NOISE RESILIENT SPEECH SIGNAL ANALYSIS USING NON-STATIONARY SIGNAL PROCESSING TECHNIQUES

Ph.D. Thesis

by

Pooja Jain

DISCIPLINE OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

APRIL 2015

NOISE RESILIENT SPEECH SIGNAL ANALYSIS USING NON-STATIONARY SIGNAL PROCESSING TECHNIQUES

A report submitted in partial fulfillment of the requirements for the award of the degree of

Doctor of Philosophy

by

Pooja Jain

under the guidance of

Dr. Ram Bilas Pachori

DISCIPLINE OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

APRIL 2015

CHECKLIST

#	Items	Declaration	
1.	Is the thesis/report bound as specified?	Yes	No
2.	Is the Cover page in proper format as given in Annexure 1 of guidelines for thesis preparation?	Yes	No
3.	Is the Title page (Inner cover page) in proper format?	Yes	No
4.	**I.** Is the Certificate from the Supervisor in proper format? **II.** Has it been signed by the Supervisor?	Yes	No
5.	**I.** Is the Abstract included in the thesis/report properly written within 400 to 600 words? **II.** Have the technical keywords (not more than six) specified properly?	Yes	No
6.	Have you included the List of Abbreviations/Acronyms in the thesis/report?	Yes	No
7.	Does the thesis/report contain a summary of the literature survey?	Yes	No
8.	Does the Table of Contents include page numbers? **I.** Are the Pages numbered properly? (Chapter 1 should start on page number 1) **II.** Are the Figures numbered properly? (Figure Numbers and Figure Titles should be only at the bottom of the figures) **III.** Are the Tables numbered properly? (Table Numbers and Table Titles should be only at the top of the tables) **IV.** Are the Titles for the Figures and Tables proper and sources acknowledged? **V.** Are the Appendices numbered properly? Are their titles appropriate?	Yes	No
9.	Have you incorporated feedback received during various stages of evaluation?	Yes	No
10.	Is the Conclusion of the thesis/report based on discussion of the work?	Yes	No
11.	**I.** Are References or Bibliography given at the end of the thesis/report? **II.** Have the References been cited properly inside the text of the thesis/report? **III.** Is the citation of References in proper format?	Yes	No
12.	Is the thesis/report format and contents are according to the guidelines?	Yes	No

CANDIDATE'S DECLARATION

I hereby certify that I have properly checked and verified all the items as prescribed in the checklist and ensure that my thesis/report is in proper format as specified in the guideline for thesis preparation.

I also declare that the work containing in this report is my own work. I, understand that plagiarism is defined as any one or combination of the following:

1. To steal and pass off (the ideas or words of another) as one's own
2. To use (another's production) without crediting the source
3. To commit literary theft
4. To present as new and original an idea or product derived from an existing source.

I understand that plagiarism involves an intentional act by the plagiarist of using someone else's work/ideas completely/partially and claiming authorship/originality of the work/ideas. Verbatim copy as well as close resemblance to some else's work constitute plagiarism.

I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programmes, experiments, results, web-sites, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.

I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report/dissertation/thesis are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable. My faculty supervisor(s) will not be responsible for the same.

Signature:

Name: Pooja Jain

Roll No.: 1010205

Date: 30 April, 2014

THESIS CERTIFICATE

I hereby certify that the work, which is being presented in the report/thesis, entitled **Noise Resilient Speech Signal Analysis using Non-stationary Signal Processing Techniques**, in fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** and submitted to the institution is an authentic record of my/our own work carried out during the period **January-2011** to **April-2014** under the supervision of **Dr. Ram Bilas Pachori**. I also cited the reference about the text(s)/figure(s)/tables(s) from where they have been taken.

To the best of my knowledge, the matter presented in this thesis has not been submitted elsewhere for the award of any other degree or diploma from any Institutions.

Dated: Signature of the Candidate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Dated: Signature of the Thesis Supervisor

ACKNOWLEDGMENT

I am extremely grateful to my supervisor Dr. Ram Bilas Pachori for his guidance and support throughout my research work. I thank him to provide this wonderful opportunity to do research under his guidance. His enthusiasm encouraged me to work diligently and in a focused manner. He exposed me to non-stationary signal analysis techniques which developed my interest in this area. He gave me a new perspective to understand the behavior of physical signals. Frequent and lengthy discussions held with him helped me to gain valuable insights into my research problems. He set high standards to foster quality research which made me strive hard to come up with innovative and efficient solutions. He stressed on consistency and contemplation. He also helped in the development of other skills required for a professional to become versatile. His alacrity towards solving complex problems has a great influence in the way I have and will approach my professional/personal duties at present and in future.

I am thankful to Dr. Amod Umarikar for his kind support during challenging times in my research. The thoughtful questions raised by him during my research work presentations have helped me to gain new perspectives on my research problems. He encouraged me to keep on working hard to attain my research goals. I am also thankful to Dr. Narendra S. Chaudhari for sparing his invaluable time to evaluate the progress and quality of my research work.

I express my gratitude to Dr. Prakash Vyavahare for motivating me to join PhD at Indian Institute of Technology Indore. He has been a constant source of positivity and inspiration for me. I am also thankful to him for providing all the information about travel to Europe for presenting my research paper at ISPA-2013. I thank Dr. Abhishek Srivastav for answering my queries on travel to Europe. I extend my gratitude to the SERB, a statutory body of DST for approving and granting funds for the travel to Europe enabling me to present my research paper at ISPA-2013. I am grateful to Dr. G. F. Mayer for promptly providing the Keele pitch evaluation database used to assess and compare the performance of the proposed method for pitch determination in this work with other existing methods.

I had a memorable time with my colleagues cum friends Jaya Thomas, Shivnarayan

Patidar, Suneel Yadav, Neetesh Saxena, Prateek Jain and Varun Bajaj. I thank all of them for recreation and fun to relieve stress. My warm thanks to Jaya, a patient listener and a cheerful person, for all her valuable time she spared to listen and suggest solutions to distressing issues. I am thankful to Neetesh and Jaya for assisting me in performing various chores related to the international travel to present my research paper. I am grateful to Shivnarayan Patidar for helping me searching and locating crucial reference papers for my research work. I express my sincere thanks to Suneel Yadav for engaging in meaningful discussions on the mathematical theory and its presentation in my research work. I am thankful to Varun for providing me the required software, important reference papers and all the material to carry out my research work smoothly. I thank Prateek for his cheerful demeanor which helped relieve stress.

I express my sincere thanks to I.I.T. Indore library and its staff for providing access to valuable resources and subscription to crucial research websites which helped me to understand underlying concepts and state of the art methods pertaining to my research area. I express my warm thanks to my B.E. and M.Tech classmates, Ritesh Solanki, Kapil Vyas, Rhicha Masih for their moral support and encouragement. I extend my gratitude to all researchers who have designed, developed and publicized the databases, software tools that were employed for comparative studies in this research work. I am grateful to all the anonymous reviewers of my publications for their valuable insights, comments and suggestions which helped me a lot to improve the presentation of this research work.

Last but not the least, I am indebted to my parents for their unconditional love, care and support throughout my research work. I am grateful to them for their constant motivation and encouragement. I am thankful to them for their care during my illness. This work would not have been possible without their help. Finally, I would like to dedicate my thesis to my parents and Lord Shiva.

Dated: 30 April 2014 (Pooja Jain)

Dedicated to my parents and Lord Shiva

Abstract

Speech signal processing applications have gradually found place in diverse fields such as mobile phones, text reader applications, GPS, human-computer interactions, wireless communications, voice pathology detection. Real time language translation and natural language interpretation are the new emerging areas that employ speech signal processing. The scope of speech signal processing applications is expected to expand and grow in the coming years.

This thesis focuses on the noise resilient analysis of the speech signal using non-stationary signal processing techniques. Speech signal analysis in the low frequency range (LFR) is shown to be advantageous for robust determination of glottal characteristics pertaining to the voiced regions of a speech signal. It is useful for many applications such as text to speech synthesis, speaker recognition and emotion recognition. This thesis proposes noise resilient and accurate algorithms for instantaneous V/NV detection, extraction of the time-varying F_0 component of a voiced speech signal and GCI identification. A novel technique for decomposition of a multi-component non-stationary signal (such as speech signal) into AM-FM mono-component signals is proposed in the last chapter of this thesis. It is employed for formant analysis of the voiced speech signal.

The proposed V/NV detection algorithm exploits the property that in the LFR, the energy over the time-frequency plane is present only during voiced regions of the speech signal. The proposed iterative algorithm caters to the challenging problem of reliable extraction of the time-varying F_0 component of a voiced speech signal in the presence of noise, without the need of time-varying filters. The proposed GCI identification method locates GCIs reliably and accurately by employing negative cycles of the extracted time-varying F_0 component of a voiced speech signal to provide coarse estimate of the intervals

where GCIs are likely to occur. Finally, a novel iterative decomposition approach is proposed to extract either only strong or strong cum weak AM-FM mono-component signals from a multi-component non-stationary signal (such as a voiced/unvoiced speech signal). The proposed iterative decomposition approach efficiently extracts the formant components of a voiced speech signal. The proposed iterative decomposition approach when used along with discrete energy separation algorithm (DESA) performs efficient and noise resilient formant analysis.

Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

A/D	Analog to Digital
AM	Amplitude Modulated
AM-FM	Amplitude-Frequency Modulated
DEGG	Differenced Electroglottograph Signal
DESA	Discrete Energy Separation Algorithm
DFT	Discrete Fourier Transform
DYPSA	Dynamic Programming Projected Phase Slope Algorithm
EGG	Electroglottograph Signal
EEMD	Ensemble Empirical Mode Decomposition
EMD	Empirical Mode Decomposition
EOS	Equal and Opposite in Sign
EVD	Eigenvalue Decomposition
FFR	Full Frequency Range
FFT	Fast Fourier Transform
FM	Frequency Modulated
GCI	Glottal Closure Instant
GPS	Global Positioning System
HHT	Hilbert-Huang Tansform
IVR	Integrated Voice Response
LCM	Least Common Multiple
LFR	Low Frequency Range (50 Hz - 500 Hz)
LOMA	Lines of Maximal Amplitude

LP	Linear Prediction
MEDT	Marginal Energy Density With Respect to Time
MSC	Mono-component Signal Criteria
PWVD	Pseudo Wigner-Ville Distribution
SEDREAMS	Speech Event Detection Using the Residual Excitation and a Mean Based Signal
SNR	Signal to Noise Ratio
VAD	Voiced Activity Detection
V/NV	Voiced/Non-voiced
VOIP	Voice Over Internet Protocol
V-UV-S	Voiced-Unvoiced-Silence
V/UV	Voiced/Unvoiced
WVD	Wigner-Ville Distribution
YAGA	Yet Another GCI Algorithm
ZFR	Zero Frequency Resonator

Chapter 1

Introduction

Speech is essential to establish verbal communication between human beings for conveying information. It is one of the most primitive modes of communication used by humans to facilitate socialization and interaction. Humans feel comfortable to express themselves and exchange information using speech even when at distance from each other. This lead to the invention of applications like fixed and mobile telephony, voice over Internet protocol (VOIP), requiring speech signal processing. The demand for human-computer interactions using speech also motivated the development of a variety of speech signal processing applications like interactive voice response (IVR) systems, voice-activated GPS navigation systems, text readers, voice dialing enabled mobile phones, voice security systems etc. The need for non-invasive detection of voice disorders and objective measurement of vocal conditions in the medical context paved the way of development of innovative solutions using speech signal processing techniques.

Speech signal processing refers to the study of characteristics of speech signals and development of efficient algorithms for their processing. Here, we assume the speech signal to have a digital representation. The diverse speech signal processing applications can be categorized into speech analysis/synthesis, speech coding, speech recognition, speaker recognition, speaker diarization, speech enhancement, voice analysis. This thesis proposes noise resilient algorithms for acoustic analysis of the speech signal using both existing and innovative non-stationary signal processing techniques. The thesis demonstrates the significance of performing V/NV detection and GCI identification by analyzing the speech signal in the low frequency range (LFR: 50 Hz - 500 Hz) to achieve high accuracy and noise

robustness. An innovative iterative approach for decomposition of a multi-component non-stationary signal (e.g. speech signal) into amplitude-frequency modulated (AM-FM) mono-component signals is presented in the last chapter of this thesis. The proposed iterative decomposition approach is suitably employed for formant analysis of the voiced speech signal. Before stating the objectives of this thesis and explaining the motivation behind them, brief description of theory and discrete-time system modeling of human speech production are provided in the next section for better understanding of the underlying concepts.

1.1 Speech Signal Production and Modeling

Speech waveform is a sound pressure wave. The air flow forced from the lungs passes through the trachea and gets modulated in the larynx. The modulated air flow from the larynx is spectrally shaped by the movement of organs in the vocal tract system comprising of pharyngeal cavity (throat), oral cavity (mouth) and nasal cavity (nose). The spectrally shaped air flow is finally radiated at lips converting the velocity waveform to pressure waveform. The block diagram of human speech production is depicted in Fig. 1.1 [1].

The modulated air flow coming out of the larynx acts as an excitation to the vocal tract system. The two elemental types of source excitation are voiced and unvoiced [1]. Voiced speech is produced when the vocal folds situated in the larynx vibrate in a quasi-periodic fashion, chopping the air flow passing through them, resulting in the source excitation to take the form of quasi-periodic puffs of air. The space located between the vocal folds is known as glottis and the vibration of vocal folds causes repeated opening and closing of the glottis. Within a glottal cycle, the excitation to the vocal tract system is maximum at the instant of closure of glottis (GCI). The variation of glottal flow volume velocity and its first derivative with respect to time represented by $u(t)$ and $u'(t)$ respectively, are shown in Fig. 1.2 for the duration of a glottal cycle [2]. The glottal pulses in successive glottal cycles are quasi-periodic in nature. The time duration between successive vocal fold openings is called as the fundamental period of voiced speech. The inverse of the fundamental

period provides the rate of vibration of vocal folds known as the fundamental frequency (F_0). F_0 is a time-varying quantity and lies in the range of 50 Hz - 500 Hz [1]. Pitch is a subjective psychoacoustical attribute of the voiced speech which is closely related to F_0 of the voiced speech [1]. In this thesis, we have used the words pitch and fundamental frequency interchangeably. Unvoiced speech is produced when the air flow passes through a narrow constriction in the vocal tract system, rendering the source excitation to be noise-like in nature, resulting in a random output. Silence durations occur in the speech in the absence of any excitation to the vocal tract system [1].

The speech pressure waveform is converted to an electrical signal and vice versa using microphone and loudspeaker respectively. The analog speech signal is converted into digital signal by a series of processes: low pass filtering, sampling, quantization and A/D conversion. The voiced speech signal is a quasi-periodic waveform characterized by high amplitude, high autocorrelation between samples of successive glottal cycles, low zero-crossing rate and periodic structure in its magnitude spectrum. The unvoiced speech signal is a random waveform characterized by low amplitude, low autocorrelation between samples, high zero-crossing rate and aperiodic spectrum [1,5]. Silence durations of a clean speech signal have no signal strength and contain the background noise for a noisy speech signal.

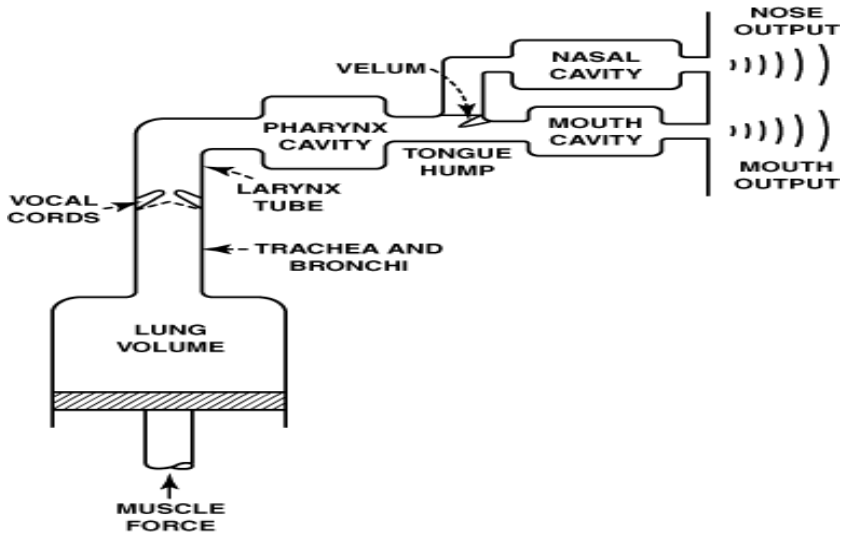

Figure 1.1: Block diagram of human speech production [1].

The understanding of system behavior of the human speech production requires sepa-

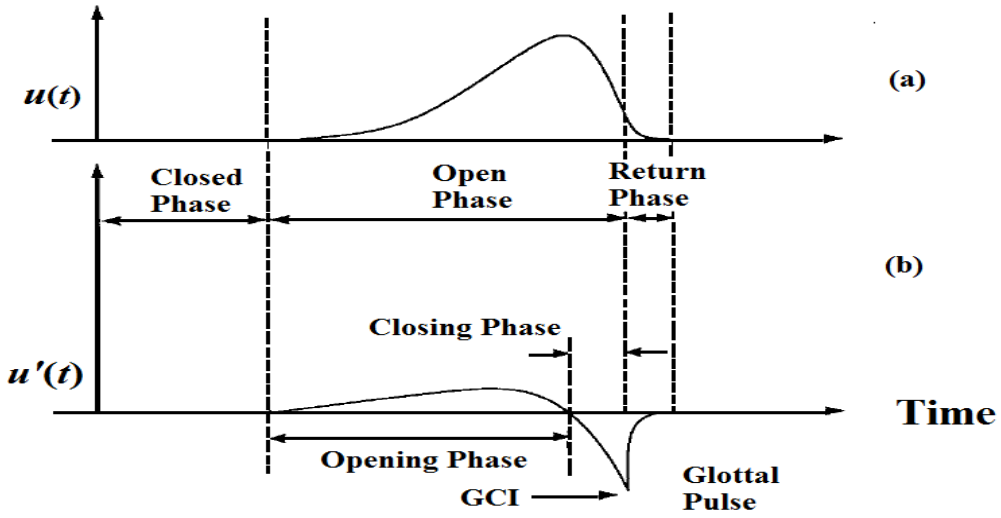

Figure 1.2: (a) Glottal flow volume velocity (b) First order derivative of glottal flow volume velocity. The waveforms are quasi-periodic in nature and are depicted for a single glottal cycle [2].

rate modeling of source excitation, vocal tract system and lip radiation. A linear discrete-time model of the human speech production is shown in Fig. 1.3 [1]. It can be inferred from Fig. 1.2 that the source excitation of the voiced speech signal can be modeled as a glottal shaping filter driven by an impulse train generator [1], as depicted in Fig. 1.3. The time duration between successive impulses in the discrete-time impulse train varies in accordance to the time-varying fundamental period of the voiced speech signal. The source excitation of the unvoiced speech signal can be modeled as a random noise generator [1]. The vocal tract system acts as a linear time-varying filter amplifying certain frequencies present in the source excitation while attenuating others [1]. The frequencies at which local peaks occur in the magnitude spectrum of the speech signal are called as formants. Thus, formants represent frequencies contained in the source excitation that are emphasized by the vocal tract system. Lets assume the speech production system to be stationary for a small time duration. The z-transform of a discrete-time speech segment, $s[n]$ spanning a small duration represented by S[z] can be expressed as [1]:

$$S[z] = \Theta_0 U[z] H[z] R[z] \tag{1.1}$$

where the z-domain transfer functions of source excitation, vocal tract system and lip

radiation are denoted by $U[z]$, $H[z]$, $R[z]$ respectively. The gain constant is represented by Θ_0 which controls the overall amplitude of the system. $H[z]$ is usually modeled as an all-pole filter [1]. The all-pole model of $H[z]$ is capable of producing a waveform that preserves the magnitude spectrum of the speech signal, sufficient for speech coding, recognition and synthesis [1]. The advantage of using an all-pole model for $H[z]$ is that it submits to the use of a powerful and simple analytic technique, linear prediction (LP) analysis. For unvoiced excitation, $U[z]$ represents the z-transform of random noise. For voiced excitation, $U[z] = E[z]G[z]$ where $E[z]$ and $G[z]$ represent the z-transform of the impulse train and z-domain transfer function of the glottal shaping filter respectively. It has been found that the lip radiation act as a differentiator and its effect is usually included in the source excitation [1]. Thus, equation (1) can be written as:

$$S[z] = \Theta_0 V[z] H[z] \tag{1.2}$$

where $V[z]$ represents the z-transform of the first order derivative of the discrete-time glottal air flow denoted by $v[n] = u'[n]$. It can be deduced from (1.2) that $u'[n]$ act as an excitation to the vocal tract system. It can be inferred from Fig. 1.2 (b) that the first order derivative of glottal air flow acting as an excitation to the vocal tract system can be closely approximated by a train of negative impulses for voiced speech [1]. The impulse-like behavior of voiced excitation is attributed to the sudden cessation of the glottal air flow during the closing phase of a glottal cycle [6]. The instants of local minima of $u'(t)$ (Fig. 1.2 (b)) are indicative of GCIs. The duration between successive GCIs varies with time and is reflected in the time-varying F_0 of voiced speech. Identification of GCIs plays a significant role in many speech processing applications. The accurate and noise resilient identification of GCIs act as a motivation for the development of algorithms presented in this thesis as explained in the next section.

1.2 Importance of Glottal Characteristics

Glottal Closure Instant (GCI) is one of the most important glottal characteristic pertaining to production of the voiced speech. During a glottal cycle, the glottal impedance drops

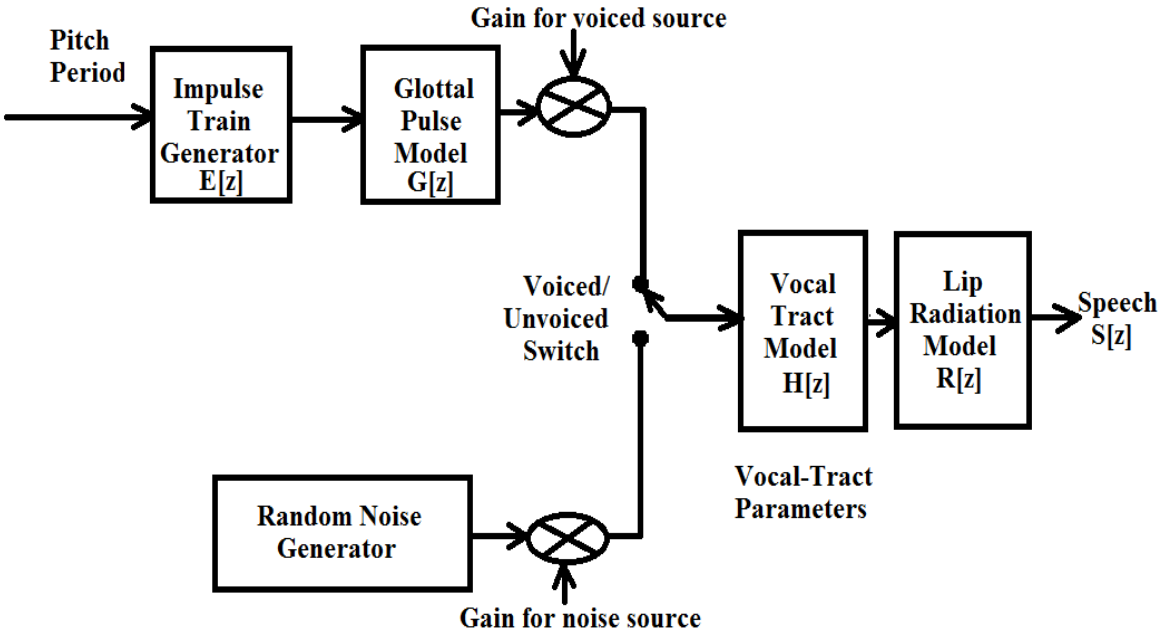

Figure 1.3: Discrete-time modeling of human speech production [1].

suddenly at the GCI, resulting in a high voiced speech signal strength. Therefore, GCIs are generally robust to noise. Identification of GCIs facilitate parametric coding of the speech signal by enabling modeling of voiced speech signal in each glottal cycle [7]. The knowledge of GCIs is helpful in identification of the closed phase of glottal cycles required by inverse filtering techniques used for estimation of the glottal source excitation [2]. The estimation of glottal source excitation finds use in determining speaker-specific features employed in applications such as speaker identification and speaker verification [8, 9].

Instantaneous F_0 can be estimated by taking inverse of the duration between successive GCIs [10]. Estimation of the instantaneous F_0 finds use in the diagnosis of pathological voice disorders, speech compression and speech enhancement [11–13]. The variations in F_0 also encode prosodic features such as intonation and stress. Stress refers to the variations in F_0 to relatively give more emphasis to certain syllables in a word or certain words in a phrase or sentence. Intonation implies variation of the pitch contour with respect to time that signifies whether an utterance is a statement or question, emotional state of a speaker, presence of sarcasm/humor, taunt in an utterance or any other information which cannot be included using grammatical rules of a language [1]. Prosody manipulation

performed in applications such as text-to-speech synthesis, voice conversion, expressive speech synthesis [14–16] can be carried out in a pitch-synchronous manner by employing the identified GCIs as markers of the corresponding pitch periods. Prosodic features have been modeled to perform speaker recognition [17] in the telephone data. It was concluded in [17] that features derived from the F_0 contour were the most useful among all features employed for automatic speaker recognition in [17]. Classification of emotions has been accomplished using the feature set comprising of cepstral analysis of the F_0 contour, instantaneous F_0 and strength of excitation around GCIs [18–20]. Thus, accurate identification of GCIs is crucial to a variety of speech signal processing applications. Many of the above mentioned applications entail the algorithms of GCI identification and instantaneous F_0 estimation to be noise resilient.

Many algorithms for instantaneous F_0 estimation and GCI identification reported in the literature assume prior information of the boundaries of voiced regions in the speech signal [21–24]. The algorithm in [25] detect voiced regions after locating GCIs but the main drawback of this method is that it requires a fixed level of noise (signal to noise ratio (SNR): 10 dB) to be added to the speech signal already degraded with noise. Some methods for F_0 estimation also perform voicing decision but require many thresholds to be set on various parameters [26]. Moreover, the performance of such methods degrade in the presence of noise. Thus, it is preferable and logical to reliably locate voiced regions prior to identification of GCIs or estimation of the instantaneous F_0, especially in the presence of noise.

1.3 Motivation

Many methods have been reported in the literature to perform voiced-unvoiced-silence (V-UV-S) classification, voice activity detection (VAD) and voiced/unvoiced (V/UV) classification [5, 27–33]. V-UV-S classification implies categorization of the speech signal into three separate classes based on the type of excitation namely: voiced, unvoiced and silence (absence of any excitation). VAD involves detection of regions of speech activity (voiced or unvoiced) in the speech signal. Thus, VAD detects intervals in an utterance with pres-

ence of human speech. V-UV classification can be performed after VAD to classify speech activity intervals into two categories depending on the type of excitation namely: voiced and unvoiced. Few methods have also been reported to accomplish voiced/non-voiced (V/NV) detection [25, 34]. V/NV detection refers to detection of durations in the speech signal for which the source excitation was of 'voiced' type i.e. detection of durations in the speech signal during which the vocal folds were vibrating. Many methods for GCI identification and the time-varying F_0 estimation require prior information of the boundaries of voiced regions of speech signal which can be provided by a noise resilient V/NV detection method at a less computational expense than V-UV-S classification methods.

