797-6801, 2013.

[27] F. Plante, G. Meyer and W. Ainsworth, “A pitch extraction reference database,” in *EUROSPEECH'95*, Madrid, pp. 837-840, 1995.

[28] D. G. Childers, B. Yegnanarayana and K. Wu, “Voice conversion: Factors responsible for quality,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Florida, pp. 748-751, 1985.

[29] B. Yegnanarayana, J. Naik and D. Childers, “Voice simulation: Factors affecting quality and naturalness,” in *Proc. of the 10th international conference on Computational linguistics*, Association for Computational Linguistics, pp. 530-533, 1984.

[30] Y. Stylianou, O. Cappé and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. on Speech and Audio Processing,* vol. 6, no. 2, pp. 131-142, March 1998.

[31] Y. Stylianou, “Voice transformation: A survey,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, pp. 3585-3588, 2009.

[32] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *the Journal of the Acoustical Society of America,* vol. 55, no. 6, pp. 1304-1312, 1974.

[33] T. Drugman, “GLOttal Analysis Toolbox (GLOAT),” Downloaded November, 2012, Available [Online] : http://tcts.fpms.ac.be/drugman/{Last Accessed: March 25, 2015}.

[34] Perceptual Evaluation of Speech Quality (PESQ), http://www.itu.int/rec/T-REC-P.862 {Last Accessed: March 25, 2015}.

[35] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, USA, pp. 285-288, 1998.

[36] T. Toda, A. W. Black and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2222-2235, 2007.

[37] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis system,” in *Proc IEEE International Conference on Acoustics, Speech, and Signal Processing*

(ICASSP), Istanbul, Turkey, pp. 1315-1318.June 2000.

[38] S. Imai, K. Sumita and C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications),* vol. 66, no. 2, pp. 10-18, 1983.

[39] T. Toda, A. Black and K. Tokuda, "TRANSFORM : flexible voice synthesis through articulatory voice transformation," [Online]. Available: http://festvox.org/transform /transform.html {Last Accessed: March 25, 2015}.

[40] Mean Opinion Score (MOS), http://www.itu.int/rec/T-REC-P.800-199608-I/en {Last Accessed: March 25, 2015}.

[41] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Victoria, pp. 125-128, 1993.

Publications from the Thesis

[1] P. B. Bachhav, H. A. Patil and T. B. Patel, "A novel filtering based approach for epoch extraction," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4784-4788, Brisbane, Australia, 2015.

[2] P. B. Bachhav, N. J. Shah and H. A. Patil, “A novel filtering based pitch extraction algorithm for voice conversion”, article under preparation for *ICASSP, China*, 2016.