Various time domain features such as: zero-crossing rate, short-term energy estimates, features extracted from the LP analysis, frequency domain features exploiting the periodic structure of the magnitude spectrum of the voiced speech signal and energy of the zero frequency resonator (ZFR) filtered signal have been employed to detect voiced regions of the speech signal [5,27,33,34]. The LP analysis assumes the speech signal to be stationary for about $20-25$ ms which is not true for quickly varying phonemes such as plosives [35]. The drawbacks of these methods are prerequisite of training data [27, 28], requirement of prior information of the average pitch period [25, 34] and noise sensitivity. One of the main limitation of the above mentioned methods lie in their inability to provide the V/NV detection at each sample instant of the speech signal. *The above mentioned limitations of existing methods and considerable scope of further improvement in performance act as strong motivations to develop an efficient algorithm for instantaneous and noise robust detection of voiced regions in the speech signal.*

The state of the art for GCI identification includes numerous methods. There are methods based on autocovariance matrix, LP residual, Frobenius norm, harmonic superposition, lines of maximum amplitudes (LOMA) of the wavelet transform, dynamic programming projected phase slope algorithm (DYPSA), AM-FM signal model, yet another GCI algorithm (YAGA) [23, 36–42]. These methods suffer from deterioration in performance in the presence of noise. The speech event detection method using the residual excitation and a mean based signal (SEDREAMS) [24] and the method based on the ZFR filtered speech signal [6] were demonstrated to be robust to different noise environ-

ments in [43] but leave ample scope for further enhancement in the GCI identification rate and detection accuracy. It was stated in [39] that negative cycles of the time-varying F_0 component of voiced speech signal provide reliable coarse estimate of intervals where GCIs are likely to occur. However, the harmonic superposition method presented in [39] have not considered the extraction of the time-varying F_0 component and its harmonics at moderate to low SNRs in noisy environments. The extraction of the time-varying F_0 component of voiced speech signal is a challenging task because of high F_0 variations possible during the course of a voiced region, substantially less energy of the time-varying F_0 component in comparison to formant components and distortion caused by noise. *The aim to develop an accurate and noise resilient GCI identification method based on devising a novel technique to reliably extract the time-varying F_0 component of the voiced speech signal in the presence of noise is the key motivation for this thesis.*

As stated in the previous section, many speech signal processing applications require estimation of the time-varying F_0 at each glottal cycle. There are two types of F_0 estimation methods which can capture variations in the time-varying F_0 at the desired time resolution, namely: instantaneous methods and event-based methods. The instantaneous F_0 estimation methods estimate the F_0 value at each sample instant of the voiced speech signal. Instantaneous F_0 estimation methods based on the time-frequency analysis techniques [22, 26, 44] and modeling of the time-varying F_0 using a B-spline expansion [21] work well for clean speech signals but their performances degrade in the presence of noise. On the other hand, event-based methods mark the occurrence of a characteristic event in each glottal cycle such as the GCI and F_0 is computed as the inverse of the time interval between successive GCIs. Thus, the event-based methods estimate the F_0 value at the time-resolution of a glottal cycle. We have already deduced from (1.2) that the voiced excitation resembles a train of negative impulses occurring at discrete-time instants (instants of glottal closure); therefore, variations in F_0 within a glottal cycle are sufficient to describe the continuous change in F_0 [44]. Event-based methods based on the GCI determination signal, wavelet transform [45, 46] are strongly dependent on the shape of the speech signal waveform which can get excessively distorted by the noise environment at moderate to low SNRs. The event-based method presented in [10] derived the final

pitch contour from the positive zero crossings of the ZFR filtered voiced speech signal and the positive zero crossings of the ZFR filtered Hilbert envelope of the voiced speech signal. The performance of [10] was shown to be robust to different noise environments. However, there is ample scope to further reduce the F_0 estimation error rates in clean and noisy environments. *The motivation here is to achieve better performance than the existing event-based methods for F_0 estimation in clean and noisy environments by employing an accurate and robust GCI identification method.* With regard to above discussions, the objectives of this thesis are stated in the next sub-section.

1.4 Objectives

Having understood the importance of precisely locating GCIs in a voiced speech signal in the area of speech signal analysis and processing, the need to locate voiced regions in a speech signal prior to identification of GCIs and limitations of existing methods, the following objectives have been identified:

1. To explore novel features which can efficiently discriminate between voiced and non-voiced regions of the speech signal even in the presence of different noise environments. To design a robust algorithm for instantaneous detection of voiced regions in the speech signal.

2. To design a technique for accurate and noise resilient extraction of the time-varying fundamental frequency (F_0) component from the detected voiced region in the presence of additive noise.

3. To design a method to precisely and reliably identify GCIs in the voiced speech signal by employing the extracted time-varying F_0 component and to estimate the instantaneous F_0 from identified GCIs.

1.5 Contributions

In the process of attaining the objectives formulated in section 1.4, the salient contributions made by this thesis are as follows:

1. A generalized amplitude-frequency modulated (AM-FM) signal model of the speech signal in the LFR is derived. The derived AM-FM signal model signifies that during voiced regions, the energy is present only at around the time-varying F_0 and its harmonics present in the LFR while negligible energy is present in the LFR during non-voiced regions of the speech signal.

2. A novel feature, marginal energy density with respect to time (MEDT) over the LFR is proposed to instantaneously detect voiced regions in the speech signal. The proposed feature has significant values only during voiced regions of the speech signal and negligible values for non-voiced regions of the speech signal.

3. The significance of the LFR in performing the speech analysis is demonstrated. The MEDT over the LFR provided substantially better discrimination between voiced and non-voiced regions than the MEDT over the full frequency range. It is shown that the speech signal analysis in the LFR provides robustness against noise.

4. The method based on the Kullback-Leibler divergence (KLD) and cumulative histogram of the MEDT over the LFR is devised to automatically determine the value of threshold on the MEDT over the LFR for reliable detection of voiced regions in the presence of additive noise.

5. The conditions on the square Hankel matrix size to accurately extract the constant amplitude/frequency harmonically related mono-component signals of a multi-component signals using eigenvalue decomposition (EVD) of the Hankel matrix are derived.

6. An iterative algorithm is developed to robustly extract the time-varying F_0 component of noise deteriorated voiced speech signal by repeatedly performing EVD of the Hankel matrix. The Hankel matrix is initially constructed from the samples of the LFR filtered voiced speech signal.

7. An accurate and noise-resilient method is devised for GCI identification that relies on the extracted time-varying F_0 component of a voiced speech signal to provide coarse estimates of intervals of where GCIs are likely to occur.

8. A robust event-based method for instantaneous F_0 estimation employing the identified GCIs as markers of respective fundamental periods is proposed.

9. A generalized approach to decompose a multi-component non-stationary signal into AM-FM mono-component signals based on the iterative EVD of the Hankel matrix is proposed. The Hankel matrix is initially constructed from the samples of the multi-component non-stationary signal. The proposed iterative decomposition approach can extract strong or strong cum weak components of a multi-component non-stationary signal. The proposed approach is used along with DESA to perform formant analysis of the voiced speech signal.

1.6 Organization of Thesis

The subsequent chapters of the thesis are structured as follows:

Chapter 2 proposes a novel feature, MEDT over the LFR to efficiently discriminate voiced and non-voiced regions of the speech signal. The MEDT over the LFR is computed from the energy distribution of speech signal over the time-frequency plane. The chapter discusses the advantage of performing V/NV detection in the LFR. A solution for an automatic threshold determination on the MEDT over the LFR is provided to reliably perform the V/NV detection in the presence or absence of noise. The chapter finally describes the proposed robust instantaneous V/NV detection method and presents a quantitative performance evaluation of the proposed method in clean and noisy environments.

Chapter 3 presents the derivation of a generalized AM-FM signal model of the speech signal in the LFR. The conditions on the square Hankel matrix size for reliable extraction of harmonically-related constant amplitude/frequency mono-component signals from a multi-component signal by performing EVD of the Hankel matrix are derived. The chapter extends the theory developed for the extraction of harmonically related constant amplitude/frequency mono-component signals contained in a multi-component signal and proposes a noise resilient iterative algorithm for extraction of the time-varying F_0 component of voiced speech signal based on repeatedly performing EVD of Hankel matrix, initially constructed from the samples of the LFR filtered speech signal.

Chapter 4 proposes an accurate and robust method for GCI identification which relies on the extracted time-varying F_0 component of voiced speech signal to provide coarse estimate of intervals where GCIs are likely to occur. The chapter provides an objective performance comparison of the proposed method with some of state of the art methods at various SNRs in different noise environments.

Chapter 5 proposes an event-based method for F_0 estimation based on the employment of identified GCIs as markers of the respective fundamental periods. The chapter presents the quantitative performance evaluation of the proposed method and provides an assessment of the variations in the obtained results with respect to gender of the speaker. It also provides an objective comparison of the gross F_0 estimation errors obtained by the proposed method and existing methods at various levels of degradation in different noise environments.

Chapter 6 proposes an approach for decomposing a multi-component non-stationary signal into AM-FM mono-component signals by iteratively performing EVD of the Hankel matrix. The Hankel matrix is initially constructed from the samples of the multi-component non-stationary signal. The efficacy of the proposed iterative decomposition approach is manifested by decomposing different kinds of synthetic and natural multi-component non-stationary signals. The proposed iterative decomposition approach is employed along with DESA to perform formant analysis of the voiced speech signal.

Chapter 7 presents the conclusions drawn with respect to findings of the study of glottal characteristics of voiced regions and performance evaluation of proposed features and proposed algorithms on the speech databases. It discusses salient features and advantages offered by the algorithms proposed in this thesis. It also mentions the scope of future work.

Chapter 2

Voiced/Non-voiced Detection

This chapter proposes a novel feature, the MEDT over the LFR to accomplish the instantaneous V/NV detection in the presence of noise. The MEDT over the LFR is computed from the energy distribution of the speech signal on the time-frequency plane, obtained by computing the pseudo Wigner-Ville distribution (PWVD) coefficients of the analytic speech signal over the low frequency range (LFR). The significance of the LFR is manifested by the MEDT over the LFR providing considerably better discrimination between voiced and non-voiced regions in comparison to the MEDT over the full frequency range. A method based on the KLD and cumulative histogram of the MEDT over the LFR is presented to automatically determine the threshold value on the MEDT over the LFR to reliably detect voiced regions in the presence of noise. The experiments were performed on speech signals of the CMU-Arctic database in different noise environments at various levels of degradation. The performance of the proposed method is demonstrated to be resilient to additive noise. The quantitative comparison of experimental results shows that the proposed method achieves a significant performance improvement over some state of the art methods.

2.1 Introduction

Voiced/Non-voiced (V/NV) detection refers to identification of regions in the speech signal with strong vocal fold activity. During the production of voiced speech, the vocal tract system is excited by vibration of the vocal folds, resulting in a quasi-periodic speech

signal. The unvoiced speech is produced when the air is passed through a narrow constriction in the wind pipe, generating a noise like random output signal. Silence occurs in the absence of any excitation to the vocal tract system and contains only background noise. Non-voiced speech includes unvoiced speech and silence. While speech signal processing applications like language identification [47], multi-rate speech coders [48, 49], speech signal modeling [50], require classification of the speech signal into voiced, unvoiced and silence (V-UV-S) regions, there are some prominent speech signal analysis applications like identification of glottal closure instants (GCIs) [23], pitch frequency estimation [21, 51], which require knowledge of only voiced regions of the speech signal. The prerequisite of boundaries of voiced regions of these applications can be catered by a V/NV detection method requiring much less computational complexity than V-UV-S classification methods. Detection of voiced regions from the speech signal in the presence of noise finds use in automatic speech recognition (ASR) [52]. Applications like speech enhancement [13], diagnosis of pathological voice disorders [11, 53], emotion recognition [18, 19] rely on the estimation of pitch frequency and detection of GCIs from noisy speech signals. A noise resilient V/NV detection method can provide reliable detection of voiced regions for pitch frequency determination and extraction of GCIs from speech signals distorted by noise.

Several methods have been proposed in the literature to distinguish V/NV regions in the speech signal. Various time domain parameters like zero crossing rate (ZCR), short-term energy estimates have been used to separate voiced/unvoiced (V/UV) regions of the speech signal [5]. However, the method is susceptible to noise. Features extracted from the linear prediction (LP) analysis of the speech signal such as the first predictor coefficient, LP residual energy have been considered to perform V-UV-S classification in [27]. The normalized low frequency energy ratio and merit of periodicity evaluated from the LP residual, harmonicity measure computed from the LP residual have been employed to decide V/UV regions in the noisy speech signal [32, 54]. The reliable estimation of the parameters of the assumed statistical distributions of multiple features used in [27] to achieve the V-UV-S classification requires large amount of training data. In order to enable the adaptive modification of the classifier; multilayer feedforward network was employed in [28] and the feature vector comprising of waveform features and cepstral

coefficients derived from LP coefficients and the LP residual energy was used to accomplish the V-UV-S classification. The LP based analysis assumes the speech signal to be stationary for about 20 - 25 ms which is not true for quickly varying phonemes such as plosives [35]. Methods based on frequency domain parameters exploit the periodic structure of the magnitude spectrum of voiced regions of the speech signal, such as in [33], the harmonic measure computed from the instantaneous frequency amplitude spectrum (IFAS) was used to perform the V/UV detection and in [52], the similarity between the shape of the signal's short-term magnitude spectrum and the spectrum of the frame analysis window was employed for voicing detection. The Gabor atomic decomposition was proposed in [55] and the generalized likelihood ratio test which measures the ratio of the energy of the harmonic part of the signal to the energy of the complementary orthogonal non-harmonic part of the signal was proposed in [56] to distinguish V/UV regions in the speech signal degraded with noise. The method in [56] requires training of the radial basis function neural network for different types of background noises. The energy of the zero frequency resonator (ZFR) filtered signal was shown to provide efficient characterization of the glottal activity in the presence of noise [34]. However, an estimate of the pitch period is a prerequisite for this method. The property of noise robustness of GCIs present during voiced regions of the speech signal was explored in [25] to detect V/NV regions. The limitation of the method is that it requires fixed level of noise (SNR: 10 dB) to be added to the noisy speech signal. Moreover, thresholds have to be set on various parameters such as GCI drift, jitter, pitch period and excitation strength.

One of the major drawbacks of methods mentioned above, is that they cannot provide V/NV decision at each sample instant of the speech signal. Hence, there lies a strong motivation to develop an instantaneous V/NV detection technique to detect voiced regions in the speech signal. This chapter presents a robust instantaneous V/NV detection method based on the analysis of the speech signal over the low frequency range (LFR). The proposed method does not require any prior information about the pitch frequency or GCIs. The proposed method exploits the property that vibration of the vocal folds during the production of the voiced speech, produces substantial energy only around the pitch frequency and its few harmonics (included in the LFR). The energy in the LFR

is negligible during non-voiced regions of the speech signal. The MEDT over the LFR computed using the pseudo Wigner-Ville distribution (PWVD) coefficients of the analytic speech signal is employed as a feature to detect voiced regions of the speech signal. The reason for choosing the PWVD technique for the time-frequency analysis of the speech signal over the LFR is that it offers excellent time-resolution at all frequencies (including the LFR) without introducing cross-terms between nonconcurrent auto-components of the multi-component speech signal. The cross-terms introduced by the PWVD technique between concurrent auto-components of the speech signal in the LFR aid in the V/NV detection as explained in detail in Section 2.3. This chapter is organized as follows: Section 2.2 details the computation of the MEDT using the PWVD technique. The proposed instantaneous method for the V/NV detection is explained in Section 2.3. The experimental results obtained by the proposed method in various noise environments are presented in Section 2.4. The concluding remarks are provided in Section 2.5.

2.2 Computation of the MEDT using the PWVD technique

The Wigner-Ville distribution (WVD) is a quadratic time-frequency analysis technique that provides optimum temporal and frequency resolutions. The WVD of a discrete-time signal $s[n]$, denoted $S_{\mathrm{W}}[n, f]$ is given by [57]:

$$S_{\mathrm{W}}[n, f] = 2 \sum_{m=-\infty}^{\infty} \overline{s[n-m]} s[n+m] \exp(-j4\pi f m) \tag{2.1}$$

where $f = \dfrac{F}{F_s}$ and $(^{-})$ denotes the conjugate operator. The frequency in Hz, the normalized frequency and the sampling frequency of the discrete-time signal $s[n]$ are represented by F, f and F_s respectively. One of the major limitations of the WVD technique is that it introduces cross-terms between auto-components of a multi-component signal occurring at mid-time and mid-frequency of auto-components [58]. Cross-terms are introduced for each pair of auto-components. Let $s[n]$ be a discrete-time multi-component signal which

can be expressed as a linear sum of its K auto-components as follows:

$$s[n] = \sum_{k=1}^{K} s_k[n] \tag{2.2}$$

The WVD $S_W[n, f]$ of the multi-component signal $s[n]$ in (2.2) is given by [59]:

$$S_{\mathrm{W}}[n, f] = \sum_{k=1}^{K} S_{\mathrm{W}}^{k}[n, f] + \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} 2\Re\left(S_{\mathrm{W}}^{k,l}[n, f]\right) \tag{2.3}$$

The first term and the second term on the right hand side of (2.3) are associated with the WVD of auto-components of $s[n]$ and cross-terms respectively. There will be $\binom{K}{2}$ cross-terms present in the WVD $S_{\mathrm{W}}[n, f]$ of the multi-component signal $s[n]$. A technique to reduce cross-terms in the WVD based on the separation of auto-components of a multi-component signal using the Fourier-Bessel (FB) expansion was suggested in [59]. In order to eliminate cross-terms introduced between nonconcurrent auto-components of a multi-component signal in the WVD technique, a window function is employed in the PWVD to emphasize the signal properties near the time of interest compared to far away times. The PWVD of a discrete-time signal $s[n]$ denoted by $S_{\mathrm{PW}}[n, f]$ is defined as [60]:

$$S_{\mathrm{PW}}[n, f] = 2 \sum_{m=-M}^{M} w[m]\overline{s[n-m]}s[n+m]\exp(-j4\pi fm) \tag{2.4}$$

where $w[m]$ represents the real frequency smoothing window function (FSWF) with finite time support of $2M+1$ samples. The purpose of the FSWF is to make the summation in (2.1) numerically computable and to eliminate the cross-terms occurring between nonconcurrent auto-components of a multi-component signal. However, cross-terms still occur at mid-frequencies of concurrent auto-components in the PWVD technique [61]. The spectrum of the real signal consists of both negative and positive frequency components. In order to further eliminate the cross-terms occurring at mid-frequencies of negative and positive frequencies associated with the concurrent auto-components of a real signal, the analytic signal is used to compute the PWVD. Also note that the periodicity of the frequency variable in (2.4) is half of the sampling frequency F_s; therefore, in order to avoid

aliasing, the PWVD is computed for the analytic signal [62]. The discrete-time analytic signal, denoted $z[n]$ associated with the discrete-time real signal $s[n]$ spanning $(0, N-1)$ samples is given by [62]:

$$z[n] = s[n] + j\hat{s}[n] \qquad n = 0, 1, ...N-1 \tag{2.5}$$

where $\hat{s}[n]$ represents the Hilbert transform of $s[n]$. The Hilbert transform of $s[n]$ can be determined by employing a finite impulse response (FIR) filter [62] but the error introduced in the computation of the Hilbert transform increases with a decrease in the employed FIR filter length [63]. Hence, we have used a frequency domain method described in [64] to obtain $z[n]$ from $s[n]$. The computation of the discrete-time analytic signal consists of the following steps:

1. Compute the N-point discrete-time Fourier transform (DTFT) denoted by $S[v]$ at N discrete normalized frequencies $\left(f_v = \frac{v}{N}\right)$ of the discrete-time real signal $s[n]$ consisting of N samples by using the following equation:

$$S[v] = \sum_{n=0}^{N-1} s[n]\exp\left(\frac{-j2\pi vn}{N}\right) \qquad v = 0, 1, ..., N-1 \tag{2.6}$$

2. Form the N-point one-sided discrete-time analytic signal transform represented by $Z[v]$ as follows:

$$Z[v] = \begin{cases} S[0], & v = 0 \\ 2S[v], & 1 \leq v \leq \frac{N}{2} - 1 \\ S\left[\frac{N}{2}\right], & v = \frac{N}{2} \\ 0, & \frac{N}{2} + 1 \leq v \leq N-1 \end{cases} \tag{2.7}$$

3. Compute the N-point inverse DTFT to obtain the discrete-time analytic signal denoted by $z[n]$ with the same sampling rate as the original signal $s[n]$ by using the following equation:

$$z[n] = \frac{1}{NT_s}\sum_{v=0}^{N-1} Z[v]\exp\left(\frac{j2\pi vn}{N}\right) \qquad n = 0, 1..., N-1 \tag{2.8}$$

where T_s denotes the sampling time period of $s[n]$. The fast Fourier transform (FFT) algorithm and inverse FFT are used for the computation of the DTFT and the inverse DTFT respectively. The PWVD of a discrete-time analytic signal $z[n]$, denoted $Z_{PW}[n, f]$ is obtained by substituting $z[n]$ in the place of $s[n]$ in (2.4) as follows:

$$Z_{\mathrm{PW}}[n, f] = 2 \sum_{m=-M}^{M} w[m]\overline{z[n-m]}z[n+m]\exp(-j4\pi fm) \tag{2.9}$$

The cross-terms appearing at mid-frequencies of the positive frequencies associated with the concurrent auto-components of the multi-component analytic signal $z[n]$ aid in the V/NV detection as demonstrated in the next section. Let $E^A[n, f]$ represents the energy distribution of the discrete-time analytic signal $z[n]$ on the time-frequency plane. $E^A[n, f]$ can be computed from the magnitude of PWVD coefficients, $Z_{PW}[n, f]$ as follows [65]:

$$E^A[n, f] = |Z_{\mathrm{PW}}[n, f]| \tag{2.10}$$

The marginal energy density with respect to time (MEDT) denoted by $E^A[n]$ of the analytic signal $z[n]$ over the frequency range (f_1, f_2) can be derived from the energy distribution of the analytic signal on the time-frequency plane $E^A[n, f]$ as:

$$E^A[n] = \sum_{f=f_1}^{f_2} E^A[n, f] \tag{2.11}$$

2.3 Proposed V/NV Detection Method based on the MEDT over the LFR

It has been demonstrated in [66] that the discrete-time speech signal in the LFR, denoted $s_{\mathrm{LF}}[n]$, is a multi-component signal which can be expressed using the amplitude and frequency modulated (AM-FM) signal model as:

$$s_{\mathrm{LF}}[n] = \sum_{k=1}^{I} A_k[n] \cos\left(2\pi k f_0[n] n + \theta_k[n]\right) \tag{2.12}$$

where I denotes the number of harmonic components in the LFR. The normalized time-varying fundamental frequency or pitch frequency is denoted by $f_0[n]$. The time-varying amplitude and phase of the k^{th} harmonic of the pitch frequency are denoted by $A_k[n]$ and $\theta_k[n]$ respectively. In the LFR, the time-varying amplitude $A_k[n]$ has significant values only for voiced regions of the speech signal and is nearly zero for non-voiced regions of the speech signal. The pitch frequency of speech signals ranges from 50 Hz - 500 Hz [1]. In order to suppress the DC component, formants, remove noise energy present outside the LFR and include the pitch frequency component and its few harmonics, we have chosen the range of frequencies from 50 Hz - 500 Hz as the LFR. The speech signal segment and its spectrogram in the LFR are shown in Fig. 2.1. Note that the speech signal segment in Fig. 2.1 (a) consists of voiced, unvoiced and silence regions. It is observed from the spectrogram depicted in Fig. 2.1 (c) that substantial energy is present around the pitch frequency component (200 Hz) and the second harmonic component (400 Hz) on the time-frequency plane at instants corresponding to voiced regions while negligible energy is present during non-voiced regions of the speech signal. The reference voiced region is derived from the differenced electroglottograph (DEGG) signal depicted in Fig. 2.1 (b). The DEGG signal is the first-order derivative of the electroglottograph (EGG) signal.

The energy distribution on the time-frequency plane over the LFR obtained from the PWVD coefficients of the analytic speech signal segment using (2.9) and (2.10) is depicted in Fig. 2.2. The analytic speech signal segment is computed from the real speech segment shown in Fig. 2.1 (a) using (2.6), (2.7), (2.8). Note the presence of cross-terms that have occurred at mid frequencies of the pitch frequency component and its harmonics (cross-terms at 300 Hz and 500 Hz) during the voiced region in Fig. 2.2. These cross-terms increase the energy during voiced regions of a speech signal and aid in discrimination of voiced and non-voiced regions of the speech signal. The MEDT over the LFR can be derived from the PWVD coefficients of the analytic speech signal segment as follows:

$$E^{A}_{\mathrm{LF}}[n] = \sum_{f \,\epsilon\, \mathrm{LFR}} |Z_{\mathrm{PW}}[n, f]| \tag{2.13}$$

The speech segment and the MEDT over the LFR derived from the energy distribution

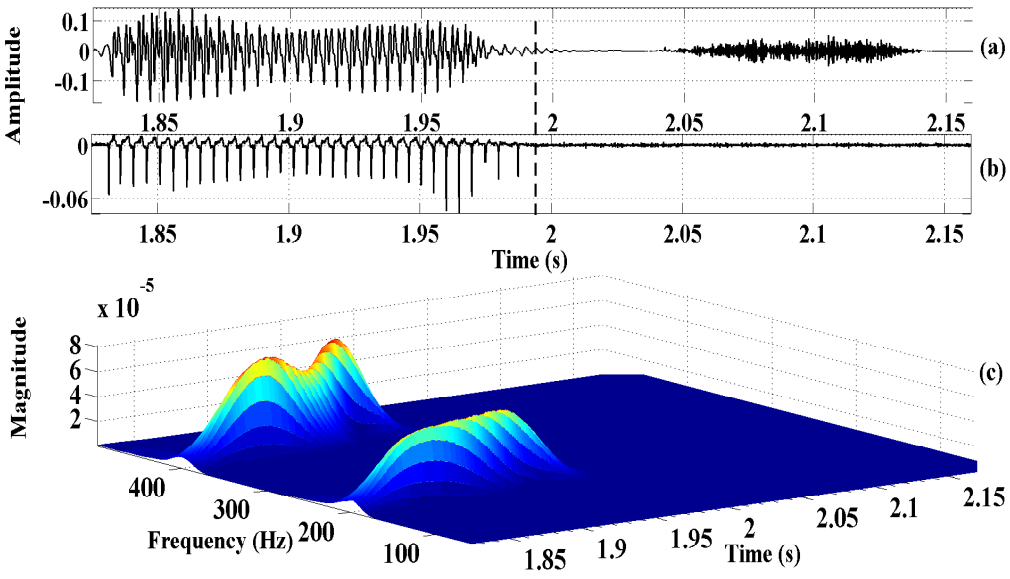

Figure 2.1: (a) Speech segment (b) DEGG signal (c) Spectrogram of the speech segment. The reference voiced region is marked by the dashed line.

of the analytic speech segment using (2.6), (2.7), (2.8), (2.9) and (2.13) are shown in Fig. 2.3. The reference voiced region is detected using the DEGG signal. It is clear from Fig. 2.3 (c) that the MEDT over the LFR, $E^{A}_{\mathrm{LF}}[n]$ computed from the PWVD coefficients of the analytic speech segment can be used as a feature to provide the instantaneous V/NV decision by applying a suitable threshold, such that the samples having the value of the MEDT over the LFR above the threshold are considered as voiced and the samples having the value of the MEDT over the LFR below the threshold are considered as non-voiced. Automatic threshold determination and the advantage of choosing the LFR for the time-frequency analysis of the speech signal to accomplish the V/NV detection are discussed in the sub-section 2.3.1.

2.3.1 Automatic threshold determination

A set of speech signals is created by randomly selecting 300 speech signals from the phonetically balanced CMU-Arctic speech database [67,68] for determination of a suitable threshold on the MEDT over the LFR to achieve the V/NV detection in the absence or

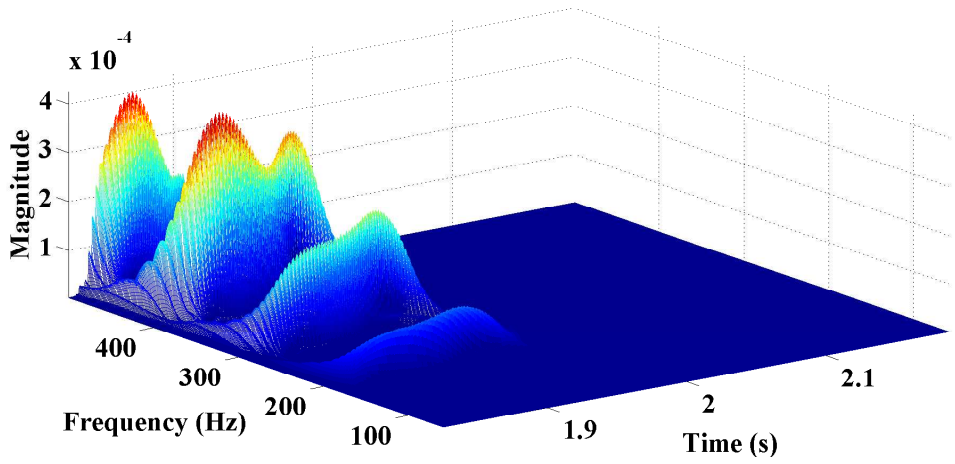

Figure 2.2: Energy distribution over the LFR of the analytic speech segment using the PWVD technique. The speech segment is shown in Fig. 2.1 (a).

presence of noise. The speech signals are of about 3 s duration with a sampling frequency of 32 kHz. The sampling frequency of speech signals is reduced to 8 kHz. The cumulative distribution functions (CDFs) of the MEDT over the LFR for voiced and non-voiced regions obtained for the set of speech signals are shown in Fig. 2.4 (a) and Fig. 2.4 (b) respectively. The CDF computation of the MEDT over the LFR is based on the cumulative histogram. It is evident from Fig. 2.4 that the values of the MEDT over the LFR for the voiced regions is very large (nearly 25 times) compared to the insignificant values of the MEDT over the LFR for the non-voiced regions. It can be inferred from Fig. 2.4 (b) that there is 99% probability that the value of the MEDT over the LFR for non-voiced regions of all speech signals from the set is below 0.0002 and it can be observed from Fig. 2.4 (a) that 95.5% values of the MEDT over the LFR during voiced regions of all speech signals from the set are above the value 0.0002. It implies that the MEDT over the LFR provides excellent discrimination between voiced and non-voiced regions of the speech signal. The analysis of the speech signal in the LFR also facilitates the removal of high frequency components which may be present in the noise environment. Speech signals contain significant energy in the frequency range of (0 Hz - 3400 Hz) [69]. The CDFs of the MEDT over the frequency range (0 Hz - 3400 Hz) for voiced and non-voiced

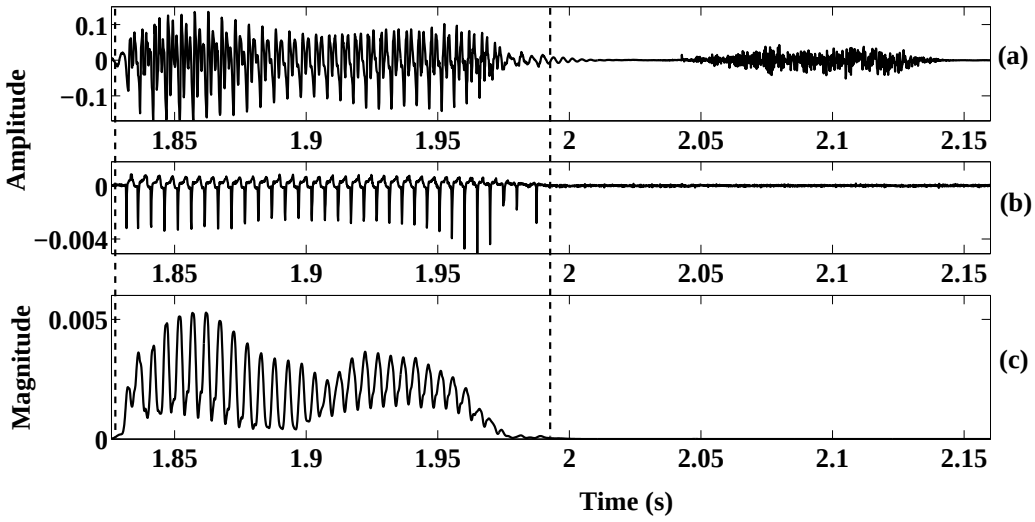

Figure 2.3: (a) Speech segment (b) DEGG signal (c) MEDT over the LFR derived from the PWVD coefficients of the analytic speech segment. The reference voiced region is shown by the dashed line.

regions obtained for the set of speech signals are shown in Fig. 2.5 (a) and Fig. 2.5 (b) respectively. It can be noted from Fig. 2.5 that the MEDT over the frequency range (0 Hz - 3400 Hz) has comparable values for both voiced and non-voiced regions. It can be deduced from Fig. 2.5 (b) that there is 99% probability that the value of the MEDT over the frequency range (0 Hz - 3400 Hz) for non-voiced regions of all speech signals from the set is below 0.00075 and it is evident from Fig. 2.5 (a) that 18.29% values of the MEDT over the frequency range (0 Hz - 3400 Hz) for voiced regions of all speech signals from the set are below 0.00075. There is a considerable overlap between the values of the MEDT over the frequency range (0 Hz - 3400 Hz) for voiced and non-voiced regions of speech signals from the set. Hence the LFR (50 Hz - 500 Hz) is chosen as the desired frequency range for the V/NV detection. The ratio of the minimum value of the MEDT over the LFR to the maximum value of the MEDT over the LFR for voiced regions of any speech signal taken from the set has been found out to be 0.00924 (nearly 1% percent) on an average. The percentage of non-voiced samples that lie below 1% of the maximum value of the MEDT over the LFR for any speech signal taken from the set has been found out to be 98.41% on an average. Thus, for the set of speech signals created from the CMU-Arctic database, we define the first preliminary threshold R_{TH1} to be one percent of the maximum value of the MEDT over the LFR computed for the speech signal under

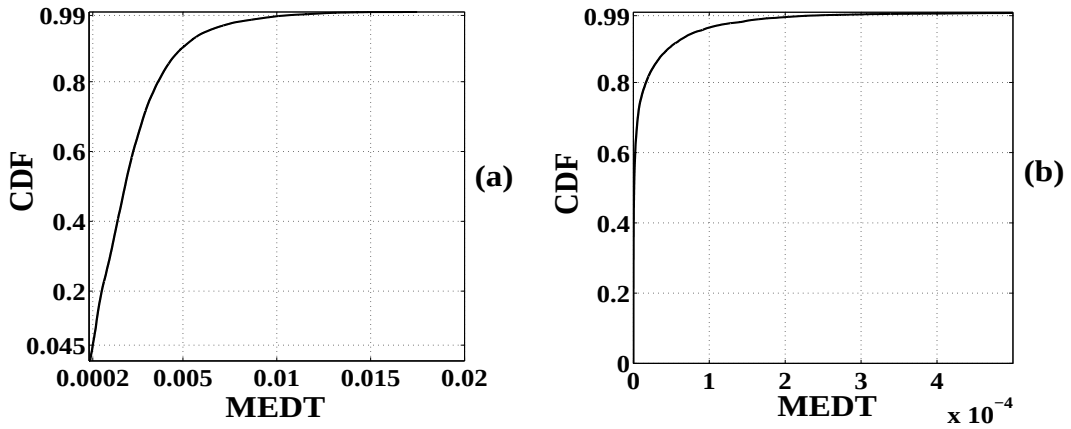

Figure 2.4: CDF of the MEDT over the LFR for (a) Voiced regions (b) Non-voiced regions.

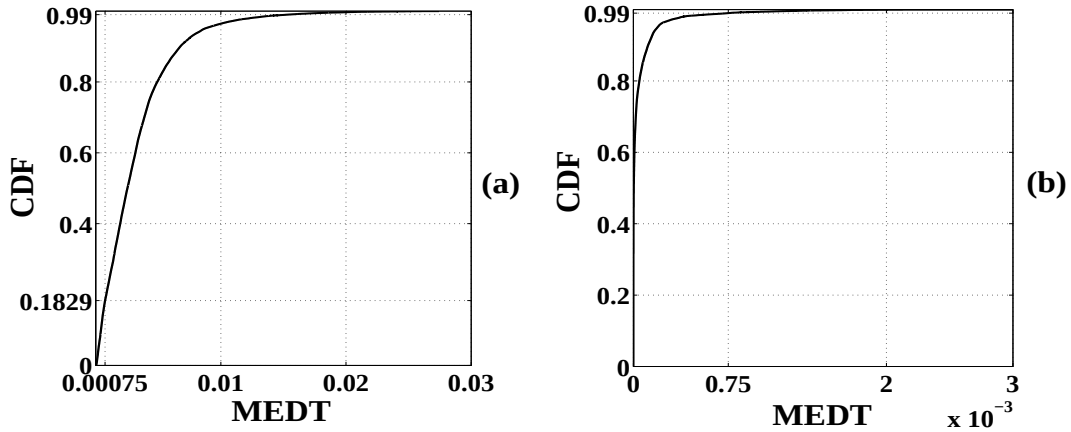

Figure 2.5: CDF of the MEDT over the frequency range (0 Hz - 3400 Hz) for (a) Voiced regions (b) Non-voiced regions.

consideration as: $R_{\text{TH1}} = 0.01 \times \max(E^A_{\text{LF}}[n]) \ \ \forall\ n$.

While recording, the speech signal is distorted by the ambient noise. In applications like diagnosis of pathological disorders [11, 53], emotion recognition [18, 19], speaker recognition [70], the recording conditions can be controlled to ensure stationary noise conditions. In the presence of the additive noise, the energy of the noise signal is distributed across the entire time-frequency plane and augment to the energy distribution of the speech signal. In order to simulate and estimate the noise floor for stationary noise environments, we have appended the silence duration at the beginning of the speech signal and then added the noise realization from the NOISEX-92 database [71] after reducing the sampling frequency of the noise signal to 8 kHz. Thus, the resultant noisy speech signal contains only the noise during the appended silence duration. The length of the silence duration appended at the beginning of the speech signal is critical for reliable estimation of the underlying noise process. The duration length of the appended silence must be such that the noise contained in it must be able to characterize the noise added to the entire duration of the speech signal. The Kullback-Leibler divergence (KLD) denoted by $\text{KLD}(p_1||p_2)$ is a non-symmetric measure of the difference between two probability density functions (PDFs) represented by $p_1(x)$ and $p_2(x)$ and is given by [72]:

$$\text{KLD}(p_1||p_2) = \int p_1(x) \log_{\text{e}} \left(\frac{p_1(x)}{p_2(x)} \right) dx \tag{2.14}$$

The value of the $\text{KLD}(p_1||p_2)$ becomes zero when both the PDFs are identical. It can be deduced from Table 2.1 that the noise contained in 500 ms of the appended silence duration is optimum to characterize the noise added to the entire duration of the speech signal for different types of noise realizations, as the value of the KLD between the PDF of the noise contained in the appended silence duration and the PDF of the noise added to the entire duration of the speech signal is below or approximately equal to 1%. We define the second preliminary threshold represented by R_{TH2} as follows:

$$P\left(E^A_{\text{LF}}[n] \leq R_{\text{TH2}}\right) = 0.99, \quad n \quad \in \quad \text{appended silence duration} \tag{2.15}$$

where P denotes the probability operator. The value of the second preliminary threshold

Table 2.1: Kullback-Leibler Divergence (KLD) between the PDF of the noise contained in the appended silence duration and the PDF of the noise added to the entire duration of the speech signal for different types of noises and various silence durations.

Silence Duration (ms)	AWGN	Babble	Vehicular
31.25	0.0732	0.0576	0.0932
62.50	0.0450	0.0411	0.0920
125.0	0.0328	0.0304	0.0805
250.0	0.0203	0.0210	0.0515
312.5	0.0149	0.0177	0.0361
375.0	0.0108	0.0106	0.0269
437.5	0.0082	0.0089	0.0165
500.0	0.0071	0.0076	0.0112

computed as in (2.15) ensures that, on an average 99% of the samples belonging to silence regions of the speech signal have the value of the MEDT over the LFR below the value of R_{TH2} in the presence of noise. It is already shown in Fig. 2.3 (c) and Fig. 2.4 (b) that the MEDT over the LFR during unvoiced regions of the speech signal has negligible values. Thus, the second preliminary threshold allows for rejection of nearly 99% of non-voiced samples. The final threshold Υ is chosen as the greater of the two preliminary thresholds, R_{TH1} and R_{TH2}. At high SNRs, there is negligible energy in the appended silence duration and the value of R_{TH1} will be greater than R_{TH2}. At low SNRs, the noise contained in the appended silence region has significant energy and hence the value of R_{TH2} will be greater than R_{TH1}. The proposed V/NV detection algorithm is explained in the next subsection.

2.3.2 V/NV detection algorithm

The proposed algorithm for the V/NV detection consists of the following steps:

1. Append 500 ms of silence duration at the beginning of the speech signal $s[n]$ spanning $(0, Q-1)$ samples. At a speech signal sampling frequency of 8 kHz, 4000 samples of silence duration gets appended at the beginning of the speech signal.

2. The noisy speech signal $y[n]$ resulted from the distortion of the speech signal $s[n]$ by the additive noise $\xi[n]$ can be expressed as:

$$y[n] = s[n] + \xi[n] \qquad n = 0, 1, ..., Q-1 \tag{2.16}$$

3. Divide the noisy speech signal into segments of 250 ms. Let $y_l[n]$ denotes the l^{th} noisy speech signal segment spanning $(0, N-1)$ samples.

4. Compute the analytic speech signal segment $z_l[n]$ of the noisy speech signal segment $y_l[n]$ using (2.6), (2.7), (2.8).

5. Compute the PWVD coefficients $Z_{\mathrm{PW},l}[n, f]$ over the LFR for the analytic speech segment $z_l[n]$ using (2.9).

6. Compute the MEDT over the LFR, $E_{\mathrm{LF},l}[n]$ for the analytic speech segment $z_l[n]$ from its PWVD coefficients $Z_{\mathrm{PW},l}[n, f]$ using (2.13).

7. Repeat steps 4-6 for each noisy speech signal segment and then obtain the MEDT over the LFR, $E^A_{\mathrm{LF}}[n]$ for the entire duration of the speech signal by concatenating the MEDT over the LFR for each noisy speech signal segment, $E_{\mathrm{LF},l}[n]$ for $l = 1, 2, ..., L$, one after the other as follows:

$$E^A_{\mathrm{LF}}[n] = (E_{\mathrm{LF},1}[n] \;\; E_{\mathrm{LF},2}[n] ... \;\; E_{\mathrm{LF},L}[n]) \tag{2.17}$$

   where L denotes the total number of noisy speech signal segments.

8. Obtain the smoothed MEDT over the LFR denoted by $SE^A_{\mathrm{LF}}[n]$ by applying the moving average filter as follows:

$$SE^A_{\mathrm{LF}}[n] = \frac{1}{2C+1} \sum_{m=-C}^{C} E^A_{\mathrm{LF}}[n+m] \tag{2.18}$$

   The length of the moving average filter $2C + 1$ is not critical and is chosen to be equal to the highest possible pitch period duration of 20 ms corresponding to the pitch frequency of 50 Hz. We have kept the length of the moving average filter fixed.

9. Determine two preliminary thresholds R_{TH1} and R_{TH2}. The value of R_{TH1} is calculated as 1% of the maximum value of the $SE^A_{\mathrm{LF}}[n]$ and the value of R_{TH2} is computed from the values of $SE^A_{\mathrm{LF}}[n]$ obtained during the appended silence duration (refer to

step 1), as follows:

$$\begin{aligned} R_{\text{TH1}} &= 0.01 \times \max(SE^{A}_{\text{LF}}[n]) \qquad n = 0, 1, ..., Q-1 \\ P(SE^{A}_{\text{LF}}[n] \leq R_{\text{TH2}}) &= 0.99 \qquad n = 0, 1, ..., 3999 \end{aligned} \tag{2.19}$$

The final threshold Υ is equal to the greater of the two preliminary thresholds.

10. Identify voiced regions in the speech signal by comparing the smoothed MEDT over the LFR, $SE^{A}_{\text{LF}}[n]$ with the final threshold Υ at each sample instant as follows:

$$\begin{aligned} SE^{A}_{\text{LF}}[n] &\geq \Upsilon,\ y[n]\ \epsilon\ \text{Voiced} \\ SE^{A}_{\text{LF}}[n] &< \Upsilon,\ y[n]\ \epsilon\ \text{Non-Voiced} \end{aligned} \tag{2.20}$$

Please note here that the V/NV decision is taken instantaneously; i.e., at each sample instant of the speech signal.

2.4 Experimental Results and Discussion

The experimental results of the proposed method are obtained on clean and noisy speech signals. The SNR in dB denoted by SNR_{dB} is defined as:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \frac{P_s}{P_\xi} \tag{2.21}$$

where P_s and P_ξ represents the power in the speech signal and the noise signal respectively. For clean speech signals; $y[n] = s[n]$ in (2.16). The power in the speech signal $s[n]$ is calculated without taking into account the appended silence duration i.e if $s[n]$ contains 500 ms (4000 samples at $F_s = 8$ kHz) of appended silence duration at its beginning and spans (0, Q - 1) samples, then P_s is calculated as:

$$P_s = \frac{1}{(Q - 4000)} \sum_{n\ =\ 4000}^{Q-1} s^2[n] \tag{2.22}$$

In all subsequent figures illustrating the results of the proposed method on clean and noisy speech signals, the detected voiced regions are shown in dotted lines and the reference

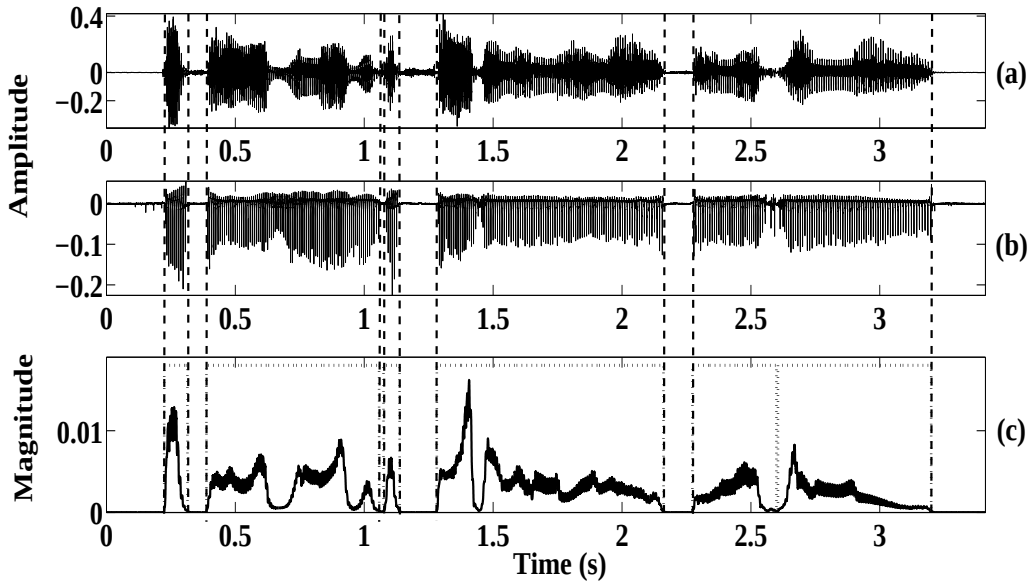

Figure 2.6: (a) Clean male speech signal (b) DEGG signal (c) Smoothed MEDT over the LFR using the PWVD technique.

voiced regions derived from the respective DEGG signal are marked in dashed lines.

The results obtained by the proposed method on clean male and female speech signals are shown in Fig. 2.6 and Fig. 2.7 respectively. As evident from the figures, the SMEDT over the LFR obtained using the PWVD technique shown in Fig. 2.6 (c) and Fig. 2.7 (c) has significant values during the voiced regions and negligible values during the non-voiced regions of the speech signal. The detected voiced regions obtained for clean male and female signals in Fig. 2.6 (c) and Fig. 2.7 (c) match very well with the reference voiced regions. Fig. 2.8 depicts the result obtained by the proposed method on a male speech signal at a low SNR of 0 dB in a white noise environment. Please note that the false detection (detection of non-voiced regions as voiced regions) has occurred (Fig. 2.8 (c)) at around 1 ms and the missed detection (detection of voiced regions as non-voiced regions) has occurred (Fig. 2.8 (c)) during some regions of the speech signal but still the proposed method was able to detect many of the voiced regions, obtaining a good detection accuracy of 96.98%. The detection accuracy is defined later in this section. The result obtained by the proposed method on a female speech signal at low SNR of 0 dB in a white noise environment is shown in Fig. 2.9. Please note that the false detection

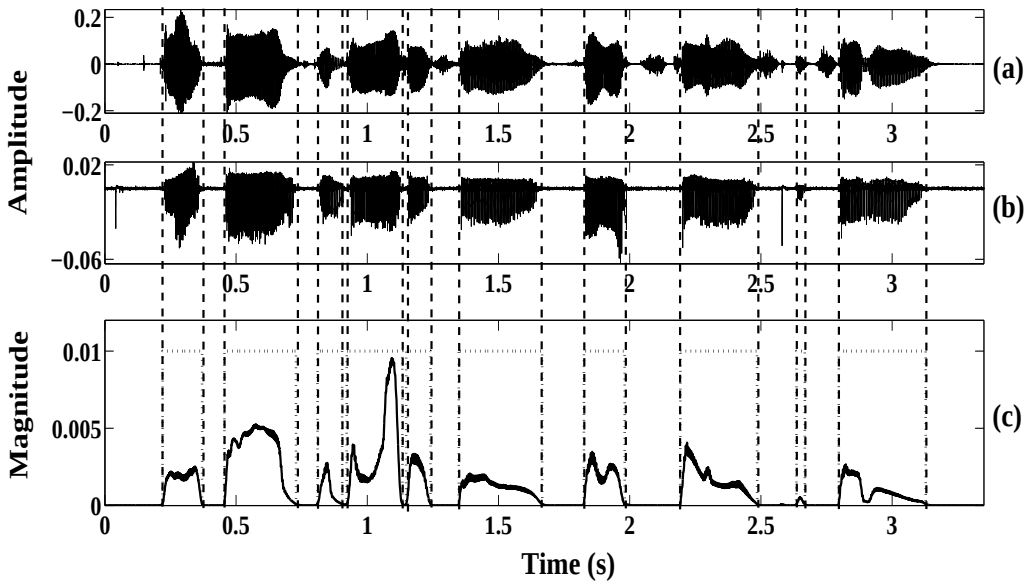

Figure 2.7: (a) Clean female speech signal (b) DEGG signal (c) Smoothed MEDT over the LFR using the PWVD technique.

has occurred (Fig. 2.9 (c)) at around 1 ms, 2.56 s, 2.58 s and the missed detection has occurred (Fig. 2.9 (c)) during some regions of the speech signal. A detection accuracy of 97.13% was obtained by the proposed method in this case.

The results obt ained by the proposed method on male and female speech signals at 5 dB SNR in a babble noise environment are shown in Fig. 2.10 and Fig. 2.11 respectively. It can be observed in Fig. 2.10 (c) and Fig. 2.11 (c) that some durations belonging to voiced regions of the speech signals are detected as non-voiced regions leading to missed detection. The false detection has occurred at around 0.07 s, 2.24 s, 3.31 s and 3.37 s in Fig. 2.10 (c). In Fig. 2.11 (c), the false detection has occurred at around 0.07 s, 1.74 s, 1.81 s, 2.05 s, 2.51 s and 2.74 s. The proposed method has obtained the V/NV detection accuracies of 94.52% and 93.92% in experimental cases depicted in Fig. 2.10 and Fig. 2.11 respectively. Fig. 2.12 and Fig. 2.13 depict the results obtained by the proposed method on male and female speech signals respectively, at 5 dB SNR in the vehicular noise environment. It is evident in Fig. 2.12 (c) and Fig. 2.13 (c) that some durations belonging to voiced regions of speech signals are detected as non-voiced regions leading to missed detection. The false detection has occurred at around 0.03 s, 0.1 s, 1.18 s, 2.18

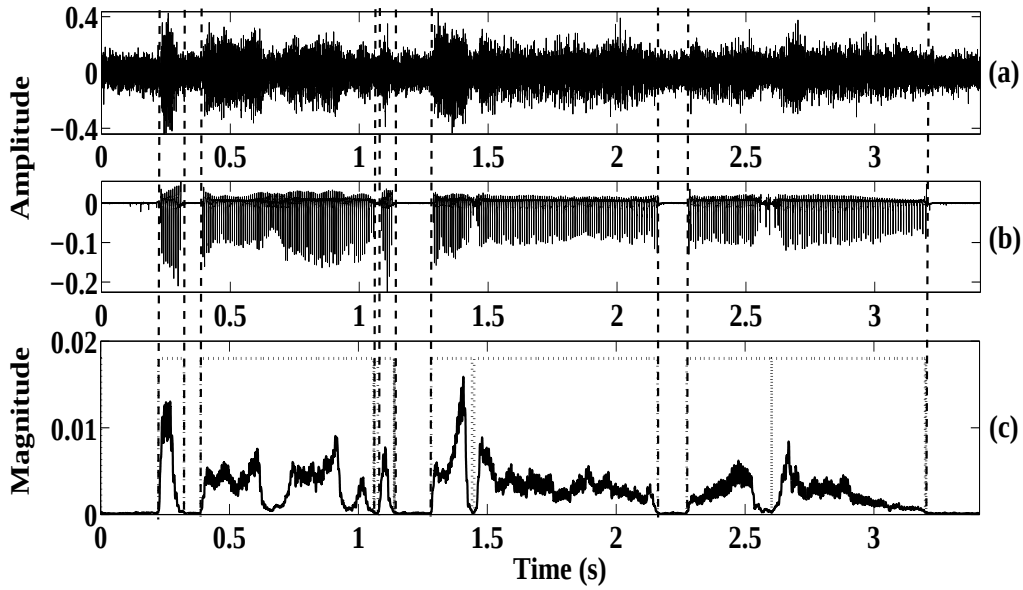

Figure 2.8: (a) Male speech signal at 0 dB SNR (white noise) (b) DEGG signal (c) Smoothed MEDT over the LFR using the PWVD technique.

s, 3.28 s in Fig. 2.12 (c). In Fig. 2.13 (c), the false detection has occurred at around 0.09 s, 1.14 s, 2.17 s and 2.73 s. The proposed method has obtained the V/NV detection accuracy of 93.81% and 94.03% in experimental cases shown in Fig. 2.12 and Fig. 2.13 respectively.

The performance of the proposed method is evaluated by performing experiments on a set of 300 speech signals containing 100 speech signals spoken by each of the two male speakers and one female speaker, randomly selected from the CMU-Arctic database [67,68] in different noise environments taken from the NOISEX-92 database [71] at various SNRs. The CMU-Arctic database consists of around 1150 phonetically balanced sentences of about 3 s duration, sampled at 32 kHz, spoken by five male and two female speakers with simultaneous recordings of EGG signals available for two male speakers and one female speaker. In order to compensate for the larynx to microphone delay which was determined to be 0.7 ms, the time alignment of speech signals and EGG signals was performed in the CMU-Arctic database. The NOISEX-92 database consists of various noise environments sampled at 19.98 kHz. The sampling frequency of speech and noise signals is reduced to 8 kHz. We have selected white, babble and vehicular noise environments to evaluate the

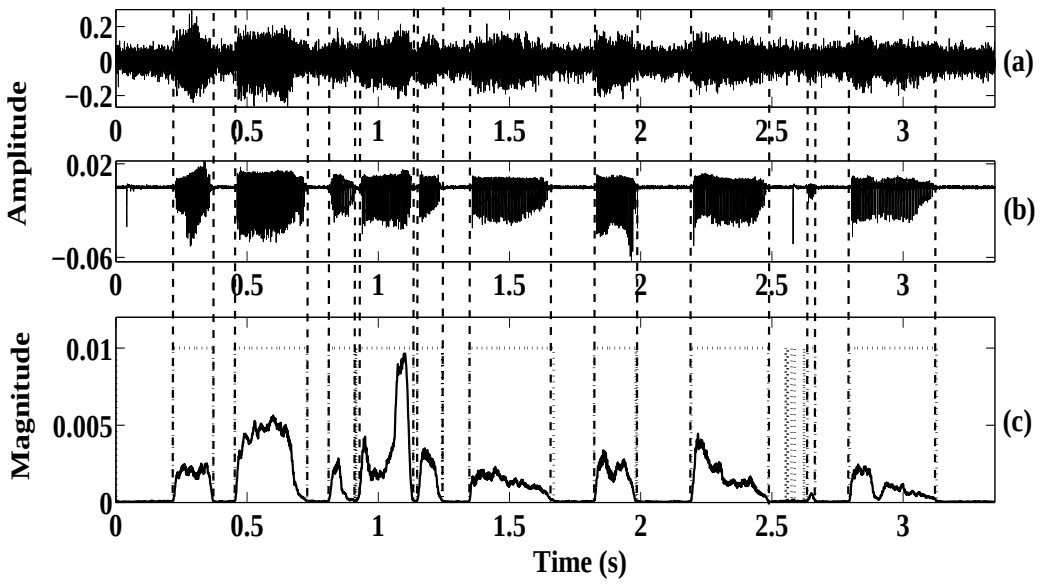

Figure 2.9: (a) Female speech signal at 0 dB SNR (white noise) (b) DEGG signal (c) Smoothed MEDT over the LFR using the PWVD technique.

performance of the proposed method. The white noise environment is characterized by a flat power spectral density (PSD) and the multi-talker babble noise and the vehicular noise environments contain higher energy in the low frequency components. The performance measures mentioned in [25] have been used which were defined in terms of number of percentage of epochs (also known as GCIs) detected as voiced or non-voiced. The lowest time resolution achieved by the V/NV method in [25] is the minimum duration between two successive epochs because the V/NV decision is taken at each epoch location. The duration between two successive epochs contain many samples. The lowest time resolution achieved by the proposed method is the sampling period because the V/NV decision is taken at each sample instant. Therefore, in order to meet the requirements for the instantaneous V/NV detection offered by the proposed method, the definition of the performance measures in [25] are slightly modified here by expressing them in terms of number of samples detected as voiced or non-voiced. The performance measures are defined below:

- Percentage Detection Accuracy (P_d): It is the ratio of correctly detected samples to total number of samples in the speech signal. A correct decision implies that the

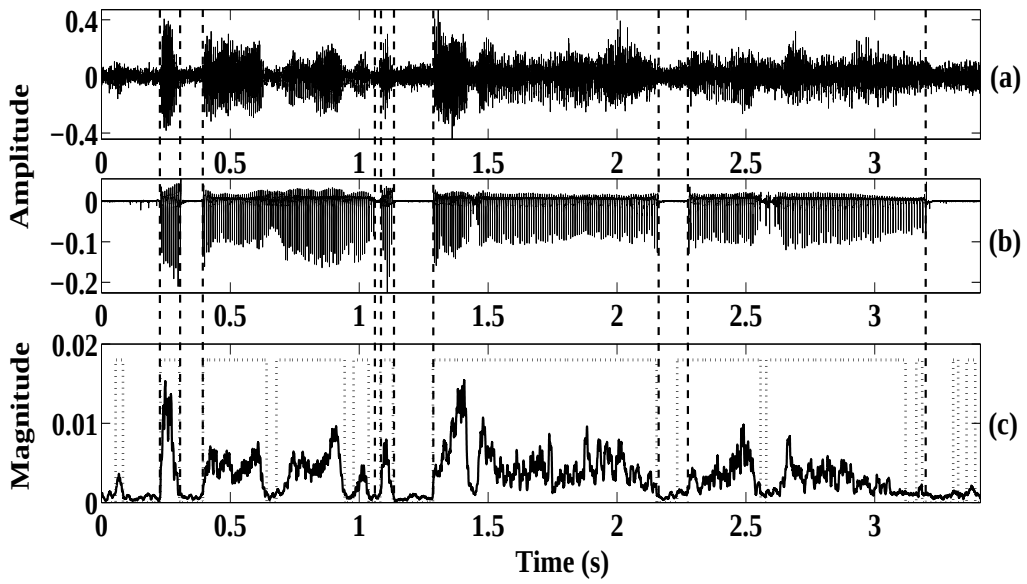

Figure 2.10: (a) Male speech signal at 5 dB SNR (babble noise) (b) DEGG signal (c) Smoothed MEDT over the LFR using the PWVD technique.

sample belonging to the voiced region is detected as voiced and the sample belonging to the non-voiced region is detected as non-voiced.

- Missed Detection Percentage (P_m): It the ratio of the samples that belong to voiced regions but are incorrectly detected as non-voiced to the total number of samples in the speech signal.

- False Alarm Percentage (P_f): It is the ratio of the samples that belong to non-voiced regions but are incorrectly detected as voiced to the total number of samples in the speech signal. All measures are expressed as percentage.

The percentage detection accuracy of 98.45% has been obtained by the proposed method for clean speech signals from the set. The missed detection and the false detection percentage of 0.2% and 1.35% respectively were obtained for clean speech signals. We have compared the performance of the proposed method with other state of the art methods: Wavesurfer [73, 74], the method based on the robustness of GCIs [25] and the method based on the ZFR filtered signal energy [34]. Wavesurfer is an open source utility which relies on the normalized cross correlation based pitch tracking refined by dynamic pro-

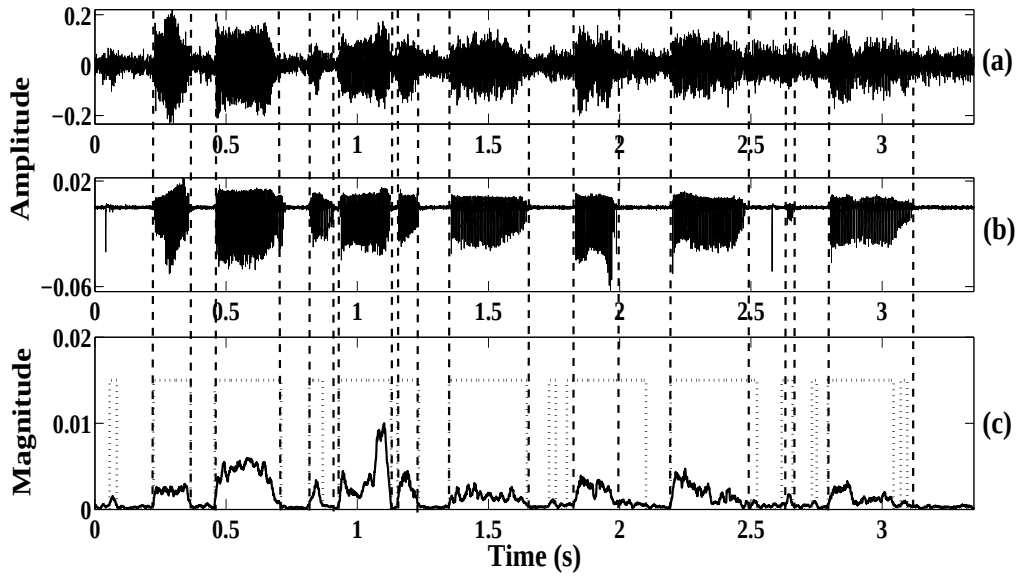

Figure 2.11: (a) Female speech signal at 5 dB SNR (babble noise) (b) DEGG signal (c) Smoothed MEDT over the LFR using the PWVD technique.

gramming. The method in [25] requires the detection of GCIs from the noisy speech signal using the ZFR for two different realizations of the Gaussian white noise added to the speech signal at the SNR of 10 dB; irrespective of the SNR of the speech signal taken into consideration. The GCIs detected during voiced regions for two different noise realizations show small drift as compared to the large drift incurred by the GCIs detected during non-voiced regions. Voiced regions are thus detected in [25] as regions in the speech signal with low values for GCI drift and jitter. The method in [34] is a frame based approach for the V/NV detection where the speech signal is divided into 20 ms frames at the rate of 100 frames/s and for each frame the ZFR filtered signal energy is compared against a threshold to decide in favor of voiced or non-voiced frame. The method does not present an automatic way to determine the threshold value to be used on the energy of the ZFR filtered signal. The threshold value where equal error percentages are obtained for the false and missed detection was chosen to evaluate the performance of the method. The percentage detection accuracy obtained by Wavesurfer, the method based on robustness of epochs and the method based on the ZFR filtered signal energy on clean speech signals from the set are 95.9%, 97.1% and 94.8% respectively. Table 2.2 shows the comparison of

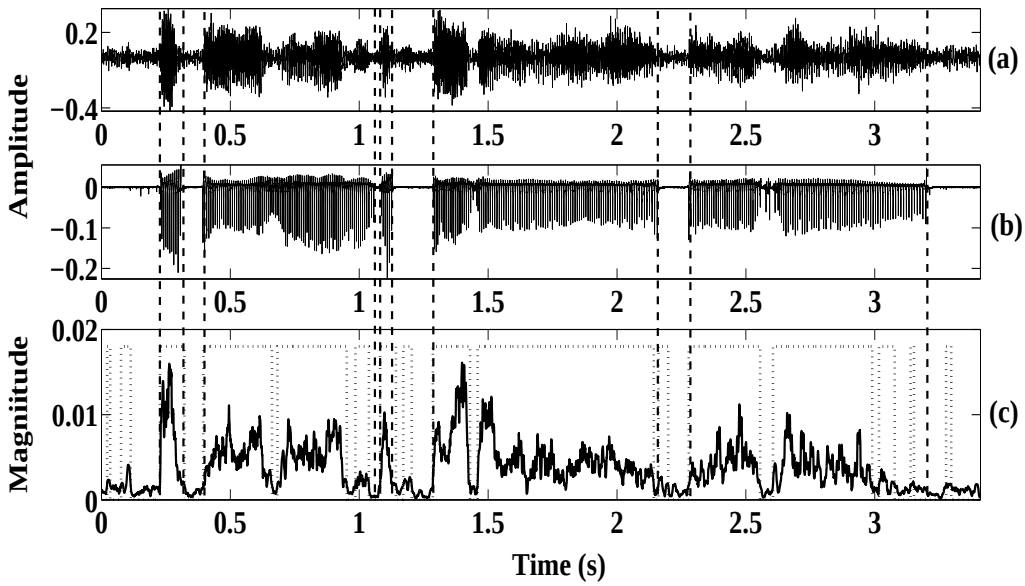

Figure 2.12: (a) Male speech signal at 5 dB SNR (vehicular noise) (b) DEGG signal (c) Smoothed MEDT over the LFR using the PWVD technique.

performance of the proposed method with three other methods for the V/NV detection in various noise environments. Please note that in Table 2.2, NA implies that the method is not able to efficiently perform the V/NV detection at the specified SNR in the given noise environment. It can be inferred from the results displayed in Table 2.2; that the proposed method has outperformed Wavesurfer and the method based on the ZFR filtered signal energy, achieving a significant performance improvement in the V/NV detection accuracy in all noise scenarios. The proposed method has provided improvement in the V/NV detection accuracy over the method based on the robustness of GCIs in the white noise environment. A marginal enhancement in the V/NV detection accuracy has been obtained by the proposed method over the method based on the robustness of GCIs in babble and vehicular noise environments. It is noted that the performance of the proposed method worsens in vehicular and babble noise environments as compared to the white noise environment because babble and vehicular noise environments contain more energy in the LFR than the white noise environment.

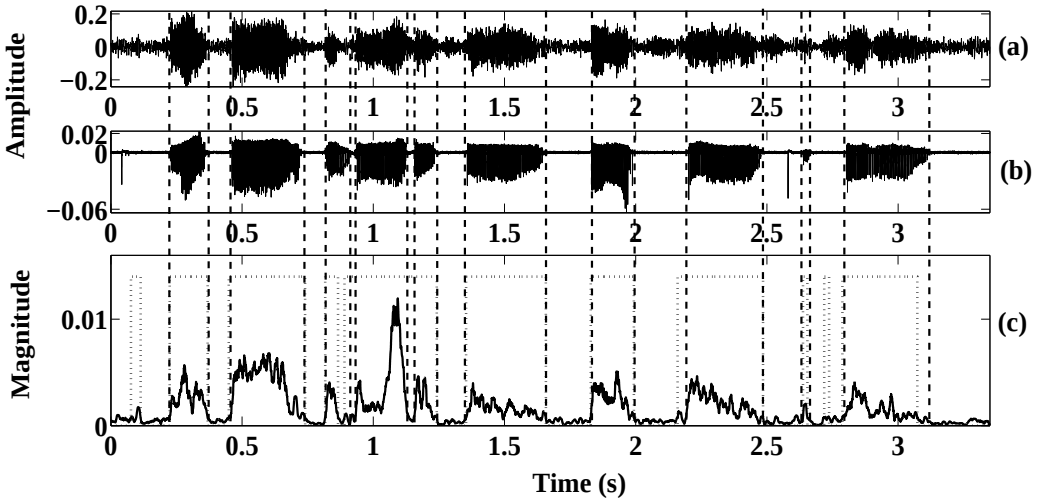

Figure 2.13: (a) Female speech signal at 5 dB SNR (vehicular noise) (b) DEGG signal (c) Smoothed MEDT over the LFR using the PWVD technique.

2.5 Conclusion

The chapter describes the proposed instantaneous V/NV detection method based on the time-frequency analysis of the speech signal over the low frequency range (LFR) using the PWVD. The PWVD technique has offered good time-resolution in the LFR and cross-terms introduced by the PWVD between the concurrent auto-components during voiced regions of a speech signal have facilitated the V/NV detection by increasing the energy during voiced regions of the speech signal. As only voiced regions of the speech signal contain significant energy in the LFR, the analysis of the speech signal in the LFR has been proven to be efficient for providing reliable discrimination between voiced and non-voiced regions in speech signals. The analysis of the speech signal in the LFR also allows the removal of the high frequency noise components. The proposed method has provided significantly better results than existing methods in terms of the V/NV detection accuracy in different noise environments. One of the main advantage of the proposed method as compared to earlier frame based methods for the V/NV detection, lies in the use of the MEDT over the LFR as a feature which has allowed instantaneous detection of voiced regions. The proposed method does not require knowledge of pitch frequency or GCIs of the speech signal in advance.

Table 2.2: Comparison of performance of the proposed V/NV detection method with existing methods in different noise environments

Method	Noise Environment	P_d (P_m, P_f)				
	SNR	20 dB	10 dB	5 dB	0 dB	-5 dB
Proposed method	White	98.2 (0.21, 1.59)	97.5 (0.35, 2.15)	97.1 (0.54, 2.36)	96.8 (0.67, 2.53)	94.9 (1.92, 3.18)
	Babble	97.5 (0.54, 1.96)	96.5 (0.84, 2.66)	95.0 (2.12, 2.88)	92.8 (3.77, 3.43)	NA
	Vehicular	97.2 (0.62, 2.18)	96.1 (1.07, 2.83)	94.3 (2.41, 3.29)	91.4 (4.54, 4.06)	NA
Wavesurfer	White	95.3 (1.30, 3.40)	92.8 (2.36, 4.84)	89.1 (4.27, 6.63)	84.8 (7.08, 8.12)	NA
	Babble	94.9 (1.33, 3.77)	92.0 (2.75, 5.25)	88.3 (4.94, 6.76)	85.2 (6.64, 8.16)	NA
	Vehicular	95.1 (1.37, 3.53)	91.8 (2.52, 5.68)	88.6 (4.57, 6.83)	85.7 (6.16, 8.14)	NA
Method based on the robustness of GCIs	White	96.3 (1.24, 2.46)	95.5 (1.84, 2.66)	94.6 (2.49, 2.91)	90.7 (4.45, 4.85)	NA
	Babble	95.9 (1.51, 2.59)	94.3 (2.25, 3.45)	92.2 (3.28, 4.52)	89.2 (5.16, 5.64)	NA
	Vehicular	95.8 (1.72, 2.48)	94.5 (2.19, 3.31)	92.3 (3.34, 4.36)	88.9 (5.40, 5.70)	NA
Method based on the ZFR filtered signal energy	White	94.2 (2.90, 2.90)	93.2 (3.41, 3.41)	91.9 (4.07, 4.07)	89.6 (5.18, 5.18)	NA
	Babble	93.6 (3.20, 3.20)	91.5 (4.26, 4.26)	87.5 (6.25, 6.25)	81.4 (9.31, 9.31)	NA
	Vehicular	93.9 (3.04, 3.04)	92.1 (3.95, 3.95)	88.4 (5.82, 5.82)	83.1 (8.43, 8.43)	NA

In order to deal with the stationary noise environment, the silence duration has been appended at the beginning of the speech signal to estimate the noise floor. It should be noted that in practical noisy scenarios, the silence duration containing only the background noise must be recorded by the system before recording the speech signal implementing the proposed method. The recording conditions of the speech signal for applications like diagnosis of pathological disorders, speaker verification and emotion recognition can be controlled to maintain a stationary noise environment.

Chapter 3

Extraction of the Time-Varying F_0 Component of a Voiced Speech Signal

This chapter presents derivation of the AM-FM signal model of the voiced speech signal in the low frequency range (LFR) which indicates the presence of energy only around the time-varying fundamental frequency (F_0) and its harmonics. The conditions on the Hankel matrix size to reliably and accurately extract harmonically related constant amplitude/frequency components contained in a multi-component signal by performing repeated EVD of the Hankel matrix are derived. The Hankel matrix is initially constructed from the samples of the multi-component signal. The theory and concepts developed for the extraction of harmonically related components of a multi-component stationary signal are extended and a noise resilient iterative algorithm is proposed in this chapter to reliably extract the time-varying F_0 component from the LFR filtered noisy voiced segment. The Hankel matrix is initially constructed from the samples of the LFR filtered noisy voiced segment. A *Distance Metric* based criterion is employed in the iterative algorithm to reliably select the dominant frequency suitable for estimating the F_0 range of a noisy voiced speech segment in an *Iteration*. A *Mono-component Signal Criteria* is introduced to ensure that the contamination of higher harmonic components of F_0 and noise is considerably reduced in the extracted time-varying F_0 component of the noisy voiced speech segment. The experiments are performed on a synthetic multi-component stationary signals, a synthetic multi-component non-stationary signal and a LFR filtered voiced speech signal in clean and noisy environments to demonstrate the efficacy of the proposed iterative algorithm.

3.1 Introduction

Voiced speech is produced when the vocal folds vibrate and chops the air flow from the lungs in a quasi-periodic manner. The vocal tract system act as a time-varying filter to the quasi-periodic excitation. It spectrally shapes the quasi-periodic excitation in a time-varying manner. Voiced speech signal is a quasi-periodic waveform. The rate of vibration of the vocal folds is comprehended as the fundamental frequency (F_0) of the voiced speech signal. F_0 is a time-varying quantity and varies with gender, age, health condition, accent, emotional condition of the speaker. The prosodic features namely: intonation, stress and rhythm are incorporated in the speech signal by varying F_0 [1]. F_0 varies in the range of 50 Hz - 250 Hz and 150 Hz - 500 Hz for adult males and adult females respectively. Children can have F_0 values as high as 500 Hz [1].

It was demonstrated in [39] that negative cycles of the time-varying F_0 component provide reliable coarse estimate of the intervals where GCIs are likely to occur. Therefore, reliable extraction of the time-varying F_0 component from noise deteriorated voiced speech signal can aid in accurately locating GCIs. However, extraction of the time-varying F_0 component of a noisy voiced speech signal is a challenging task due to high F_0 variations (as high as 80%) observed across the entire course of some voiced regions, substantially lower energy of the time-varying F_0 component than the formant components and distortion caused by noise. The extraction of the time-varying F_0 component from a clean speech signal was accomplished using a bank of two band pass filters and the time-order representation (TOR) in [26] and [66] respectively. However, both methods are sensitive to noise. The harmonic superposition method presented in [39] extracted harmonic components of a voiced speech signal occurring below the frequency of 1 kHz using the discrete Fourier transform (DFT) of a windowed voiced speech signal. However, the method in [39] did not consider the extraction of the time-varying F_0 component and its harmonics at moderate to low SNRs in noisy environments. The use of an adaptive resonance filter suggested in [75] requires tracking of bandwidth, center frequency of the time-varying F_0 component in the presence of noise and entails the use of an adaptive all-zero filter to attenuate leakages of noise and other harmonics present in the LFR filtered noisy voiced

speech signal. These limitations act as a strong motivation to develop a robust algorithm to extract the time-varying F_0 component of voiced speech signal without the need to design filters. This chapter, presents a robust algorithm for extraction of the time-varying F_0 component of a voiced speech signal. The proposed iterative algorithm does not require designing of filters and relies on the extraction of mono-component signals contained in a voiced speech signal using EVD of the Hankel matrix, initially constructed from the samples of the LFR filtered voiced speech signal. The filtering of voiced speech signal in the LFR, neglecting the non-significant eigenvalue pairs, the use of *Distance Metric* while determining the F_0 range of voiced speech signal account for the noise resilience of the proposed iterative algorithm.

This chapter is organized as follows: Section 3.2 presents the derivation of the AM-FM signal model of the voiced speech signal in the LFR. In Section 3.3, the conditions on the Hankel matrix size for accurate extraction of harmonically related constant amplitude/frequency components contained in a multi-component signal using EVD of Hankel matrix are derived. Section 3.4 presents the proposed iterative algorithm for extraction of the time-varying F_0 component of voiced speech signal. Section 3.5 demonstrates experimental results of the proposed iterative algorithm on a synthetic multi-component non-stationary signal and a voiced speech signal. Section 3.6 concludes the chapter.

3.2 AM-FM Model of the Voiced Speech Signal in the LFR

From (1.2), it has been learnt that the z-transform of the voiced speech signal, $s[n]$ can be modeled as:

$$S[z] = \Theta_0 V[z] H[z] \tag{3.1}$$

where Θ_0 controls the overall amplitude during the voiced speech production. $S[z]$ and $V[z]$ denote the z-transforms of $s[n]$ and the first-order derivative of the glottal flow respectively. The z-domain transfer function of the vocal tract system is represented by $H[z]$. The assumption of the speech production system to be stationary is made in (3.1).

The differentiated glottal flow $v[n] = u'[n]$ act as an excitation to the vocal tract system. The excitation $u'[n]$ is characterized by a large negative impulse-like pulse during the return phase of each glottal cycle [76]. In order to facilitate the derivation of the AM-FM signal model of the voiced speech signal in the LFR, we start with approximating $u'[n]$ by a periodic train of negative impulses $\hat{u}[n]$ with period of N_0 samples as [1]:

$$\hat{u}[n] = -D \sum_{i=-\infty}^{\infty} \delta[n - iN_0] \tag{3.2}$$

The fundamental period of $\hat{u}[n]$ is represented by N_0. The strength of each impulse is denoted by D and δ denotes the unit sample sequence which is defined in [62]. As $\hat{u}[n]$ is real and even, its discrete-time Fourier series expansion [62] is:

$$\hat{u}[n] = -\frac{D}{N_0} \sum_{k=0}^{N_0-1} \cos(2\pi k f_0 n) \quad \text{where} \quad f_0 = \frac{1}{N_0} = \frac{F_0}{F_s} \tag{3.3}$$

where the fundamental frequency in Hz, sampling frequency in Hz and normalized fundamental frequency are denoted by F_0, F_s and f_0 respectively. In reality, the vocal tract system act as a time-varying filter which changes the amplitude and phase of each harmonic component in the excitation in a time-varying manner [1]. Thus, by removing the constraint of the vocal tract system to be stationary, the voiced speech signal $s[n]$ can be approximated as:

$$\hat{s}[n] = -\frac{D}{N_0} \sum_{k=0}^{N_0-1} |H[\omega_k, n]| \cos(\omega_k n + \text{angle}(H[\omega_k, n])) \tag{3.4}$$

where $\omega_k = 2\pi k f_0$ represents the normalized angular frequency of k^{th} harmonic component of the normalized fundamental frequency f_0. The time-varying magnitude and phase spectrum of the vocal tract system are denoted by $|H[\omega, n]|$ and $\text{angle}(H[\omega, n])$ respectively. In reality, the interval between successive negative impulses in the differentiated glottal flow $v'[n]$ is not constant and varies with time, resulting in the F_0 to be time-varying in nature. Therefore, if we relax the constraint of constant fundamental

frequency in (3.4), we get:

$$\hat{s}[n] = \sum_{k=0}^{N_0-1} A_k[n] \cos\left(2\pi k f_0[n] n + \theta_k[n]\right), \quad f_0[n] = \frac{F_0[n]}{F_s} \tag{3.5}$$

where $f_0[n]$, $F_0[n]$, F_s, $A_k[n]$ and $\theta_k[n]$ denote the time-varying normalized fundamental frequency, time-varying fundamental frequency in Hz, sampling frequency in Hz, the time-varying amplitude and phase of k^{th} harmonic component of $f_0[n]$ respectively. Hence, the voiced speech signal in the LFR (50 Hz - 500 Hz), denoted $s_{\text{LF}}[n]$ can be approximated using the AM-FM signal model as follows:

$$\hat{s}_{\text{LF}}[n] = \sum_{k=1}^{K} A_k[n] \cos\left(2\pi k f_0[n] n + \theta_k[n]\right) \tag{3.6}$$

where K denotes the number of harmonic components of $f_0[n]$ present in the LFR. It can be inferred from (3.6) that voiced speech signal in the LFR is a multi-component signal with the presence of energy only around $F_0[n]$ and its few harmonics. The spectrogram of the voiced speech signal (Fig. 3.1 (a)) over the LFR is shown in Fig. 3.1 (b) which shows the presence of energy only at around the $F_0[n]$ (175 Hz) and its second and third harmonics in accordance with the derived AM-FM signal model of the voiced speech signal in the LFR given by (3.6).

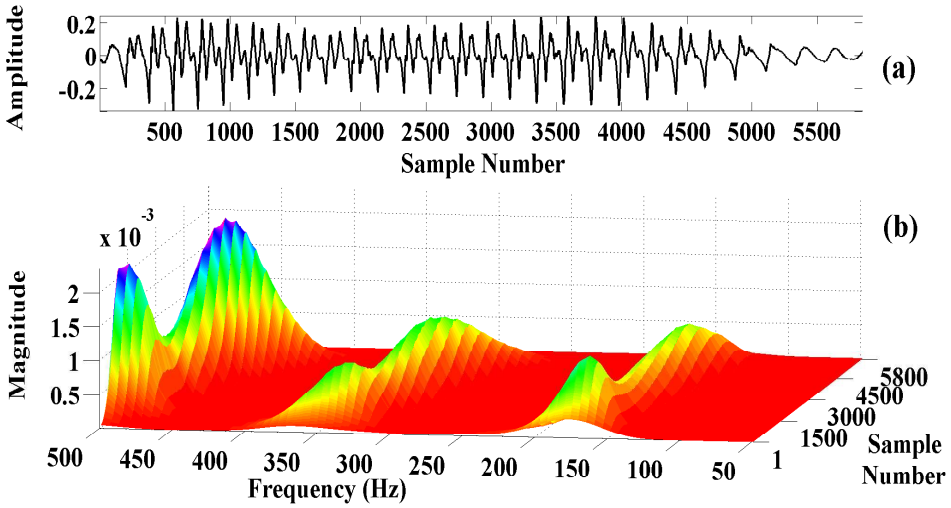

Figure 3.1: (a) Clean voiced speech signal (b) Spectrogram over the LFR.

The sinusoidal functions are eigenfunctions of a linear time-invariant (LTI) system [1]. The Hankel matrix offers interesting properties which are described in the next section. It motivated us to derive the conditions on the Hankel matrix size for accurate and reliable extraction of harmonically related constant amplitude/frequency components contained in a multi-component signal using repeated eigenvalue decomposition (EVD) of the Hankel matrix. The Hankel matrix is initially constructed from the samples of the multi-component signal. The theory and concepts developed in the next section are extended in Section 3.4 to develop a robust iterative algorithm for extraction of the time-varying F_0 component of a noisy voiced speech signal.

3.3 Extraction of Constant Amplitude/Frequency Harmonically Related Components using Eigenvalue Decomposition of the Hankel Matrix

The square Hankel matrix of size $N \times N$, H_N^y can be constructed from a real signal $y[n]$ spanning $(0, 1, ..., Q-1)$ samples, as follows [77]:

$$H_N^y = \begin{bmatrix} y[0] & y[1] & . & . & . & y[N-1] \\ y[1] & y[2] & . & . & . & y[N] \\ . & . & . & . & . & . \\ y[N-1] & y[N] & . & . & . & y[2N-2] \end{bmatrix} \tag{3.7}$$

We assume that $Q \geq 2N-1$ and N is an even number throughout the rest of this thesis. The Hankel matrix offers interesting properties because of its peculiar structure. The square Hankel matrix constructed from a real signal is a symmetric matrix, i.e. $H_N^y = (H_N^y)^T$, where T denotes the transpose operator. The EVD of the square matrix H_N^y can be expressed as [77]:

$$H_N^y = U_y \Lambda_y U_y^T \tag{3.8}$$

where Λ_y is a diagonal matrix with real scalar eigenvalues $\lambda_{y,i}$ and U_y is an orthogonal matrix, consisting of real eigenvectors, $\vec{u}_{y,i}, i = 1, 2, ..., N$ as its columns, each column

consisting of N elements. Any two eigenvectors, $\vec{u}_{y,i}$ and $\vec{u}_{y,j}$ corresponding to different eigenvalues are orthogonal [77].

Let $y[n]$ be the sum of K harmonically related constant amplitude/frequency mono-component signals as follows:

$$y[n] = \sum_{k=1}^{K} y_k[n] = \sum_{k=1}^{K} A_k \cos(2\pi k f_0 n + \theta_k),\ n = 0, 1, ..., Q-1 \tag{3.9}$$

such that $A_k \neq A_l$ for $k \neq l$, where $k, l = 1, 2, ..., K$. The normalized fundamental frequency is represented by f_0, where $f_0 = \dfrac{F_0}{F_s}$. The amplitude and phase of the k^{th} harmonic component of f_0 are denoted by A_k and θ_k respectively. It is assumed that $F_s > 2KF_0$ to avoid aliasing. Using (3.7) and (3.9), the Hankel matrix of $y[n]$, H_N^y can be expressed as sum of Hankel matrices of its mono-component signals $H_N^{y_k}$ as:

$$H_N^y = \sum_{k=1}^{K} H_N^{y_k} \qquad \text{where} \quad H_N^{y_k} = (H_N^{y_k})^T \tag{3.10}$$

The characteristic equation of H_N^y is given by [77]:

$$\det(H_N^y - \lambda I) = \lambda^N - \mathrm{Tr}(H_N^y)\lambda^{N-1} + ... + \det(H_N^y) = 0 \tag{3.11}$$

where Tr(.) and det(.) denote the trace and determinant of the matrix respectively. The number of sinusoids contained in H_N^y and $H_N^{y_k}$ are K and 1 respectively; therefore, irrespective of the value of N, the ranks and non-zero eigenvalues of H_N^y and $H_N^{y_k}$ cannot be greater than $2K$ and 2 respectively. Thus, the characteristic equation of $H_N^{y_k}$ can be written as follows:

$$|H_N^{y_k} - \lambda I| = \lambda^{N-2}(\lambda^2 - Tr(H_N^{y_k})\lambda + \kappa_{N-2}) = 0 \tag{3.12}$$

where κ_{N-2} denotes the coefficient associated with λ^{N-2}. It is evident from (3.12) that when $Tr(H_N^{y_k})$ becomes zero, the two real eigenvalues of $H_N^{y_k}$ become equal and opposite in sign (EOS). The trace of the square matrix can be expressed in terms of its eigenvalues

as follows [77]:

$$\begin{aligned} \mathrm{Tr}(H_N^y) &= \sum_{i=1}^{N} \lambda_{y,i} \\ \mathrm{Tr}(H_N^{y_k}) &= \sum_{i=1}^{2} \lambda_{y_k,i} \qquad , N \geq 2 \end{aligned} \tag{3.13}$$

We now consider two different cases based on the value of N.

3.3.1 Case (i): when Hankel matrix size is an integer multiple of the fundamental period

In this case, $N = \dfrac{\sigma}{f_0} = \sigma N_0$, where σ is a positive integer. Using (3.7), (3.9) and (3.10), $\mathrm{Tr}(H_{\sigma N_0}^{y_k})$ and $\mathrm{Tr}(H_{\sigma N_0}^{y})$ are given by:

$$\begin{aligned} \mathrm{Tr}(H_{\sigma N_0}^{y_k}) &= A_k \sum_{n=0}^{\sigma N_0 - 1} \cos(2\pi k f_0 2n + \theta_k) \\ &= A_k \, \Re\left(e^{j\theta_k} \sum_{n=0}^{\sigma N_0 - 1} e^{j2\pi k f_0 2n} \right) \\ &= \quad 0 \quad \forall \; k \\ \mathrm{Tr}(H_{\sigma N_0}^{y}) &= \sum_{k=1}^{K} \mathrm{Tr}(H_{\sigma N_0}^{y_k}) = 0 \end{aligned} \tag{3.14}$$

The inner product of i^{th} row/column of $H_{\sigma N_0}^{y_k}$ and j^{th} row/ column of $H_{\sigma N_0}^{y_l}$ denoted by $\left\langle H_{\sigma N_0}^{y_k}, H_{\sigma N_0}^{y_l} \right\rangle_{i,j}$ is given by:

$$\begin{aligned} &\left\langle H_{\sigma N_0}^{y_k}, H_{\sigma N_0}^{y_l} \right\rangle_{i,j} \\ &= A_k A_l \sum_{n=0}^{\sigma N_0 - 1} \Big(\cos(2\pi k f_0 (n+i-1) + \theta_k) \times \cos(2\pi l f_0 (n+j-1) + \theta_l) \Big) \\ &= \frac{A_k A_l}{2} \, \Re\Bigg(e^{j(2\pi f_0 m_1 + \theta_k + \theta_l)} \sum_{n=0}^{\sigma N_0 - 1} e^{j2\pi (k+l) f_0 n} + e^{j(2\pi f_0 m_2 + \theta_k - \theta_l)} \sum_{n=0}^{\sigma N_0 - 1} e^{j2\pi (k-l) f_0 n} \Bigg) \\ &= 0, \qquad\qquad i,j = 1,2,...,\sigma N_0 \;\text{ and }\; k \neq l \end{aligned} \tag{3.15}$$

where $m_1 = k(i-1) + l(j-1)$ and $m_2 = k(i-1) - l(j-1)$. It can be deduced from (3.15) that rows and columns of $H_{\sigma N_0}^{y_k}$ and $H_{\sigma N_0}^{y_l}$ for $k \neq l$ are orthogonal to each other. In such

scenario, the $2K$ non-zero eigenvalues and corresponding eigenvectors of $H^{y}_{\sigma N_0}$ are equal to the set consisting of eigenvalues and corresponding eigenvectors of $H^{y_k}_{\sigma N_0}$ as follows:

$$\begin{aligned} \lambda_{y,(2k+j-2)} &= \lambda_{y_k,j} \\ u_{y,(2k+j-2)} &= u_{y_k,j}, \qquad k = 1, 2, ..., K,\ j = 1, 2 \end{aligned} \tag{3.16}$$

Moreover, using (3.13) and (3.14), it can be deduced that:

$$\lambda_{y_k,1} = -\lambda_{y_k,2} \tag{3.17}$$

It can be inferred from (3.16) and (3.17) that the k^{th} mono-component signal of $y[n]$ can be extracted by creating a modified eigenvalue diagonal matrix $\tilde{\Lambda}_{y_k}$ which preserves only the k^{th} non-zero eigenvalue pair of Λ_y as follows:

$$\tilde{\Lambda}_{y_k} = \text{diag}(0, ..., 0, \lambda_{y_k,1}, -\lambda_{y_k,1}, 0, ..., 0) \tag{3.18}$$

where diag(.) denotes diagonal matrix. Construct $\tilde{H}^{y_k}_N$ as follows:

$$\tilde{H}^{y_k}_N = U_y \tilde{\Lambda}_{y_k} U_y^T \tag{3.19}$$

where $N = \sigma N_0$, $\tilde{H}^{y_k}_N = H^{y_k}_N$. The average of skew diagonal elements of $\tilde{H}^{y_k}_N$ gives $\tilde{y}_k[n]$, where $\tilde{y}_k[n] = y_k[n]$ for $N = \sigma N_0$. Here's an example:

Example 1: $y[n] = \sum_{k=1}^{3} y_k[n] = \sum_{k=1}^{3} A_k \cos\left(\frac{2\pi k 100 n}{32000}\right)$, n = 0, 1, ..., 638

where $A_1 = 2$, $A_2 = 3$, $A_3 = 1$. The value of N is chosen to be equal to $N_0 = 320$. The non-zero eigenvalue pairs corresponding to the three mono-component signals of $y[n]$ contained in H^{y}_{320} found using MATLAB are {(320, -320), (480, -480), (160, -160)}. Please note that the value of $\lambda_{y_k,1} = \frac{NA_k}{2}$ $\forall k$ is directly proportional to the amplitude of the mono-component signal. The signal $y[n]$ and its extracted mono-component signals using (3.7), (3.8), (3.18) and (3.19) are shown in Fig. 3.2.

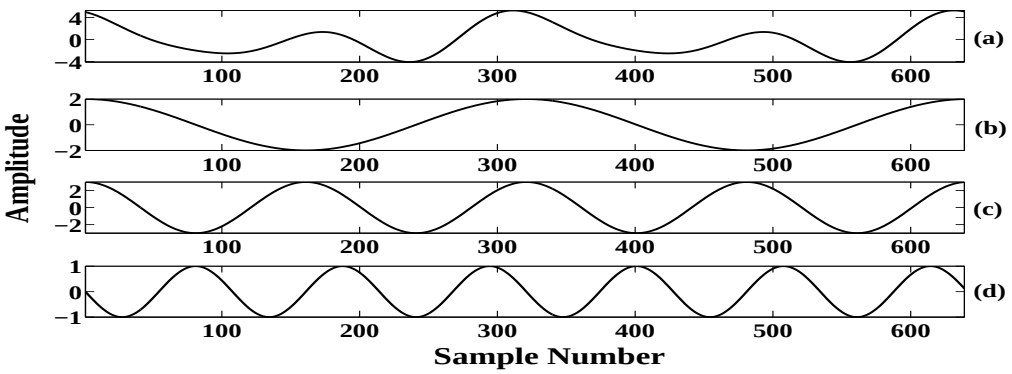

Figure 3.2: (a) Multi-component signal $y[n]$ (b) Mono-component signal $y_1[n]$ (c) Mono-component signal $y_2[n]$ (d) Mono-component signal $y_3[n]$. $N = N_0 = 320$.

3.3.2 Case (ii): when Hankel matrix size is not an integer multiple of the fundamental period

In such cases, $N \neq \sigma N_0$ and the relations in (3.14), (3.15), (3.16), (3.17) no longer hold. The value of $\langle H_N^{y_k}, H_N^{y_l} \rangle_{i,j} \neq 0$ for some values of i, j, where $i, j = 1, 2, ..., N$. Let the eigenvalues of H_N^y be now arranged in ascending order, i.e. $\lambda_{y,i+1} \geq \lambda_{y,i}$, $i = 1, 2, ..., N-1$. The modified eigenvalue diagonal matrix preserving the k^{th} eigenvalue pair denoted by $\tilde{\Lambda}_{y_k}$ is now given by:

$$\tilde{\Lambda}_{y_k} = \text{diag}(0, ..., 0, \lambda_{y,k}, 0, ..., 0, \lambda_{y,N-k+1}, 0, ..., 0) \tag{3.20}$$

The k^{th} mono-component signal of $y[n]$ corresponding to the k^{th} eigenvalue pair of H_N^y can be extracted by substituting $\tilde{\Lambda}_{y_k}$ from (3.20) in (3.19), where $N \neq \sigma N_0$ in this case. Let the k^{th} original and extracted mono-component signals of $y[n]$ be denoted by $y_k[n]$ and $\tilde{y}_k[n]$ respectively. Consider the same signal $y[n]$ as given in the Example 1, but now spanning $(0, 2N-2)$ samples. The non-zero eigenvalue pairs of H_N^y for different values of Hankel matrix size N are compiled in Table 3.1. The original and extracted mono-component signals of $y[n]$ obtained using (3.7), (3.8), (3.19) and (3.20) are depicted in Fig. 3.3, Fig. 3.4 and Fig. 3.5 for three different values of N in solid and dashed lines respectively. It can be inferred from Fig. 3.3, Fig. 3.4 and Fig. 3.5 that $\tilde{y}_k[n] \neq y_k[n]$ but $\tilde{y}_k[n]$ matches closely with $y_k[n]$ when $N > N_0$. The mean square error (MSE) between

$\tilde{y}_k[n]$ and $y_k[n]$, $\text{MSE}_N^{y_k}$ is calculated as:

$$\text{MSE}_N^{y_k} = \frac{1}{2N-1} \sum_{n=0}^{2N-2} (y_k[n] - \tilde{y}_k[n])^2 \tag{3.21}$$

The values of $\text{MSE}_N^{y_k}, k = 1, 2, 3$ are compiled in Table 3.1. It can be deduced that the MSE decreases as N increases, $\forall k$.

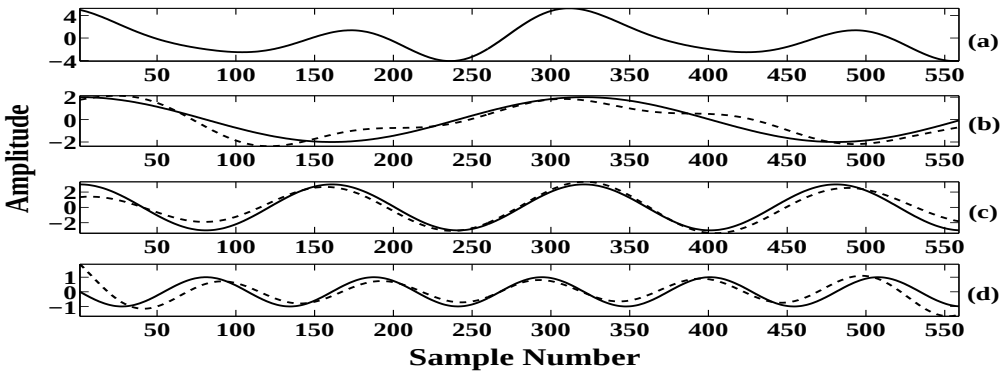

Figure 3.3: (a) Multi-component signal $y[n]$ (b) Mono-component signals $y_1[n]$ and $\tilde{y}_1[n]$ (c) Mono-component signals $y_2[n]$ and $\tilde{y}_2[n]$ (d) Mono-component signals $y_3[n]$ and $\tilde{y}_3[n]$. $N = 280$.

In practice the exact value of N_0 is not known in advance and therefore, $N \neq \sigma N_0$. We now comprehensively study the variation of error to signal ratio between the extracted and original components of a multi-component signal with respect to variation in the Hankel matrix size N. As an example, let $y[n]$ be a multi-component signal containing five constant amplitude/frequency harmonically related components as follows:

$$y[n] = \sum_{k=1}^{K} y_k[n] = \sum_{k=1}^{K} A_k \cos\left(\frac{2\pi kn}{N_0} + \theta_k\right), \quad n = 0, 1, ..., 2N-2 \tag{3.22}$$

where $A_k \neq A_l$ for $k \neq l; k, l = 1, 2, ..., K$, $F_s = 20$ kHz, $F_0 = 70$ Hz, $N_0 = F_s/F_0$, $K = 5$,

Table 3.1: Non-zero eigenvalue pairs for different values of N

N	Eigenvalue Pairs	MSE
280	(250.2, -294.9), (440.9, -421.8), (108.6, -137.9)	0.46, 0.31, 0.24
500	(490.8, -475.3), (780.3, -718.5), (227.9, -243.4)	0.14, 0.08, 0.04
1000	(1016.9, -972.3), (1507.3, -1498.9), (494.9, -501.9)	0.02, 0.02, 0.00

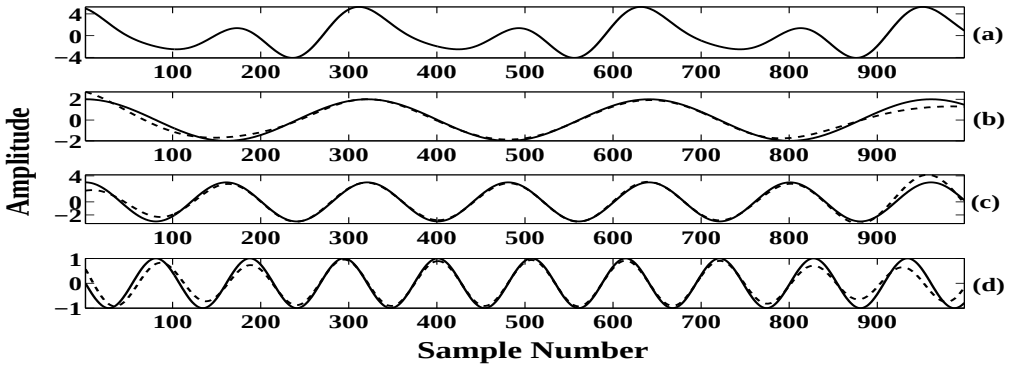

Figure 3.4: (a) Multi-component signal $y[n]$ (b) Mono-component signals $y_1[n]$ and $\tilde{y}_1[n]$ (c) Mono-component signals $y_2[n]$ and $\tilde{y}_2[n]$ (d) Mono-component signals $y_3[n]$ and $\tilde{y}_3[n]$. $N = 500$.

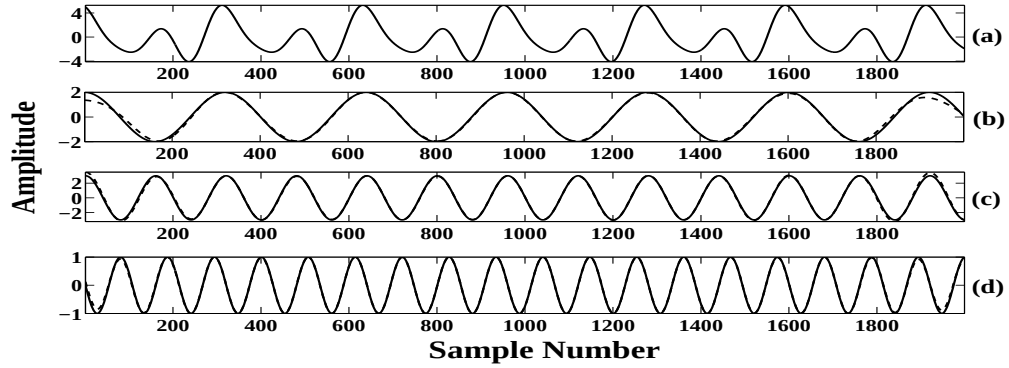

Figure 3.5: (a) Multi-component signal $y[n]$ (b) Mono-component signals $y_1[n]$ and $\tilde{y}_1[n]$ (c) Mono-component signals $y_2[n]$ and $\tilde{y}_2[n]$ (d) Mono-component signals $y_3[n]$ and $\tilde{y}_3[n]$. $N = 1000$.

$A_1 = 2$, $A_2 = 1.8$, $A_3 = 3$, $A_4 = 1.0$, $A_5 = 0.8$, $\theta_1 = \theta_5 = 0$, $\theta_2 = \pi/3$, $\theta_3 = \pi/4$, $\theta_4 = \pi/5$. Let the eigenvalues obtained by performing EVD of H_N^y be arranged in ascending order. It can be inferred from (3.19) and (3.20) that the Hankel matrix formed by preserving the p^{th} non-zero eigenvalue pair of H_N^y is given by:

$$\tilde{H}_N^{y_p} = \lambda_p \vec{u}_{y,p} \vec{u}_{y,p}^T + \lambda_{N-p+1} \vec{u}_{y,N-p+1} \vec{u}_{y,N-p+1}^T \tag{3.23}$$

where p takes values from 1, 2, ..., K. The p^{th} extracted component of $y[n]$ denoted by $\tilde{y}_p[n]$ is computed by taking the average of the skew diagonal elements of $\tilde{H}_N^{y_p}$ [4]. We define the error to signal ratio between the extracted and the original p^{th} mono-component

signal of $y[n]$, denoted by ESR_N^p, as:

$$ESR_N^p = \frac{\sum_{n=0}^{2N-2} (y_p[n] - \tilde{y}_p[n])^2}{\sum_{n=0}^{2N-2} (y_p[n])^2}, \quad p = 1, 2, ..., K \tag{3.24}$$

It is not feasible to analytically determine the variation of ESR_N^p with respect to the Hankel matrix size N because ESR_N^p is a function of $\tilde{y}_p[n]$, and it can be inferred from (3.23) that $\tilde{y}_p[n]$ is a function of the eigenvectors corresponding to the p^{th} eigenvalue pair of H_N^y. The derivation of analytical expressions for the eigenvectors of the Hankel matrix for any arbitrary value of N is not feasible. Therefore, we chose to carry out an empirical study of the variation of ESR_N^p with respect to N and draw meaningful inferences from it. Fig. 3.6 depicts a magnified view of the variation of ESR_N^p, $\forall p$ with respect to N. It can be inferred from Fig. 3.6 that the reduction of ESR_N^p, $\forall p$ with respect to N is not monotonic in nature but the successive maxima of ESR_N^p, $\forall p$ reduce as N increases. Please also observe in Fig. 3.6 that the values of $ESR_N^p = 0$ for $N = \sigma N_0$, where $N_0 \approx 286$ samples, $\sigma = 1, 2, 3$ and $p = 1, 2, ..., 5$, in accordance with the mathematically derived result in section 3.3.1 [4]. In order to understand the variation of ESR_N^p, $\forall p$ with respect to N, the combined magnitude spectrums of the eigenvectors corresponding to different eigenvalue pairs of H_N^y over the positive frequency range has been depicted in Fig. 3.7, Fig. 3.8, Fig. 3.9 and Fig. 3.10 for $N = 62$, 250, 420, 700 respectively. It can be observed in Fig. 3.9 and Fig. 3.10 that the combined magnitude spectrum of $\vec{w}_{x,p}$ and $\vec{w}_{x,N-p+1}$ attains peak at one of the harmonic frequencies contained in $y[n]$ for $N > N_0$, $\forall p$. It can be inferred from Fig. 3.7, Fig. 3.8, Fig. 3.9 and Fig. 3.10 that the frequency range over which the combined magnitude spectrum of $\vec{w}_{x,p}$ and $\vec{w}_{x,N-p+1}$ has significant value, gradually decreases as N increases, $\forall p$. Therefore, the extracted components $\tilde{y}_p[n]$, $\forall p$ eventually approach the original sinusoidal functions contained in $y[n]$ and $ESR_N^p \approx 0$, $\forall p$ when $N >> N_0$. However, it is evident from the comparison of Fig. 3.9 with Fig. 3.10 that the frequency range over which the combined magnitude spectrum of $\vec{u}_{y,p}$ and $\vec{u}_{y,N-p+1}$ has significant value does not reduce monotonically with

respect to N, where p takes values from 1, 2, ..., K. This accounts for the non-monotonic variation of $ESR_N^p,\ \forall p$ with respect to N, as shown in Fig. 3.1.

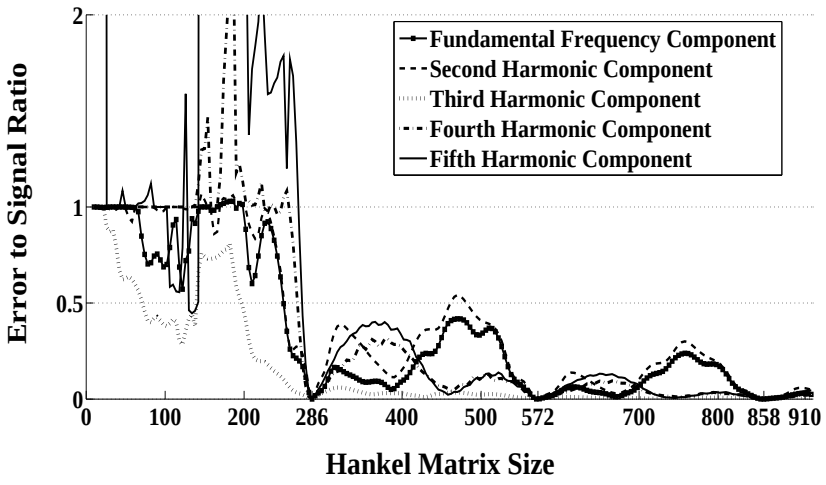

Figure 3.6: Error to signal ratio with respect to the Hankel matrix size (N) after the first *Iteration*. $N_0 \approx 286$ samples.

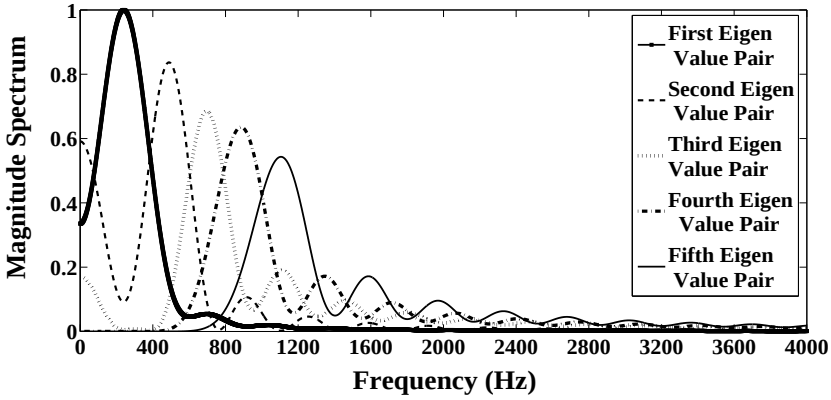

Figure 3.7: Combined magnitude spectrum of the eigenvectors corresponding to different eigenvalue pairs of H_N^x for $N = 62$ after the first *Iteration*.

It is apparent from Fig. 3.7, Fig. 3.8, Fig. 3.9 and 3.11 that the extracted components corresponding to different eigenvalue pairs of H_N^y may not be mono-component signals because the combined magnitude spectrums of the eigenvectors (acting as basis functions for the extracted components) corresponding to different eigenvalue pairs of H_N^y may have significant value around more than one harmonic frequency contained in $y[n]$. In other words, the extracted components may have significant contributions from more than one

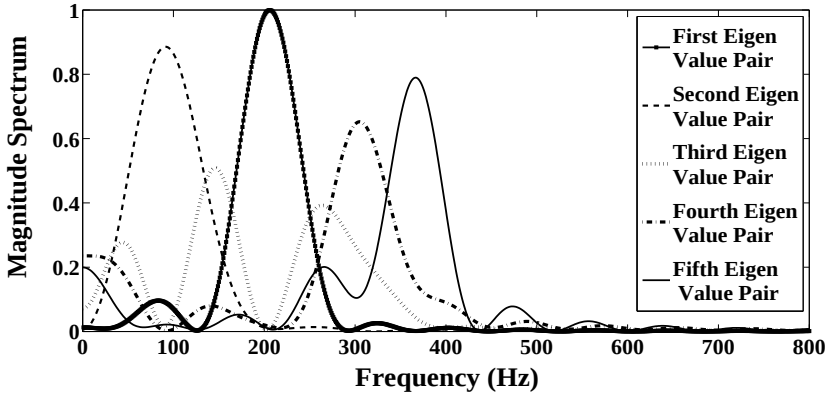

Figure 3.8: Combined magnitude spectrum of the eigenvectors corresponding to different eigenvalue pairs of H_N^x for $N = 250$ after the first *Iteration*.

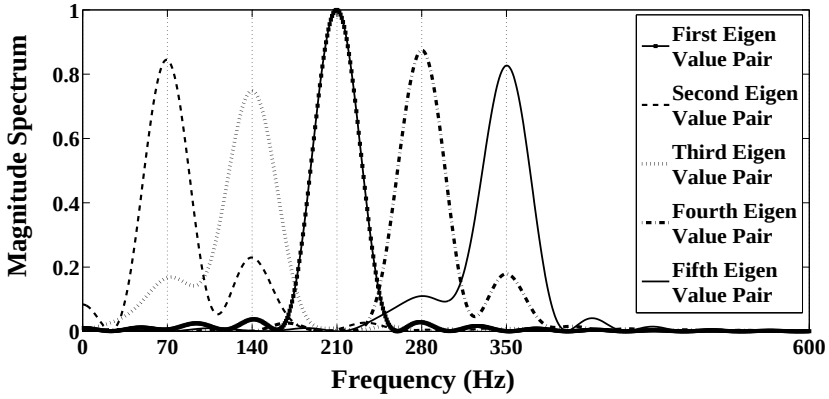

Figure 3.9: Combined magnitude spectrum of the eigenvectors corresponding to different eigenvalue pairs of H_N^x for $N = 420$ after the first *Iteration*.

harmonic components of $y[n]$. In order to determine whether an extracted component is a mono-component signal or not, we define the *Mono-component Signal Criteria* in the next subsection.

Mono-component Signal Criteria

The *Mono-component Signal Criteria* to find out whether an extracted component is a mono-component AM-FM signal or not is defined as:
(1) The magnitude of the difference between the number of zero-crossings and local extrema (minima and maxima) of the extracted component denoted by D_n is less than or equal to one.

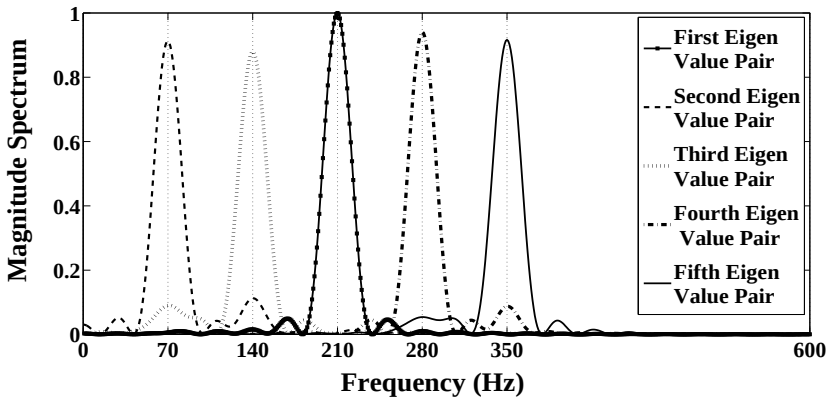

Figure 3.10: Combined magnitude spectrum of the eigenvectors corresponding to different eigenvalue pairs of H_N^x for $N = 700$ after the first *Iteration*.

(2) The number of significant eigenvalue pairs obtained by performing EVD of the Hankel matrix constructed from the samples of the extracted component denoted by D_r is equal to one.

Please note that we have considered an eigenvalue pair of the Hankel matrix to be significant if magnitude of one of the eigenvalues constituting the pair is greater than or equal to one-fourth of the maximum eigenvalue of the Hankel matrix. Please also note that the first part of the *Mono-component Signal Criteria* defined above is first stated and used in [78] to extract intrinsic mode functions (IMFs) from a multi-component signal using empirical mode decomposition (EMD).

Multiple Iterations

It is evident from Fig. 3.11 that even when $N > N_0$, the combined magnitude spectrum of $\vec{u}_{y,p}$ and $\vec{u}_{y,N-p+1}$ may have significant side lobes at harmonic frequencies contained in $y[n]$, other than the harmonic frequency at which it attains the peak value, where $p = 1, 2, ..., K$. The extracted components corresponding to such eigenvectors contain contributions from more than one harmonic component of $y[n]$ and hence, do not satisfy the *Mono-component Signal Criteria* defined in the previous subsection. In order to attenuate the side lobes, such extracted components are treated as multi-component signals for the second *Iteration* and EVD is again performed on the Hankel matrices constructed from their samples. Let's assume that the eigenvalues obtained by performing EVD are arranged in ascending

order. Only the component corresponding to the first eigenvalue pair (highest energy) is extracted in the second *Iteration*. It is apparent from the comparison of Fig. 3.11 with Fig. 3.12 that the side lobes in the combined magnitude spectrum of the eigenvectors corresponding to different eigenvalue pairs of H_N^y are attenuated after the second *Iteration*. The substantial improvement obtained in the $ESR_N^p,\ \forall p$ after the second *Iteration* is apparent from the comparison of Fig. 3.6 with Fig. 3.13. Please observe in Fig. 3.13 that the ESR_N^p decreases considerably for $N > N_0,\ \forall p$ and the value of error to signal ratio for the F_0 component of $x[n]$ is less than 0.2 (-7 dB) for all values of N. We have obtained similar improvement in the error to signal ratio for the F_0 component of $x[n]$ with respect to N, for about 300 different combinations of the values of K, A_k, θ_k and N_0 in (3.24), where we have kept $K \leq 7, 40 \leq N_0 \leq 400, A_1 \geq 0.25 \max(A_k), k = 2, 3, ..., K$ and $|A_k - A_l| \geq (0.1 \max(A_k, A_l))$ for $k = 1, 2, ..., K-1$ and $l = k+1, k+2, ..., K$. We have kept $A_1 \geq 0.25 \max(A_k), k = 2, 3, ..., K$ to ensure that the F_0 component has significant energy relative to other harmonic components of $x[n]$. The time-varying F_0 component of a voiced speech signal has substantial energy relative to other harmonic components in the LFR [4, 44].

It is also important to note here that the main aim is not to achieve a very low value of the error to signal ratio for the F_0 component of $x[n]$. The objective is that the intervals marked by the positive and negative cycles of the extracted F_0 component should match with the intervals marked by the positive and negative cycles of the original F_0 component of $y[n]$ respectively. Negative cycles of the time-varying F_0 component of a voiced speech signal provide a reliable coarse estimate of the intervals where GCIs are likely to occur. Therefore, we define a quantity, *Intervals Matching Percentage* (IMP) to objectively measure the matching of the intervals marked by the positive and negative cycles of the extracted F_0 component and the original F_0 component of $x[n]$ as follows:

$$IMP = \frac{n_{\mathrm{NC}} + n_{\mathrm{PC}}}{2N - 1} \times 100 \tag{3.25}$$

where n_{NC} and n_{PC} denote the number of samples which are common in the negative and positive cycles of the extracted F_0 component and the original F_0 component of $y[n]$

respectively. The total number of samples in $y[n]$ is $2N-1$, as given by (3.24). It can be observed from Fig. 3.14 that IMP is greater than 91% for $N > N_0$ after the second *Iteration*. We have obtained similar results on the variation of IMP with respect to N for 300 different combinations of the values of K, A_k, θ_k and N_0 in (3.24). Hence, from this empirical study, we draw a general conclusion that the error to signal ratio for the F_0 component of $x[n]$ decreases considerably and the value of IMP increases substantially for $N > N_0$ irrespective of the values of K, A_k, θ_k, N_0, when EVD of the Hankel matrix, initially constructed from the samples of $x[n]$ is performed repeatedly until the *Mono-component Signal Criteria* defined in the previous subsection is satisfied by the extracted F_0 component of $y[n]$.

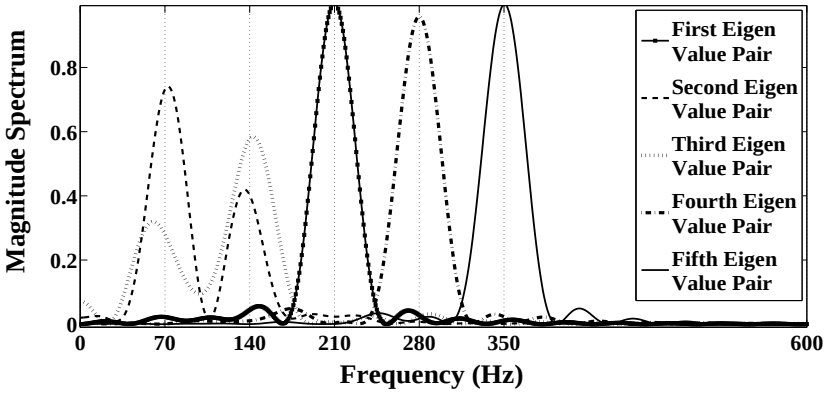

Figure 3.11: Combined magnitude spectrum of the eigenvectors corresponding to different eigenvalue pairs of H_N^x for $N = 470$ after the first *Iteration*.

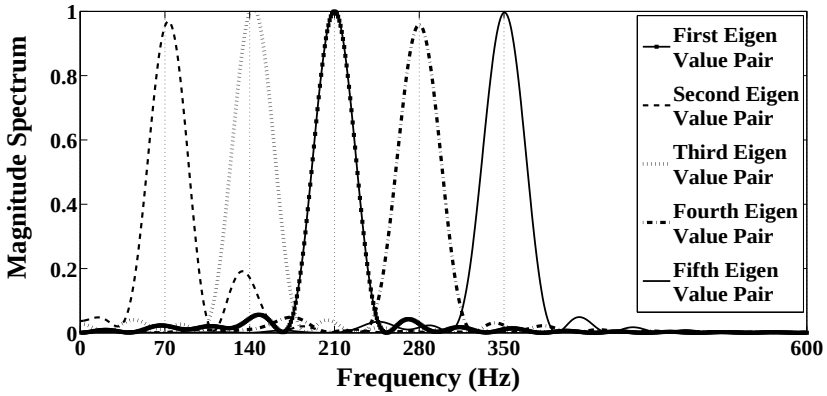

Figure 3.12: Combined magnitude spectrum of the eigenvectors corresponding to different eigenvalue pairs of H_N^x for $N = 470$ after the second *Iteration*.

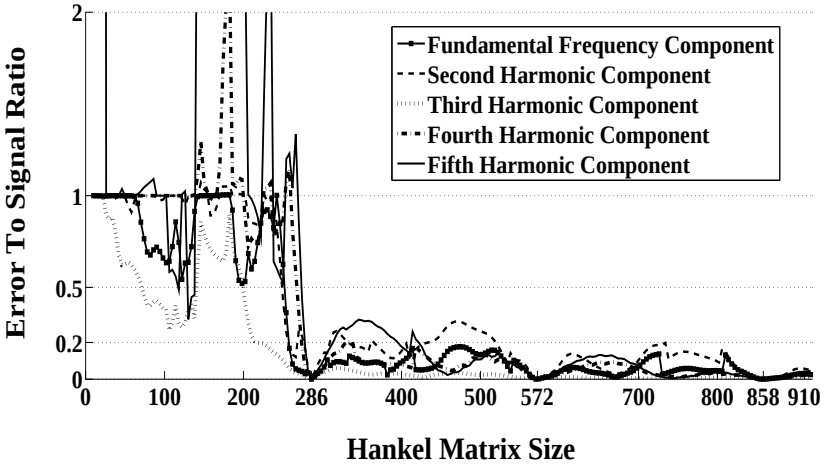

Figure 3.13: Error to signal ratio with respect to the Hankel matrix size (N) after the second *Iteration*. $N_0 \approx 286$ samples.

3.4 Extraction of the Time-varying F_0 component using Eigenvalue Decomposition of the Hankel Matrix

It is inferred from Example 1 of the previous section that the magnitude of eigenvalues constituting an eigenvalue pair corresponding to a mono-component signal contained in a multi-component signal is directly proportional to the amplitude of the mono-component signal. Here, the eigenvalues are obtained by performing EVD of the Hankel matrix constructed from the samples of a multi-component signal. The formants contain substantially higher energy than the time-varying F_0 component of the voiced speech signal [1,44]. The formants are required to be attenuated to render the magnitude of the eigenvalues constituting an eigenvalue pair corresponding to the time-varying F_0 component significant. Therefore, in order to diminish the formants of voiced speech signal and render the time-varying F_0 component discernible among its harmonics, the voiced speech signal is required to be filtered in the LFR. The next subsection describes the filtering of the voiced speech signal in the LFR.

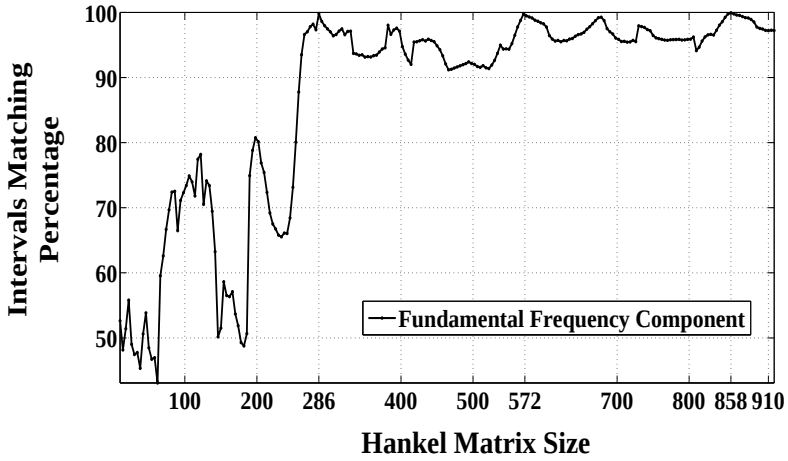

Figure 3.14: Interval matching percentage with respect to the Hankel matrix size (N) after the second *Iteration*. $N_0 \approx 286$ samples.

3.4.1 Filtering of voiced regions in the LFR

Let $x[n]$ denote a noisy voiced region spanning Q samples, detected using the method described in the previous chapter of this thesis. $x[n]$ contains the clean voiced speech signal, $s[n]$, and the additive noise $\xi[n]$; i.e, $x[n] = s[n] + \xi[n]$. The detected noisy voiced region is filtered in the LFR using the Fourier-Bessel (FB) coefficients. The FB coefficients are suitable for the analysis of a voiced speech signal because the Bessel functions used as bases in the FB series expansion of the signal are non-stationary in nature [79, 80]. The zero-order FB series expansion of a discrete-time noisy voiced speech signal $y[n]$ is given by [81, 82]:

$$x[n] = \sum_{l=1}^{Q} C_l^x J_0\left(\frac{\lambda_l n}{Q}\right), \quad n = 0, 1, .., Q-1 \tag{3.26}$$

The FB coefficients C_l^y of $y[n]$ are computed by using the following analysis equation [82, 83]:

$$C_l^x = \frac{2}{Q^2[J_1(\lambda_l)]^2} \sum_{n=0}^{Q-1} n x[n] J_0\left(\frac{\lambda_l n}{Q}\right) \tag{3.27}$$

where $J_0(.)$ and $J_1(.)$ are the zero and first order Bessel functions respectively. The ascending order positive roots of the equation, $J_0(\lambda) = 0$ are denoted by λ_l, where the order l takes values from $1, 2, .., Q$. There exists one to one correspondence between the order l of the FB coefficient and the continuous-time frequency F_l in Hz at which it attains

peak as given by [82, 83]:

$$\lambda_l \approx \frac{2\pi F_l Q}{F_s} \quad \text{where} \quad \lambda_l \approx \lambda_{l-1} + \pi \approx l\pi \tag{3.28}$$

From (3.28), it can be inferred that:

$$l \approx \frac{2F_l Q}{F_s} \tag{3.29}$$

It can be deduced from (3.29) that the order l must vary from 1 to Q (length of the discrete-time signal) in order to represent the full bandwidth of the discrete-time signal excluding the DC component; i.e., $\left(0,\ \frac{F_s}{2}\right]$ Hz. Please note that the FB coefficients cannot represent the DC component of the signal. Therefore, any DC component present in the signal must be removed prior to computing the FB coefficients of the signal. The noisy voiced speech signal is filtered in the LFR by using the synthesis equation in (3.26) with the order l varying in the range of (L_1, L_2) corresponding to the LFR as follows:

$$x_{\text{LF}}[n] = \sum_{l=L_1}^{L_2} C_l^x J_0\left(\frac{\lambda_l n}{Q}\right), \quad n = 0, 1, .., Q-1 \tag{3.30}$$

where $L_1 = \dfrac{2 \times 50Q}{F_s}$ and $L_2 = \dfrac{2 \times 500Q}{F_s}$ using (3.29) and $x_{\text{LF}}[n]$ denotes the LFR filtered noisy voiced speech signal. Using the linearity property of the FB expansion [83], (3.30) can be rewritten as:

$$x_{\text{LF}}[n] = \sum_{l=L_1}^{L_2} (C_l^s + C_l^\xi) J_0\left(\frac{\lambda_l n}{Q}\right) = s_{\text{LF}}[n] + \xi_{\text{LF}}[n] \tag{3.31}$$

where C_l^s and C_l^ξ represent the FB coefficients of $s[n]$ and $\xi[n]$ respectively. It can be inferred from (3.31) that in $x_{\text{LF}}[n]$, the harmonic structure of $s_{\text{LF}}[n]$ given by (3.6) is corrupted by the components of the additive noise present in the LFR. It can be deduced from (3.6) that filtering of the voiced speech signal in the LFR suppresses the noise energy lying outside the LFR and thus, aids in achieving robustness against noise. In the next section, the theory and concepts developed in the previous section for extraction of

harmonically related constant amplitude/frequency mono-component signals contained in a multi-component signal are extended to extract the time-varying F_0 component from the LFR filtered noisy voiced speech signal $x_{\mathrm{LF}}[n]$. The negative cycles of the extracted time-varying F_0 component of $x_{\mathrm{LF}}[n]$ provide reliable coarse estimate of intervals where GCIs are likely to occur.

3.4.2 Evolution of concepts for the time-varying case

Let $y[n] = x_{\mathrm{LF}}[n]$ where $x_{\mathrm{LF}}[n]$ is the LFR filtered noisy voiced speech signal in (3.31) spanning $(0, Q-1)$ samples. Using (3.6), (3.7) and (3.31), H_N^y can be expressed as follows:

$$H_N^y = H_N^{s_{\mathrm{LF}}} + H_N^{\xi_{\mathrm{LF}}} = \sum_{k=1}^{K} H_N^{s_{\mathrm{LF},k}} + H_N^{\xi_{\mathrm{LF}}} \tag{3.32}$$

where $s_{\mathrm{LF},k}[n]$ represents the k^{th} mono-component signal of $s_{\mathrm{LF}}[n]$; i.e, $s_{\mathrm{LF},k}[n] = A_k[n]\cos(2\pi k f_0[n]n + \theta_k[n])$ and k varies from $1, 2, ..., K$. The time-varying fundamental frequency $F_0[n]$ varies slowly with time [45]. The rank of $H_N^{s_{\mathrm{LF}}}$ is close to $2K$ if frequencies of all K mono-component signals of $s_{\mathrm{LF}}[n]$ change significantly slow over the period of N data samples [84]. Accordingly, the value of N must be kept small such that $F_0[n]$ changes significantly slow over N data samples and the ranks of $H_N^{s_{\mathrm{LF}}}$ and $H_N^{s_{\mathrm{LF,\,k}}}$ are restricted to $2K$ and 2 respectively. This condition poses an upper limit on the value of N. On the other hand, it has been explained in section 3.3.2 that the error to signal ratio for the F_0 component of $y[n]$ reduces considerably for $N > N_0$ and the value of IMP increases substantially for $N > N_0$. This observation poses a lower limit on the value of N. The maximum value of N_0 denoted by $N_{0,\max}$ for voiced speech signal can be $\frac{F_s}{50}$ corresponding to the minimum possible value of F_0, denoted as $F_{0,\min} = 50$ Hz. Thus, the value of $2N-1$ is chosen to be the smallest integer which divides the Q samples of $y[n]$ into L equal size segments denoted by $y_l[n], l = 1, 2, ..., L$ of length $2N-1$ samples, subject to the constraint that $N > N_{0,\max}$; i.e, $N > \frac{F_s}{50}$. The EVD is performed on $H_N^{y_l}\ \forall l$ and the time-varying F_0 component is extracted for each $y_l[n]$ where $l = 1, 2, ..., L$. Please note that the assumption of stationarity is not required to divide the voiced speech signal into segments.

The presence of additive noise in $y_l[n]$ increases the rank and number of non-zero eigenvalue pairs of $H_N^{y_l}$ by adding energy in the l^{th} segment of $s_{\mathrm{LF}}[n]$ at frequencies that may not coincide with $F_0[n]$ or its harmonics. The time-varying F_0 component of $y_l[n]$ has significant energy relative to other harmonic components in the LFR as evident in Fig. 3.1 (b); therefore, only components with magnitude of the eigenvalue equal to or greater than one-fourth of the maximum eigenvalue of $H_N^{y_l}$ are extracted. Ignoring the non-significant eigenvalue pairs of $H_N^{y_l}$ which may correspond to components of l^{th} segment of the LFR filtered noise $\xi_{\mathrm{LF}}[n]$ contained in $y_l[n]$ facilitates further noise suppression. Let the m^{th} extracted component of $y_l[n]$ corresponding to the m^{th} significant eigenvalue pair of $H_N^{y_l}$ be denoted by $y_{l,m}[n]$. We define two parameters associated with an extracted component as follows:

(a) Dominant frequency: The positive frequency at which the square magnitude spectrum of the extracted component attains its peak.

(b) Energy: The sum of the square values of the extracted component.

The noise environments (such as babble noise) that can introduce a narrowband component with significant energy in the frequency range lower than the frequency range of the time-varying F_0 component of the voiced speech signal introduce difficulty in determining the candidate for the time-varying F_0 component of $y_l[n]$ from among the significant extracted components of $y_l[n]$. Therefore, a *Distance Metric* based criterion is employed for noise resilient estimation of the F_0 range of a voiced speech signal as described in the next subsection.

3.4.3 *Distance Metric* based F_0 range estimation

A reliable estimate of the average value of $F_0[n]$ for the entire noisy voiced speech signal, denoted $\hat{F}_0$, can help in estimation of the F_0 range of $y_l[n]$ in a noise resilient manner. A robust estimate of $\hat{F}_0$ facilitates determination of the dominant frequency from among the dominant frequencies of all significant extracted components of $y_l[n]$, which can be used for estimating the F_0 range of $y_l[n]$ in an *Iteration*. In this thesis, the term *Iteration* refers to the entire process of performing EVD of $H_N^{y_l}$, extraction of components corresponding to the significant eigenvalue pairs and determination of the potential candidate for the

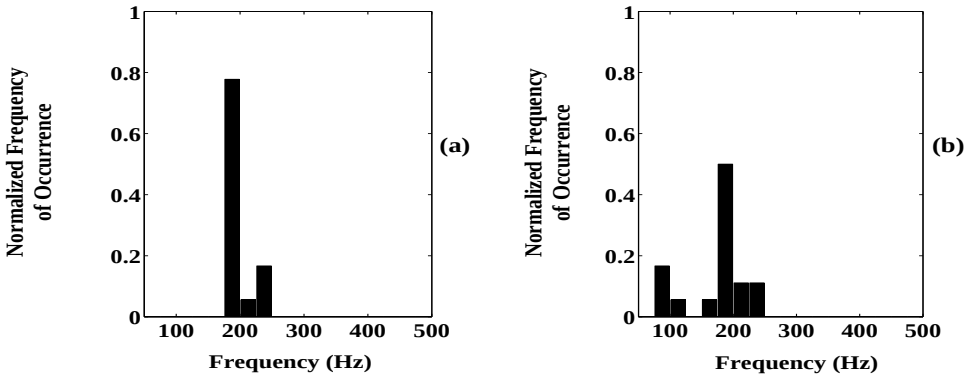

Figure 3.15: Histogram of F_0 values of short duration female speech segments estimated using the method proposed in [3] (a) in a clean environment (b) at 0 dB SNR in a babble noise environment.

time-varying F_0 component of $y_l[n]$.

There are several ways to estimate $\hat{F}_0$. The autocorrelation (ACF) method described in [3] has been used to estimate the fundamental periods (T_0) of central-clipped short duration (30 ms) voiced speech segments from the locations of the strongest peaks in their respective ACF in the interval of 2 ms - 20 ms (corresponding to the F_0 range of 50 Hz - 500 Hz). The F_0 values of voiced segments are obtained as inverse of their respective T_0 values. The histogram of the F_0 values of voiced segments is computed by dividing the F_0 range into 18 equal size bins of width 25 Hz. Let $p(F_i)$ denote the normalized frequency of occurrence of F_0 values in the bin with center F_i. The peak in the histogram of F_0 values of voiced segments provides an estimate of $\hat{F}_0$ as follows:

$$\hat{F}_0 = \arg_i \max(p(F_i)) \quad \forall i \tag{3.33}$$

Histograms of F_0 values computed for female voiced speech segments using the ACF method [3] in a clean environment and at 0 dB SNR in a babble noise environment are depicted in Fig. 3.15 (a) and Fig. 3.15 (b) respectively. The $\hat{F}_0$ value of 187.5 Hz has been estimated for the two scenarios depicted in Fig. 3.15 using (3.33). The position of the peak in the two histograms depicted in Fig. 3.15 remain unchanged irrespective of the SNR. It is understood from this example that $\hat{F}_0$ can be reliably estimated using (3.33). Let dominant frequencies of M extracted components corresponding to M significant

eigenvalue pairs of $H_N^{y_l}$ arranged in ascending order be denoted by $\breve{F}_{m,\max}, m = 1, 2, .., M$. We define a *Distance Metric*, denoted Υ_m for $\breve{F}_{m,\max}$ as the absolute value of the difference between $\hat{F}_0$ and $\breve{F}_{m,\max}$:

$$\Upsilon_m = |\hat{F}_0 - \breve{F}_{m,\max}| \quad , \; m = 1, 2, .., M \tag{3.34}$$

The average value of F_0 over $y_l[n]$ denoted by $\hat{F}_0^l$ is approximated by $\breve{F}_{m,\max}$ with the minimum value of Υ_m as follows:

$$\hat{F}_0^l \approx \breve{F}_{m_p,\max} \;\text{ if }\; \Upsilon_{m_p} = \min(\Upsilon_m) \; \forall m \tag{3.35}$$

where m_p is an integer in the set $\{1, 2, .., M\}$. As $1.5\hat{F}_0^l$ is approximately the mid-point between the time-varying F_0 of l^{th} segment of $s_{\text{LF}}[n]$ and its second harmonic frequency contained in $y_l[n]$, the F_0 range for $y_l[n]$ denoted by $(F_{0,L}^l, F_{0,H}^l)$ is defined using (3.33), (3.34) and (3.35) as follows:

$$\begin{aligned} F_{0,L}^l &= \max\left(50 \text{ Hz}, \frac{2\hat{F}_0^l}{3}\right) = \max\left(50 \text{ Hz}, \frac{2\breve{F}_{m_p,\max}}{3}\right) \\ F_{0,H}^l &= \min\left(500 \text{ Hz}, \frac{4\hat{F}_0^l}{3}\right) = \min\left(500 \text{ Hz}, \frac{4\breve{F}_{m_p,\max}}{3}\right) \end{aligned} \tag{3.36}$$

Let $y_{l,F_0}[n]$ represent the time-varying F_0 component of $y_l[n]$. The highest energy component among the significant extracted components of $y_l[n]$ with their corresponding dominant frequencies lying in the range $(F_{0,L}^l, F_{0,H}^l)$ given by (3.36) is considered as the potential candidate for $y_{l,F_0}[n]$.

The potential candidate for $y_{l,F_0}[n]$ may contain contamination of higher-order harmonic components of F_0 or noise components present in the LFR. Therefore, the *Mono-component Signal Criteria* defined in the previous subsection has been employed to ensure that the selected potential candidate is a mono-component signal. The *Mono-component Signal Criteria* defined in the previous section ensures effective removal of the leakage of other harmonics and noise from the potential candidate for $y_{l,F_0}[n]$. If the potential candidate for $y_{l,F_0}[n]$ obtained in section 3.4.3 satisfies the *Mono-component Signal Criteria*,

then it is considered as $y_{l,F_0}[n]$, otherwise the *Iteration* comprising of the entire process of EVD of the Hankel matrix, extraction of the significant components and determination of the potential candidate for $y_{l,F_0}[n]$ is repeated by treating the potential candidate obtained in section 3.4.3 as $y_l[n]$ for the next *Iteration*. The *Iterations* are repeated until $y_{l,F_0}[n]$ is extracted.

3.4.4 Proposed iterative algorithm

The steps of the proposed iterative algorithm for extracting the time-varying F_0 component of a noisy voiced speech signal represented by $x[n]$ are enumerated below:

1) Determine the average value of F_0 of $x[n]$ denoted by $\hat{F}_0$ by using (3.33) (Procedure stated in subsection 3.4.3).

2) Perform filtering of $x[n]$ in the LFR and compute $x_{\mathrm{LF}}[n]$ using (3.27), (3.29) and (3.30) (Procedure stated in subsection 3.4.2).

3) Let $y[n] = x_{\mathrm{LF}}[n]$, where $x_{\mathrm{LF}}[n]$ is the LFR filtered noisy voiced speech signal spanning Q samples given in (3.30). Divide $y[n]$ into equal size segments $y_l[n], l = 1, 2, .., L$ consisting of $2N-1$ samples, subject to the constraint that N is an even number greater than $\frac{F_s}{50}$ (Procedure stated in subsection 3.4.2).

4) Construct the Hankel matrix $H_N^{y_l}$ from $y_l[n]$ using (3.7). Perform EVD of $H_N^{y_l}$; i.e., $H_N^{y_l} = U_{y_l}\Lambda_{y_l}U_{y_l}^T$ as given by (3.8). Let the eigenvalues be arranged in increasing order i.e. $\lambda_{y_l,i+1} \geq \lambda_{y_l,i}$, $i = 1, 2, .., N-1$.

5) Determine the maximum eigenvalue denoted by $\lambda_{y_l,\max}$ among all eigen values, i.e. $\lambda_{y_l,\max} = \max(\lambda_{y_l,i}), i = 1, 2, ..., N$.

6) Determine the significant eigenvalue pairs of $H_N^{y_l}$. The m^{th} eigenvalue pair of $H_N^{y_l}$ denoted by $\zeta_{l,m}$, where $\zeta_{l,m} = (\lambda_{y_l,m}, \lambda_{y_l,N-m+1})$ is considered as significant or negligible as:

$$\begin{aligned} &\max(|\zeta_{l,m}|) \geq 0.25\ \lambda_{y_l,\max}, \quad \zeta_{l,m}\ \epsilon \text{ significant}, \\ &\max(|\zeta_{l,m}|) < 0.25\ \lambda_{y_l,\max}, \quad \zeta_{l,m}\ \epsilon \text{ negligible}. \end{aligned} \tag{3.37}$$

where m takes values from 1, 2, ..., $\frac{N}{2}$. Let there be M significant eigenvalue pairs.

7) Construct a modified eigenvalue diagonal matrix $\tilde{\Lambda}_{y_{l,m}}$ to preserve only the m^{th} significant eigenvalue pair as follows:

$$\tilde{\Lambda}_{y_{l,m}} = \text{diag}(0, .., 0, \lambda_{y_l,m}, 0, .., 0, \lambda_{y_l,N-m+1}, 0, .., 0) \tag{3.38}$$

where $m = 1, 2, ..., M$.

8) Extract the m^{th} significant component of $y_l[n]$, denoted by $\tilde{y}_{l,m}[n]$ by taking the mean of the elements of the skew-diagonals of $\tilde{H}^{y_{l,m}}$, where $\tilde{H}^{y_{l,m}}$ is constructed as:

$$\tilde{H}_N^{y_{l,m}} = U_{y_l} \tilde{\Lambda}_{y_{l,m}} U_{y_l}^T \tag{3.39}$$

9) Repeat steps 7-8 for $m = 1, 2, ..., M$.

10) Compute the R-point discrete Fourier transform (DFT), $\tilde{Y}_{l,m}(F_r)$ of $\tilde{y}_{l,m}[n]$ at frequencies $F_r = \frac{rF_s}{R}, r = 0, 1, ..., R-1$ as follows:

$$\tilde{Y}_{l,m}(F_r) = \sum_{n=0}^{2N-2} \tilde{y}_{l,m}[n] e^{-j2\pi f_r n}, \quad f_r = \frac{F_r}{F_s} \tag{3.40}$$

where $m = 1, 2, ..., M$. The value of R is chosen high to compute the DFT with good frequency resolution.

11) Determine the dominant frequency at which $|\tilde{Y}_{l,m}(F_r)|^2$ attains the maximum as:

$$|\tilde{Y}_{l,m}(F_{m,\max})|^2 = \max(|\tilde{Y}_{l,m}(F_r)|^2), \quad m = 1, 2, ..., M \tag{3.41}$$

12) Arrange the extracted components $\tilde{y}_{l,m}[n]$ in increasing order of the dominant frequencies $F_{m,\max}$ and denote them by $\breve{y}_{l,m}[n]$, $m = 1, 2, ..., M$. The dominant frequencies of $\breve{y}_{l,m}[n]$ are denoted by $\breve{F}_{m,\max}$, such that $\breve{F}_{m+1,\max} > \breve{F}_{m,\max}$ for $m = 1, 2, ..., M-1$.

13) Compute the *Distance metric* Υ_m for each dominant frequency using (3.34). The

value of $\hat{F}_0$ is computed in the step 1 of this algorithm using (3.33). Determine the approximate value of $\hat{F}_0^l$ by using (3.35).

14) Determine the F_0 range for $y_l[n]$ using (3.36). Compute energies of the extracted components whose dominant frequencies lie in the range of $(F_{0,L}^l, F_{0,H}^l)$.

15) Let $\breve{y}_{l,m_0}[n]$, where m_0 is an integer in the set $\{M_1, M_1+1, ..., M_2\}$ be the highest energy component among the extracted components $\breve{y}_{l,m}[n], m \in \{M_1, M_1+1, ..., M_2\}$ whose dominant frequencies lie in the range of $(F_{0,L}^l, F_{0,H}^l)$ where $1 \leq M_1 \leq M_2 \leq M$. Thus, $\breve{y}_{l,m_0}[n]$ is the potential candidate for $y_{l,F_0}[n]$, the F_0 component of $y_l[n]$.

16) If $\breve{y}_{l,m_0}[n]$ does not satisfy the *Mono-component Signal Criteria*, then repeat the steps 4 - 15 by treating $y_{l,m_0}[n]$ as $y_l[n]$ for the next *Iteration* (steps 3 - 16). On the other hand, if $\breve{y}_{l,m_0}[n]$ satisfies the *Mono-component Signal Criteria* then it is considered as $y_{l,F_0}[n]$.

17) Repeat steps 4 - 16 for all segments $y_l[n]$, $l = 1, 2, ..., L$ to extract their time-varying F_0 components.

18) Concatenate the extracted time-varying F_0 components of all segments to obtain the time-varying F_0 component of $y[n]$ denoted by $y_{F_0}[n]$ as follows:

$$y_{F_0}[n] = \sum_{l=1}^{L} y_{l,F_0}[n - (2N-1)(l-1)], n = 0, 1, ..., Q-1 \tag{3.42}$$

where $y_{l,F_0}[n] = 0$ for $n < 0$ and $n > 2N-2$.

The next section demonstrates the experimental results obtained by the proposed iterative algorithm on a synthetic multi-component non-stationary signal with harmonically related components and a voiced speech signal.

3.5 Experimental Results and Discussion

In order to evaluate the efficacy of the proposed iterative algorithm, experiments have been performed on synthetic and natural (voiced speech signal) multi-component non-

stationary signals with harmonically related components in clean and noisy environments. The synthetic multi-component non-stationary signal is generated at F_s of 32 kHz. The speech signal is taken from the CMU-Arctic database [67, 68] available at F_s of 32 kHz. The corresponding time-aligned EGG signals are available in the CMU-Arctic database for speech signals of some speakers. The white and babble noise signals are taken from the NOISEX database [71]. The white noise signal was acquired by sampling a high quality analog noise generator. The source of the babble noise signal was 100 people speaking in a canteen. These noise signals are available at F_s of 19.98 kHz and therefore, have been resampled to 32 kHz before adding them to synthetic and natural (speech signal) multi-component non-stationary signals.

3.5.1 Synthetic multi-component non-stationary signal

The following AM-FM signal model has been used to generate a synthetic multi-component non-stationary signal represented by $y[n]$, containing three harmonically related time-varying mono-component signals:

$$x[n] = \sum_{k=1}^{3} A_k(1 + \alpha_k n)\cos(k\omega_0(1 + \beta n)n + \theta_k) \quad , \quad n = 0, 1, ..., Q-1 \tag{3.43}$$

where $\omega_0 = 2\pi f_0$ represents the normalized angular frequency. The first component of $x[n]$ corresponding to $k = 1$ is the time-varying fundamental frequency (F_0) component of $x[n]$. Please note that in (3.43), the variation of the instantaneous frequency and amplitude of the three mono-component signals contained in $x[n]$ is linear in nature. In (3.43), the parameter β controls the rate of variation of f_0 with respect to time and the parameter α_k controls the rate of variation of instantaneous amplitude of k^{th} harmonic component of f_0 with time. The values of various parameters used in (3.43) are chosen as: $Q = 4800, \beta = \dfrac{0.296}{Q}, A_1 = 0.017, A_2 = 0.022, A_3 = 0.012, \alpha_1 = 2, \alpha_3 = 1.8, \alpha_3 = 1.4, \theta_1 = 0, \theta_2 = \pi, \theta_3 = \dfrac{\pi}{2}, f_0 = \dfrac{100}{F_s}, F_s = 32$ kHz. Using (3.43) and the given values of various parameters of the model, it can be easily found out that the the fundamental frequency is varying from 100 Hz to 129.59 Hz which corresponds to a positive change of about 30% over 150 ms duration.

The LFR filtered synthetic signal $x_{\mathrm{LF}}[n]$ is obtained from $x[n]$ using (3.27), (3.29) and (3.30). Let $y[n] = x_{\mathrm{LF}}[n]$. Subject to the constraint that the Hankel matrix size N should be greater than $\dfrac{F_s}{50} = 640$ for $F_s = 32$ kHz (refer to sub-section 3.4.1), $y[n]$ has been divided into 3 equal size segments $y_l[n], l = 1, 2, 3$ of length 1599 samples and a square Hankel matrix of size 800×800 is constructed using the samples of each segment. The EVD has been performed on all Hankel matrices. The experimental results obtained by the proposed iterative algorithm on the third segment of $y[n]$ are tabulated in Table 3.2 and depicted in Fig. 3.16 and Fig. 3.18 for the two scenarios: clean environment and at 0 dB SNR in a white noise environment respectively. In order to objectively measure the difference between the extracted time-varying F_0 component and the original time-varying F_0 component of synthetic multi-component non-stationary signal with harmonically related components, we introduce the performance measure, error to signal ratio in dB denoted by ESR_{dB}, defined as follows:

$$\begin{aligned} err[n] &= y_{F_0}[n] - \tilde{y}_{F_0}[n] \\ ESR\ (\mathrm{dB}) &= 10 \log_{10} \left(\frac{\sum\limits_{n} (err[n])^2}{\sum\limits_{n} (y_{F_0}[n])^2} \right) \end{aligned} \tag{3.44}$$

where the extracted time-varying F_0 component and the original time-varying F_0 component of $y[n]$ are denoted by $\tilde{y}_{F_0}[n]$ and $y_{F_0}[n]$ respectively.

(a) **Clean Environment**: Understanding of the results obtained by the proposed iterative algorithm on the third segment of $y[n]$ requires explanation of the tabulated results (Table 3.2) for the clean environment. The same way of interpretation will follow for the results provided in Table 3.2 for the noisy environment. The third segment of $y[n]$ is denoted as $y_3[n]$ and is depicted in Fig. 3.16 (b). The value of $\hat{F}_0$ estimated using (3.33) is found to be 137.5 Hz. Three significant components have been extracted using (3.37), (3.38) and (3.39) by performing EVD of $H_N^{y_3}$. The dominant frequencies of the extracted components have been computed using (3.40) and (3.41). The *Distance Metric* values of all dominant frequencies have been computed using (3.34). The minimum *Distance Metric* value is 0.5 Hz corresponding to the dominant frequency value of 137 Hz. Thus,

Table 3.2: Experimental results of extraction of the time-varying fundamental frequency component from the third segment of the synthetic multi-component non-stationary signal given by (3.43) obtained using the proposed iterative algorithm in a clean environment and at 0 dB SNR in a white noise environment.

Average Value of F_0 of Entire Voiced Speech Signal in Hz ($\hat{F}_0$)	Segment Number (l)	Iteration Number	Significant Eigenvalue Pairs	Dominant Frequency of Extracted Components in Hz ($F_{m,\max}$)	Distance Metric in Hz (Υ_m)	Estimated F_0 Range in Hz ($F^l_{0,L}, F^l_{0,H}$)	Energy of Extracted Components	(D_n, D_r)
CLEAN ENVIRONMENT								
137.5	3	1	(-14.32, 13.96) **(-11.23, 11.15)** (-6.35, 6.53)	275.0 **137.0** 413.0	137.5 **0.5** 275.5	**(91.3, 182.7)**	1.27 **0.68** 0.24	**(2, 1)**
		2	**(-10.33 10.07)**	**137.0**	**0.5**	**(91.3, 182.7)**	**0.67**	**(0, 1)**
WHITE NOISE ENVIRONMENT (SNR: 0 dB)								
137.5	3	1	(-13.68 13.10) **(-11.86 12.06)** (-5.63 5.80)	273.0 **138.0** 415.0	135.5 **0.5** 277.5	**(92.0, 184.0)**	1.04 **0.67** 0.20	**(2, 1)**
		2	**(-10.08, 10.03)**	**137.0**	**0.5**	**(91.3, 182.7)**	**0.64**	**(0, 1)**

by using (3.35) and (3.36), the F_0 range for $y_3[n]$ is found to be (95.9 Hz, 178.1 Hz). The component extracted using the second eigenvalue pair has the highest energy with the dominant frequency lying in the estimated F_0 range and therefore, it is the potential candidate (Fig. 3.16 (c)) for the time-varying F_0 component of $y_3[n]$. However, it does not satisfy the *Mono-component Signal Criteria* with values of $D_n = 2$ and $D_r = 1$. Therefore, the potential candidate in the first *Iteration* has been treated as $y_3[n]$ for the second *Iteration*. In the second *Iteration*, one significant component using (3.37), (3.38) and (3.39) has been extracted by performing EVD of the Hankel matrix constructed from the samples of the potential candidate obtained in the first *Iteration*. The minimum *Distance Metric* value is 0.5 Hz in the second *Iteration* corresponding to the dominant frequency value of 137 Hz. Thus, by using (3.35) and (3.36), the F_0 range for $y_3[n]$ found to be (95.9 Hz, 178.1 Hz). The component extracted using the first eigenvalue pair in the second *Iteration* (Fig. 3.16 (d)) is the time-varying F_0 component of $y_3[n]$ (Fig. 3.16 (b)) because it is the highest energy component with the dominant frequency lying in the estimated F_0 range and satisfies the *Mono-component Signal Criteria* with values of $D_n = 0$ and $D_r = 1$. It is evident from Fig. 3.16 (d) that the extracted component was able to follow the variation in the amplitude and frequency of the time-varying F_0 component of $y_3[n]$. The value of error to signal ratio for the F_0 component in dB and the value of IMP (refer to equations (3.44) and (3.25)) are found to be to be -14 dB and 97% respectively.

The results obtained by the proposed iterative algorithm on the entire synthetic multi-

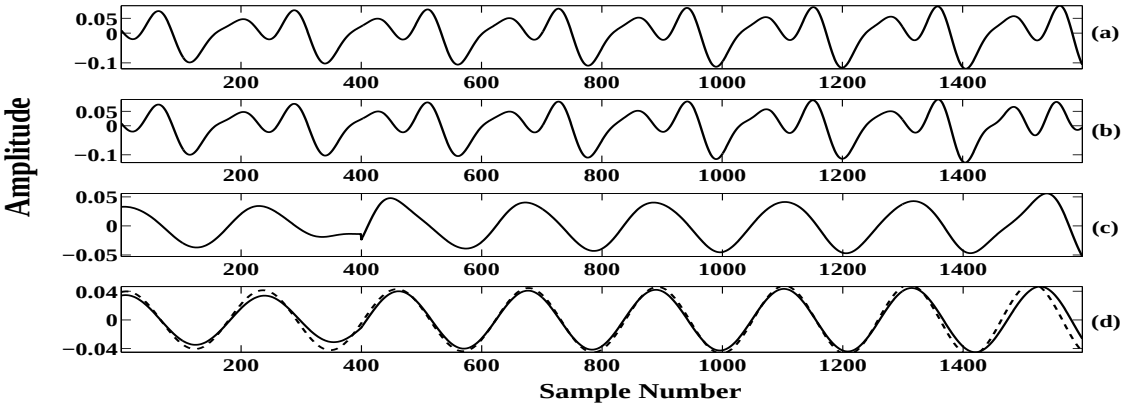

Figure 3.16: (a) Third segment of a clean synthetic multi-component non-stationary signal (b) LFR filtered synthetic multi-component non-stationary signal ($y_3[n]$) (c) Potential Candidate for the time-varying F_0 component obtained in the first *Iteration* (d) Extracted time-varying F_0 component by the proposed iterative algorithm in the second *Iteration* in solid line and the reference time-varying F_0 component in dashed line.

component non-stationary signal given by (3.40) in a clean environment is depicted in Fig. 3.17. The time-varying F_0 component of $y[n]$ depicted in Fig. 3.17 (c) has been obtained using (3.38), which involves concatenation of the time-varying F_0 components extracted from all segments of $y[n]$. It is evident from Fig. 3.17 (c) that the extracted time-varying F_0 component closely matches with the reference F_0 component of $y[n]$. The value of error to signal (refer to equation 3.44) for the F_0 component in dB computed is found to be -14.5 dB and the value of IMP is found to be 97.54% using (3.25).

(b) **Noisy Environment**: The results obtained by the proposed iterative algorithm at 0 dB SNR in the white noise environment tabulated in Table 3.2 should be interpreted in the same way as explained in the previous subsection for the clean environment. The third segment of the LFR filtered noisy synthetic multi-component non-stationary signal ($y_3[n]$) at 0 dB SNR in a white noise environment is depicted in Fig. 3.18 (b). The time-varying F_0 component of $y_3[n]$ extracted using the proposed iterative algorithm in the second *Iteration* is depicted in Fig. 3.18 (d). The value of error to signal ratio for the F_0 component in dB and the value of IMP (refer to equations (3.44) and (3.25)) are found to be -12 dB and 96.25% respectively. It is evident from 3.18 (d) and the ESR_{dB} value that the proposed iterative algorithm was able to efficiently extract the time-varying F_0 component of $y_3[n]$ in a heavily noise degraded condition also.

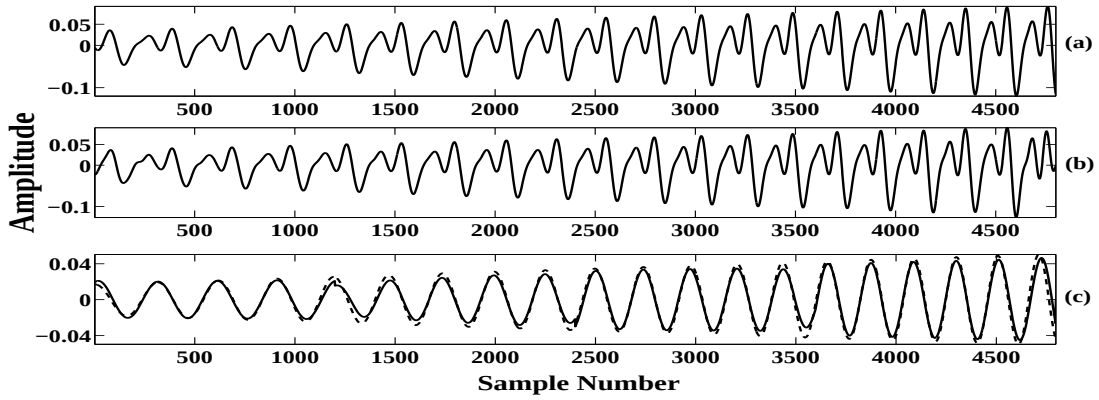

Figure 3.17: (a) Clean synthetic multi-component non-stationary signal (b) LFR filtered synthetic multi-component non-stationary signal ($y[n]$) (c) Extracted time-varying F_0 component by the proposed iterative algorithm in solid line and the reference time-varying F_0 component in dashed line.

The results obtained by the proposed iterative algorithm on the entire synthetic multi-component non-stationary signal given by (3.43) at 0 dB SNR in a white environment is depicted in Fig. 3.19. The time-varying F_0 component of $y[n]$ depicted in Fig. 3.19 (c) has been obtained using (3.42), which involves concatenation of the time-varying F_0 components extracted from all segments of $y[n]$. It is evident from Fig. 3.19 (c) that the extracted time-varying F_0 component closely matches with the reference F_0 component of $y[n]$. This result manifest the noise robustness of the proposed iterative algorithm. The value of error to signal ratio for the F_0 component in dB and the value of IMP are found to be -12.24 dB and 96.56% respectively (refer to equations (3.44) and (3.25)).

3.5.2 Voiced speech signal

The voiced regions of the speech signal of the CMU-Arctic database [67, 68] are detected using the instantaneous V/NV detection method proposed in the previous chapter of this thesis. The AM-FM signal model of voiced speech signal derived in section 3.2 assumes that the voiced speech signal in the LFR is a multi-component non-stationary signal with harmonically related components. A voiced region represented by $x[n]$ spanning 17825 samples at $F_s = 32$ kHz is selected for performing experiments using the proposed iterative algorithm. The LFR filtered voiced speech signal $x_{\mathrm{LF}}[n]$ is obtained from x[n] using (3.27),

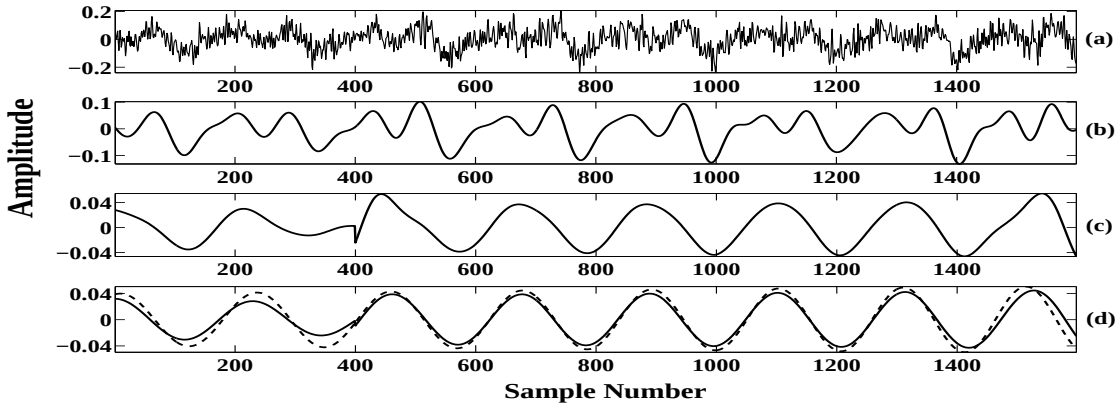

Figure 3.18: (a) Third segment of a synthetic multi-component non-stationary signal at 0 dB SNR in a white noise environment (b) LFR filtered noisy synthetic multi-component non-stationary signal ($y_3[n]$) (b) Potential Candidate for the time-varying F_0 component obtained in the first *Iteration* (c) Extracted time-varying F_0 component by the proposed iterative algorithm in the second *Iteration* in solid line and the reference time-varying F_0 component in dashed line.

(3.29) and (3.30). Let $y[n] = x_{\mathrm{LF}}[n]$. Subject to the constraint that the size of the Hankel matrix N should be greater than $\dfrac{F_s}{50} = 640$ for $F_s = 32$ kHz (refer to sub-section 3.4.1), the LFR filtered voiced speech signal $y[n]$ has been divided into 13 equal size segments $y_l[n], l = 1, 2, ..., 13$ of length 1371 samples and a square Hankel matrix of size 686×686 has been constructed from the samples of each segment. The results of extraction of the time-varying F_0 component from $y_3[n]$ obtained using the proposed iterative algorithm in a clean environment and at 0 dB SNR in a babble noise environment are tabulated in Table 3.3 and shown in Fig. 3.20 and Fig. 3.22 respectively.

(a) **Clean environment**: The third segment of $y[n]$ is denoted as $y_3[n]$ and is depicted in Fig. 3.20 (b). The results obtained by the proposed iterative algorithm on $y_3[n]$ in a clean environment, summarized in Table 3.3 are explained here. The value of $\hat{F}_0$ estimated using (3.33) is found to be 137.5 Hz. Three significant components using (3.37), (3.38) and (3.39) have been extracted by performing EVD of $H_N^{y_3}$. The dominant frequencies of the extracted components have been computed using (3.40) and (3.41). The *Distance Metric* values of all dominant frequencies have been computed using (3.34). The minimum *Distance Metric* value is 9.5 Hz corresponding to the dominant frequency value of 147

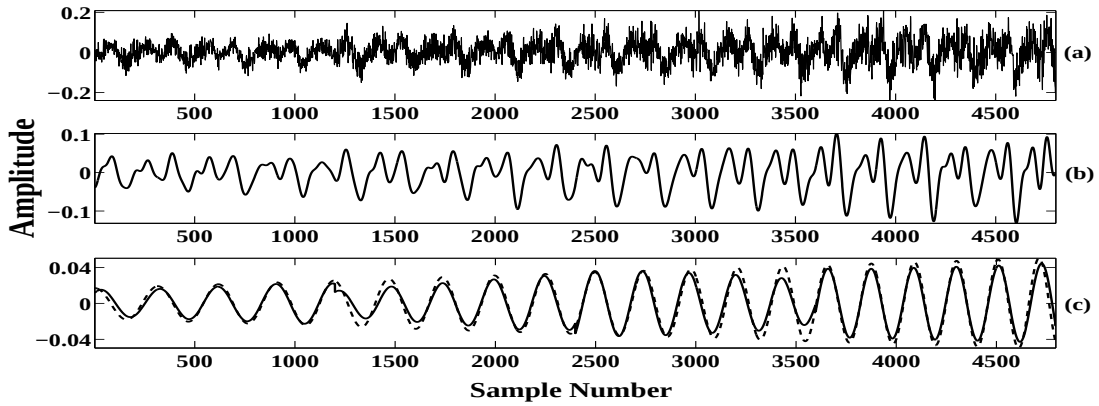

Figure 3.19: (a) Noisy synthetic multi-component non-stationary signal at 0 dB SNR in a white noise environment (b) LFR filtered noisy synthetic multi-component non-stationary signal ($y[n]$) (c) Extracted time-varying F_0 component by the proposed iterative algorithm in solid line and the reference time-varying F_0 component in dashed line.

Hz. Thus, by using (3.35) and (3.36), the F_0 range for $y_3[n]$ is found to be (98.0 Hz, 196.0 Hz). The component extracted using the third eigenvalue pair has the highest energy with the dominant frequency lying in the estimated F_0 range and therefore, it is the potential candidate selected (Fig. 3.20 (c)) for the time-varying F_0 component of $y_3[n]$ in the first *Iteration*. However, it does not satisfy the *Mono-component Signal Criteria* with values of $D_n = 11$ and $D_r = 2$. Therefore, the potential candidate in the first *Iteration* is treated as $y_3[n]$ for the second *Iteration*. In the second *Iteration*, two significant components have been extracted using (3.37), (3.38) and (3.39) by performing EVD of the Hankel matrix constructed from the samples of the potential candidate obtained in the first *Iteration*. The minimum *Distance Metric* value is 9.5 Hz in the second *Iteration* corresponding to the dominant frequency value of 147 Hz. Thus, by using (3.35) and (3.36), the F_0 range for $y_3[n]$ comes out to be (98.0 Hz, 196.0 Hz). The component extracted using the first eigenvalue pair in the second Iteration (Fig. 3.20 (d)) is the time-varying F_0 component of $y_3[n]$ (Fig. 3.20 (b)) because it is the highest energy component with the dominant frequency lying in the estimated F_0 range and satisfies the *Mono-component Signal Criteria* with values of $D_n = 1$ and $D_r = 1$. It is evident from Fig. 3.20 (d) that the negative cycles of the extracted time-varying F_0 component encompass all GCIs apparent in the differenced EGG (DEGG) signal shown in Fig. 3.20 (d) in dashed line.

The results obtained by the proposed iterative algorithm on the entire voiced speech

Table 3.3: Experimental results of extraction of the time-varying fundamental frequency component from the third segment of the LFR filtered voiced speech signal obtained using the proposed iterative algorithm in a clean environment and at 0 dB SNR in a babble noise environment.

Average Value of F_0 of Entire Voiced Speech Signal in Hz ($\hat{F}_0$)	Segment Number (l)	Iteration Number	Significant Eigenvalue Pairs	Dominant Frequency of Extracted Components in Hz ($F_{m,\max}$)	Distance Metric in Hz (Υ_m)	Estimated F_0 Range in Hz ($F^l_{0,L}, F^l_{0,H}$)	Energy of Extracted Components	(D_n, D_r)
CLEAN ENVIRONMENT								
137.5	3	1	(-44.50, 43.65)	295.0	157.5		11.84	
			(-27.39, 28.80)	443.0	305.5		3.98	
			(-25.81, 26.24)	**147.0**	**9.5**	**(98.0, 196.0)**	**2.08**	**(11, 2)**
		2	**(-18.33, 19.34)**	**147.0**	**9.5**	**(98.0, 196.0)**	**1.99**	**(1, 1)**
			(-6.16, 6.29)	441.0	303.5		0.27	
BABBLE NOISE ENVIRONMENT (SNR: 0 dB)								
137.5	3	1	**(-45.20, 48.44)**	**142.0**	**4.5**	**(94.7, 189.3)**	**11.40**	**(15, 3)**
			(-43.47, 44.32)	145.0	7.5		5.81	
			(-38.16, 37.19)	299.0	161.5		4.41	
			(-23.60, 22.91)	403.0	265.5		1.82	
			(-20.08, 21.55)	406.0	268.5		0.99	
			(-14.11, 14.96)	100.0	37.5		0.85	
		2	**(-23.50, 26.67)**	**143.0**	**5.5**	**(95.3, 190.7)**	**5.52**	**(0, 1)**
			(-15.97, 16.70)	296.0	158.5		0.30	
			(-7.37, 7.39)	442.0	304.5		2.54	

signal in a clean environment is depicted in Fig. 3.21. The time-varying F_0 component of the LFR filtered voiced speech signal $y[n]$ depicted in Fig. 3.21 (c) has been obtained using (3.42), which involves concatenation of the time-varying F_0 components extracted from all 13 segments of $y[n]$. It is evident from Fig. 3.21 (c) that the negative cycles of the extracted time-varying F_0 component using the proposed iterative algorithm encompass all GCIs apparent in the DEGG signal shown in Fig. 3.21 (d) in dashed line. The results of this experiment demonstrates the ability of the proposed iterative algorithm to efficiently extract the time-varying F_0 component from the LFR filtered clean voiced speech signal.

(b)**Noisy Environment** The results obtained by the proposed iterative algorithm at 0 dB SNR in a babble noise environment tabulated in Table 3.3 should be interpreted in the same way as explained in the previous subsection for the clean environment. The third segment of the LFR filtered noisy voiced speech signal ($y_3[n]$) at 0 dB SNR in a babble noise environment is depicted in Fig. 3.22 (b). The time-varying F_0 component of $y_3[n]$ extracted using the proposed iterative algorithm in the second *Iteration* is depicted in Fig. 3.22 (d) along with the DEGG signal in dashed line. The noise robustness of the proposed iterative algorithm is apparent in Fig. 3.22 (d), where the negative cycles of the extracted time-varying F_0 component encompass all GCIs evident in the DEGG signal shown in Fig. 3.22 (d) in dashed line.

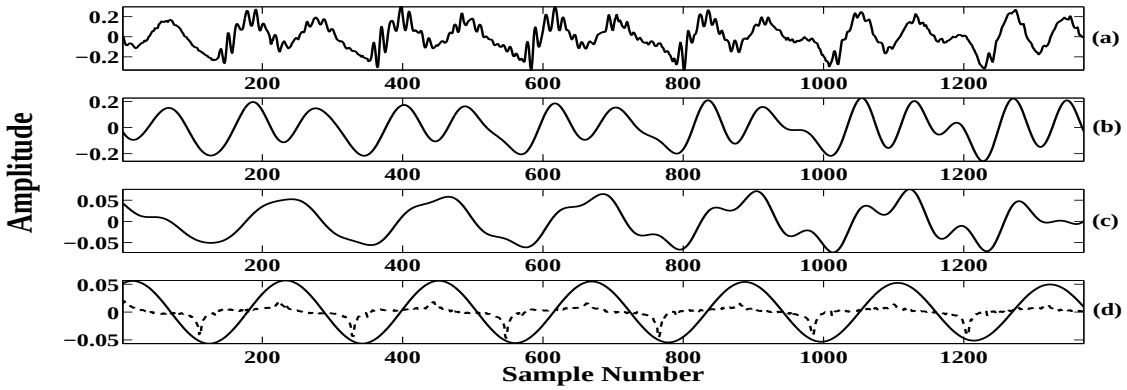

Figure 3.20: (a) Third segment of a clean voiced speech signal (b) LFR filtered clean voiced speech segment ($y_3[n]$) (c) Potential candidate for the time-varying F_0 component obtained in the first *Iteration* (d) Extracted time-varying F_0 component by the proposed iterative algorithm in the second *Iteration* and the DEGG signal are shown in solid and dashed lines respectively.

The results obtained by the proposed iterative algorithm on the entire noisy voiced speech signal at 0 dB SNR in white and babble noise environments are depicted in Fig. 3.23 and Fig. 3.24 respectively. The time-varying F_0 component of $y[n]$ depicted in Fig. 3.23 (c) and Fig. 3.24 (c) are obtained using (3.42), which involves concatenation of the time-varying F_0 components extracted from all segments of $y[n]$. The negative cycles of the extracted time-varying F_0 component shown in Fig. 3.23 (c) encompass all GCIs apparent in the DEGG signal depicted in Fig. 3.23 (c) in dashed line. It can be deduced from Fig. 3.23 (c) that the proposed iterative algorithm is resilient to white noise environment and performs efficiently at low SNRs also.

The negative cycles of the extracted time-varying F_0 component shown in Fig. 3.24 (c) are able to encompass 69 out of 74 GCIs ($\approx 93\%$) evident in the DEGG signal depicted in Fig. 3.24 (c) in dashed line. The negative cycles of the extracted time-varying F_0 component shown in Fig. 3.24 (c) fail to encompass a few GCIs around the sample numbers 9800, 13700, 14650, 14900, 16200. It can be observed in Fig. 3.24 (b) that some of these missed GCIs (around sample numbers 9800, 16200) belong to weak voiced regions. Also, an erroneous component with three negative cycles not corresponding to GCIs is extracted from the last segment (sample number range from 16453 - 17823) of the LFR filtered noisy voiced speech signal $y[n]$. It can be observed in Fig. 3.24 (b)

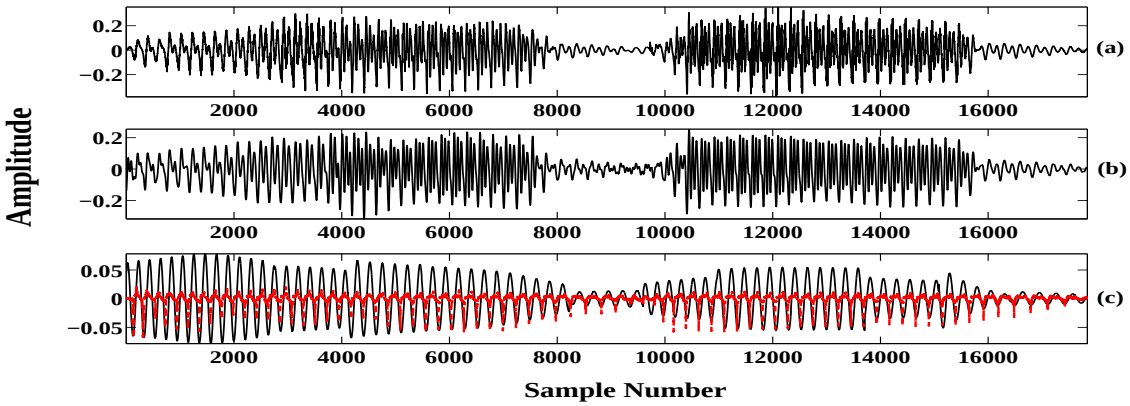

Figure 3.21: (a) Clean voiced speech signal (b) LFR filtered voiced speech signal (c) Extracted time-varying F_0 component by the proposed iterative algorithm in solid line and the DEGG signal in dashed line.

that the last segment of $y[n]$ is a weak voiced region. The results of this experiment demonstrates the robustness of the proposed iterative algorithm against the babble noise environment even at low SNRs. It can be concluded from the analysis of results depicted in Fig. 3.24 (c) that the proposed method performs fairly well in extracting the time-varying F_0 component from severely noise degraded voiced speech signal. The proposed iterative algorithm performed better in the white noise environment than the babble noise environment because the babble noise environment has higher energy in the LFR and hence, causes more signal distortion in the LFR than the white noise environment.

3.6 Conclusion

A noise resilient iterative algorithm for extraction of the time-varying F_0 component from a voiced speech signal based on EVD of the Hankel matrix has been proposed in this chapter. The Hankel matrix is initially constructed from the samples of the LFR filtered voiced speech signal. The condition on the Hankel matrix size to enable extraction of the time-varying F_0 component of the voiced speech signal using repetitive EVD has been derived. The employed *Mono-component Signal Criteria* has ensured effective removal of contamination from the potential candidate for the time-varying F_0 component of the voiced speech signal in successive *Iterations*. There is no requirement of the assumption of stationarity of the voiced speech signal over short time periods. No assumption has

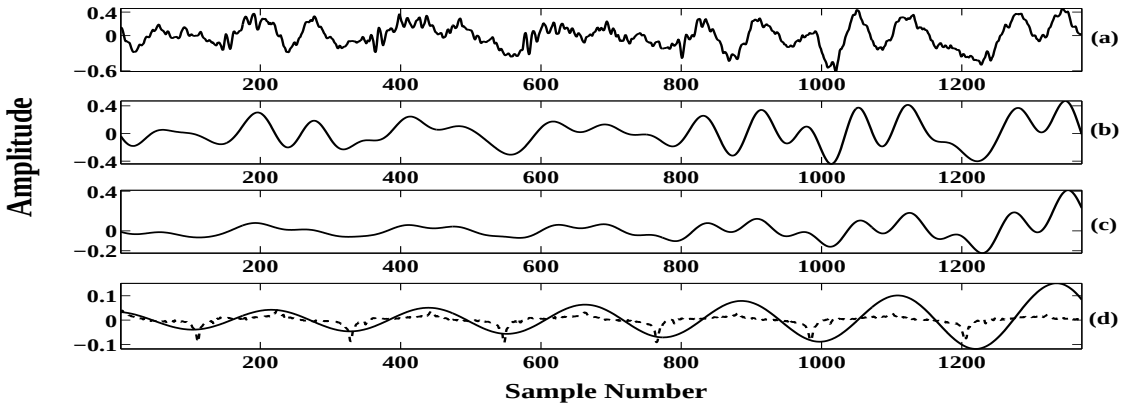

Figure 3.22: (a) Third segment of a noisy voiced speech signal at 0 dB in a babble noise environment (b) LFR filtered noisy voiced speech segment ($y_3[n]$) (c) Potential candidate for the time-varying F_0 component obtained in the first *Iteration* (d) Extracted time-varying F_0 component by the proposed iterative algorithm in the second *Iteration* and the DEGG signal are shown in solid and dashed lines respectively.

been made that the time-varying F_0 component possess the highest energy in the LFR filtered voiced speech signal.

The proposed iterative algorithm has been shown to efficiently extract the time-varying F_0 component from noise degraded voiced speech signal. The derived AM-FM model of the voiced speech signal in the LFR signifies energy only around the time-varying F_0 and its harmonics. The filtering of the voiced speech signal in the LFR rendered the time-varying F_0 component of the voiced speech signal discernible among its harmonics by attenuating formants. It also helped in achieving noise robustness by removing the noise energy lying outside the LFR. The rejection of the non-significant eigenvalue pairs of the Hankel matrix in an *Iteration*, which may correspond to LFR components of the noise signal, aids in noise suppression. The noise resilience of the proposed iterative algorithm is also attributable to the *Distance Metric* based F_0 range estimation.

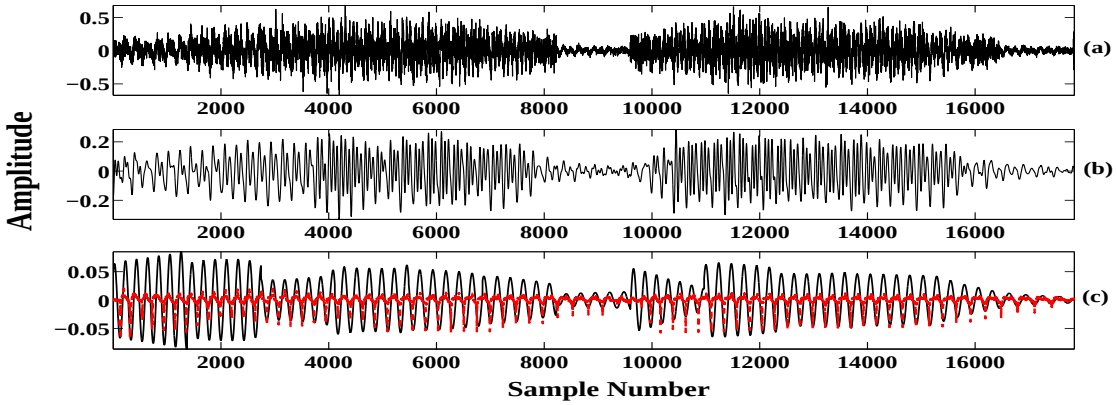

Figure 3.23: (a) Noisy voiced speech signal at 0 dB in a white noise environment (b) LFR filtered noisy voiced speech signal ($y[n]$) (c) Extracted time-varying F_0 component by the proposed iterative algorithm in solid line and the DEGG signal in dashed line.

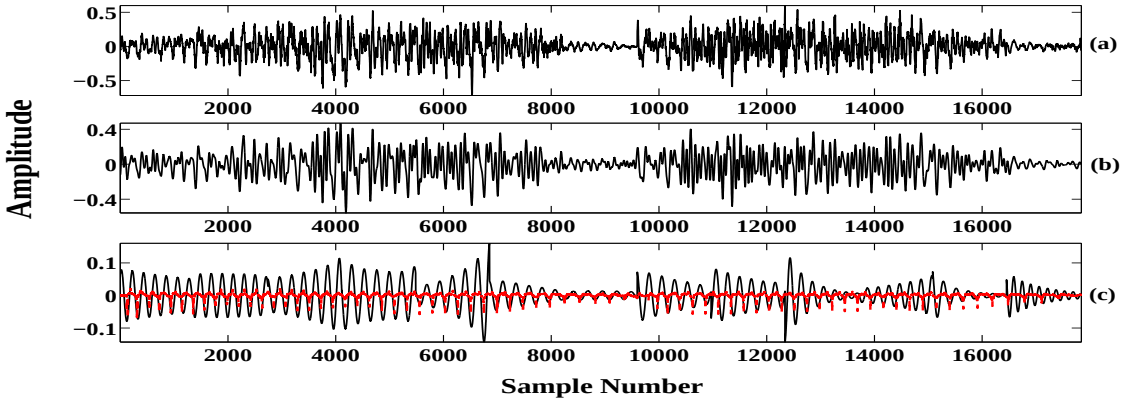

Figure 3.24: (a) Noisy voiced speech signal at 0 dB in a babble noise environment (b) LFR filtered noisy voiced speech signal ($y[n]$) (c) Extracted time-varying F_0 component by the proposed iterative algorithm in solid line and the DEGG signal in dashed line.

Chapter 4

Identification of Glottal Closure Instants

This chapter proposes a robust method to accurately identify glottal closure instants (GCIs) in the voiced speech signal. This chapter assumes that the voiced regions of speech signal are detected by using the V/NV detection method described in the second chapter of this thesis. The proposed method relies on the noise resilient extraction of the time-varying F_0 component of voiced speech signal using the iterative algorithm described in the previous chapter of this thesis, to provide reliable coarse estimates of intervals where GCIs are likely to occur. The negative cycles of the LFR filtered voiced speech signal occurring within these intervals are isolated. GCIs are identified in two steps: In the first step, GCI candidates are detected as local minima in the derivative of the falling edges of the isolated negative cycles of the LFR filtered voiced speech signal. In the second step, a selection criterion is used to discard false GCI candidates. An objective performance comparison of the proposed GCI identification method with some state of the art methods on speech signals of the CMU-Arctic database in clean and noisy environments demonstrate the superiority of the proposed method over existing methods.

4.1 Introduction

The vocal cords vibrate during the production of voiced speech, rendering the excitation to take the form of quasi-periodic puffs of air. The amplitude and phase of the harmonic components contained in the quasi-periodic excitation are modified by the vocal tract system in a time-varying manner. The rate of vibration of the vocal folds is comprehended

as the fundamental frequency (F_0) of voiced speech. Pitch of the voiced speech signal corresponds closely to its F_0. During each glottal cycle, the excitation to the vocal tract system attains the peak at the instant of closure of the glottis (GCI) which causes a sudden decrease in the glottal impedance, resulting in a high signal strength.

The accurate detection of GCIs from the speech signal found use in various speech signal processing applications. The instantaneous F_0 of voiced speech signal can be estimated as inverse of the interval between successive GCIs [10, 45]. Prosody manipulation can be performed in a pitch-synchronous manner by employing the identified GCIs as pitch period markers. Prosody manipulation finds use in applications such as text to speech synthesis, voice conversion, expressive speech synthesis [14–16]. The closed phase in each glottal cycle can be located with the help of GCIs. The inverse filtering techniques require identification of the close phase of glottal cycles for estimation of the glottal source excitation [2]. The excitation parameters characterizing the glottal flow derivative around GCIs provide speaker-specific features that aid in speaker identification and verification [8, 9]. Parametric voice coding can be performed with the knowledge of GCIs by modeling the voiced speech signal in each glottal cycle [7]. Applications like speech enhancement, speaker recognition, emotion recognition entail the detection of GCIs from noisy speech signals [13, 19, 20, 85].

Various methods have been reported in the literature for detection of GCIs from speech signals. One of the earlier methods to identify GCIs was based on the auto-covariance matrix whose elements were computed using the samples of the speech signal [36]. The peak in the linear prediction (LP) residual of the voiced speech signal within a pitch period was located as the GCI in [86]. The ambiguity arising from the presence of peaks of opposite polarities around GCIs in the LP residual was eliminated by computing its Hilbert envelope in [37]. The Frobenius norm of the voiced speech signal computed using a sliding window was used to estimate the signal energy at each sample instant and the GCI was detected as the instant with maximum energy within a pitch period [38]. An amplitude-frequency modulated (AM-FM) signal model based approach for GCI detection was proposed in [41] but it requires the band-limited (0 Hz - 300 Hz) speech signal to be a mono-component signal. The lines of maximum amplitudes (LOMA) obtained from

the amplitude maxima of the wavelet transform of the speech signal computed at various scales was used to locate GCIs in [40]. The group delay function based approach to determine GCIs from the positive zero crossings of the average slope function of the unwrapped phase of the short-time Fourier transform (STFT) of the LP residual was proposed in [87]. The dynamic programming projected phase slope algorithm (DYPSA) that determined GCI candidates from positive and projected zero crossings of the phase-slope function followed by further refinement by the dynamic programming was proposed in [23]. However, above mentioned methods suffer from low accuracy and performance degradation in the presence of noise. Lately, we proposed a time-order representation (TOR) based method for accurately detecting GCIs from the speech signal with excellent identification rate [66]. However, the main shortcoming of [66] is that the extraction of the time-varying F_0 component of a voiced speech signal by finding the first local peak in the marginal energy density with respect to frequency (MEDF) over the low frequency range (LFR) is vulnerable to noise because false local peaks arise in the MEDF around the true local peak corresponding to F_0 in the presence of noise.

The determination of GCIs from noisy speech signals was addressed in [6, 24, 46, 88]. A maximum likelihood theory based method for estimation of GCIs was proposed in [45]. Cohen's class time-frequency representation based method derived a detection function in the time-frequency plane followed by a morphological closing to detect GCIs from noisy voiced regions in [88]. Recently, a quantitative performance comparison of five contemporary GCI detection methods namely: Hilbert envelope of the LP residual based method [37], DYPSA [23], zero frequency resonator (ZFR) based method [6], yet another GCI algorithm (YAGA) [42] and speech event detection based on the residual energy and a mean based signal (SEDREAMS) [24] was performed in [43]. The positive zero crossings of the zero frequency resonator (ZFR) filtered speech signal was used to extract GCIs in [6]. The YAGA method is a union of existing GCI detection techniques using a framework based on the DYPSA algorithm [42]. In SEDREAMS method [24], intervals where GCIs are likely to occur were derived from the mean-based signal, followed by determining a precise location of the GCI within an interval by locating a discontinuity in the LP residual. The YAGA and SEDREAMS methods were shown to provide the highest accu-

racy in estimating GCIs from clean voiced speech signals [43]. The SEDREAMS and ZFR based methods were demonstrated to be noise resilient and provided high identification rates at moderate to low SNRs [43].

In order to achieve further improvement in the performance, we envisaged the employment of the time-varying F_0 component of the voiced speech signal for GCI identification. The negative cycles of the time-varying F_0 component provide reliable coarse estimate of intervals where GCIs are likely to occur [39]. This chapter presents an accurate and noise resilient method to identify GCIs in the voiced speech signal that relies on the iterative algorithm described in the previous chapter of this thesis for extraction of the time-varying F_0 component of the voiced speech signal. During each glottal cycle, the abrupt closure of the glottis (refer to Fig. 1.2) causes the glottal impedance to fall sharply; therefore, the proposed method detects GCI candidates as local minima in the derivative of the falling edges of the negative cycles of the LFR filtered voiced speech signal occurring within the intervals marked by the negative cycles of the time-varying F_0 component. The analysis of the speech signal in the LFR renders the glottal characteristics distinguishable by attenuating formants and helps to achieve noise robustness by removing the noise energy present at high frequencies. This chapter is organized as follows: The proposed method for GCI identification is described in Section 4.2. The experimental results on speech signals of the CMU-Arctic database in white and babble noise environments are presented and compared with some of state of the art methods in Section 4.3. The concluding remarks are provided in Section 4.4.

4.2 Proposed GCI Identification Method

GCIs are characteristic of the voiced speech signal; therefore, the voiced regions of speech signal are required to be detected before identifying GCIs. The noise resilient instantaneous V/NV detection method detailed in the second chapter of this thesis is used to detect voiced regions of speech signal. The detected voiced regions are filtered in the LFR using the FB coefficients as explained in subsection 3.4.1 of the previous chapter of this thesis. The filtering of voiced region in the LFR render the time-varying F_0 component

distinguishable among its harmonics, attenuates formant components, remove the DC component and noise energy present outside the LFR.

The proposed GCI identification method is based on the extraction of the time-varying F_0 component of the voiced speech signal. Let $x[n]$ represents a voiced speech signal spanning Q samples; i.e, n = 0, 1, ..., Q - 1. The time-varying F_0 component of $x[n]$ is extracted using the iterative algorithm stated in subsection 3.4.4 of the previous chapter of this thesis. The negative cycles of the extracted time-varying F_0 component represented by $x_{F_0}[n]$ provide reliable coarse estimate of intervals where GCIs are likely to occur. The negative cycles of the LFR filtered voiced speech signal represented by $x_{\mathrm{LF}}[n]$ occurring within the intervals marked by the negative cycles of $x_{F_0}[n]$ are isolated. There is a sudden decrease in the glottal impedance at GCIs, resulting in high signal strength; therefore, GCI candidates are detected as local minima in the derivative of the falling edges of the isolated negative cycles of $x_{\mathrm{LF}}[n]$. There can be only one GCI per negative cycle of $x_{F_0}[n]$ but sometimes more than one negative cycle of $x_{\mathrm{LF}}[n]$ occur within a single negative cycle of $x_{F_0}[n]$. The shape of the falling edge of a isolated negative cycle of $x_{\mathrm{LF}}[n]$ could also sometimes give rise to more than one local minima in its derivative. This leads to detection of many false GCI candidates along with the true ones. Therefore, a selection criterion is employed to retain only true GCI candidates. Out of all GCI candidates detected in the interval marked by a negative cycle of $x_{F_0}[n]$, the GCI candidate corresponding to the highest magnitude of $x_{\mathrm{LF}}[n]$ is selected. The steps of the proposed GCI identification method are as follows:

1. Let $x[n], n = 0, 1, ..., Q-1$ denote a voiced region of the speech signal detected using the V/NV detection algorithm stated in subsection 2.3.2 of the second chapter of this thesis. Let the sampling rate of $x[n]$ be denoted by F_s.

2. Let $x_{\mathrm{LF}}[n], n = 0, 1, ..., Q-1$ denote the LFR filtered voiced speech signal obtained using the FB coefficients of $x[n]$ corresponding to the LFR as described in subsection 3.4.1 of the third chapter of this thesis.

3. Divide $x_{\mathrm{LF}}[n]$ into L segments containing $2N-1$ samples, $x_l[n], l = 1, 2, ..., L$, where N is the smallest even number which divides the Q samples of $x_{\mathrm{LF}}[n]$ into L equal

size segments, subject to the constraint that $N > \frac{F_s}{50}$ (refer to subsection 3.4.2). Extract the time-varying F_0 component of $x_l[n], \forall l$ using the iterative algorithm (refer to subsection 3.4.4). Concatenate the extracted time-varying F_0 components of $x_l[n], \forall l$ denoted by $x_{l,F_0}[n], \forall l$ to obtain the time-varying F_0 component of $x[n]$, as stated in subsection 3.4.4 of the third chapter of this thesis. Let the time-varying F_0 component of $x[n]$ be denoted by $x_{F_0}[n]$.

4. Create a new signal $\chi[n]$ from $x_{F_0}[n]$ which is equal to one for only the negative cycles of $x_{F_0}[n]$ as follows:

$$\chi[n] = \begin{cases} 0 & if \quad x_{F_0}[n] \geq 0 \\ 1 & if \quad x_{F_0}[n] < 0 \quad \forall n \end{cases} \tag{4.1}$$

5. Create a new signal $\Gamma[n]$ from $x_{\mathrm{LF}}[n]$ which is equal to one for only the negative cycles of $x_{\mathrm{LF}}[n]$ as:

$$\Gamma[n] = \begin{cases} 0 & if \quad x_{\mathrm{LF}}[n] \geq 0 \\ 1 & if \quad x_{\mathrm{LF}}[n] < 0 \quad \forall n \end{cases} \tag{4.2}$$

6. Compute the differenced LFR filtered noisy voiced speech signal denoted by $x'_{\mathrm{LF}}[n]$ as:

$$x'_{\mathrm{LF}}[n+1] = x_{\mathrm{LF}}[n+1] - x_{\mathrm{LF}}[n], \qquad n = 0, 1, ..., Q-2 \tag{4.3}$$

where $x'_{\mathrm{LF}}[0] = 0$.

7. Create a new signal $\vartheta[n]$ which preserves the derivative corresponding to the falling edges of $x_{\mathrm{LF}}[n]$ as follows:

$$\vartheta[n] = \begin{cases} 0 & if \quad x'_{\mathrm{LF}}[n] \geq 0 \\ x'_{\mathrm{LF}}[n] & if \quad x'_{\mathrm{LF}}[n] < 0 \quad \forall n \end{cases} \tag{4.4}$$

8. GCI Detection Signal $\tilde{\vartheta}[n]$: Extract the derivative corresponding to the falling edges

of the peak negative cycles of $x[n]$ as:

$$\tilde{\vartheta}[n] = \vartheta[n] \times \Gamma[n] \times \chi[n] \qquad \forall n \tag{4.5}$$

Detect GCI candidates as local minima of $\tilde{\vartheta}[n]$.

9. GCI Selection Criterion: Compute the dominant frequency (refer to subsection 3.4.2) of $x_{l,F_0}[n]$ denoted by $\hat{F}_0^l$. A coarse estimate of the fundamental time period of $x_l[n]$ is given by $\hat{T}_0^l = \dfrac{1}{\hat{F}_0^l}$. If any two GCIs candidates detected in the l^{th} segment of $x[n]$ are separated by less than $\dfrac{\hat{T}_0^l}{2}$, then the GCI candidate corresponding to a higher magnitude of $x_{\mathrm{LF}}[n]$ is selected and the other is rejected.

The results of the proposed GCI identification method on a clean male voiced speech segment of duration 42.8 ms (1371 samples at $F_s = 32$ kHz) are shown in Fig. 4.1. The voiced speech signal is of 556.9 ms duration (17823 samples at $F_s = 32$ kHz) and is divided into 13 segments of length 1371 samples. Hence, the value of $2N - 1 = 1371$, where N denotes the Hankel matrix size N used by the iterative algorithm of subsection 3.4.4. The proposed GCI identification method accurately identified all GCIs apparent in the DEGG signal as shown in Fig. 4.1 (i). Fig. 4.2 depicts the results of the proposed GCI identification method on a clean female voiced speech segment of duration 50.1 ms (1603 samples at $F_s = 32$ kHz). The voiced speech signal is of 150.2 ms duration (4809 samples at $F_s = 32$ kHz) and is divided into 3 segments of length 1603 samples. Hence, the value of $2N - 1 = 802$, where N denotes the Hankel matrix size N used by the iterative algorithm of subsection 3.4.4. The proposed GCI identification method accurately identified all GCIs apparent in the DEGG signal as shown in Fig. 4.2 (i). The next section presents comprehensive results obtained by the proposed method on a speech database in clean and noisy environments.

4.3 Experimental Results and Discussion

In order to objectively assess the performance of the proposed GCI identification method and compare it with some existing noise-resilient GCI identification methods, experiments

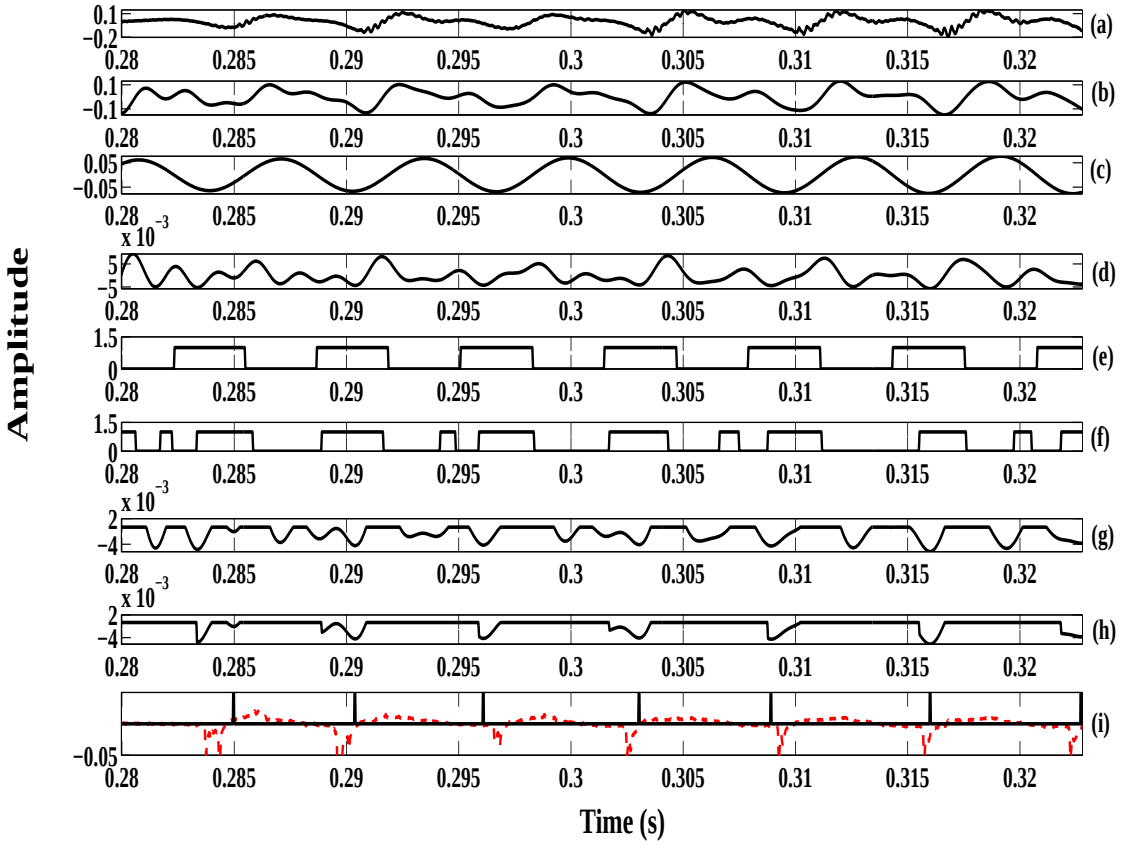

Figure 4.1: (a) Clean male voiced speech segment (b) LFR filtered voiced speech segment (c) Extracted time-varying F_0 component ($x_{1,F_0}[n]$) (d) Differenced LFR filtered voiced speech segment (e) $\chi[n]$ depicting the intervals marked by negative cycles of $x_{1,F_0}[n]$ (f) $\Gamma[n]$ depicting the intervals marked by negative cycles of the LFR filtered voiced speech segment (g) $\vartheta[n]$ depicting only negative cycles of the differenced LFR filtered voiced speech segment (e) GCI detection signal $\tilde{\vartheta}[n]$ whose local minima are detected as GCI candidates (i) GCIs identified in solid line and the DEGG signal in dashed line.

have been performed on a number of speech signals distorted with noise at various levels of degradation.

Databases: The experiments have been performed on speech signals of the CMU-Arctic database [67,68] available at a sampling rate of 32 kHz and 16 bit resolution. The database consists of around 1150 phonetically balanced sentences spoken in the English language by each of the seven speakers (five male and two female). The corresponding time-aligned electroglottograph (EGG) signals are available for speech signals of three speakers (bdl: US male, jmk: Canadian male, slt: US female). The white and babble noise signals are taken from the NOISEX-92 database [71]. The noise signals are available at a sampling

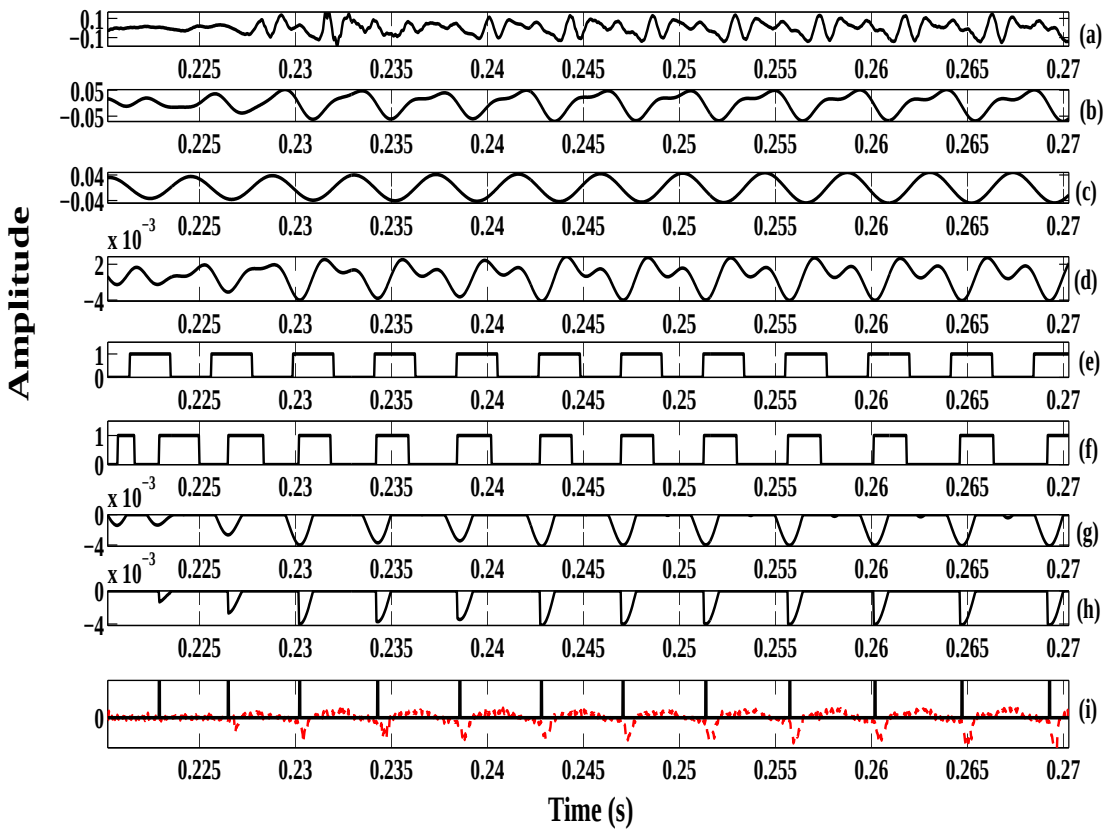

Figure 4.2: (a) Clean female voiced speech segment (b) LFR filtered voiced speech segment (c) Extracted time-varying F_0 component ($x_{1,F_0}[n]$) (d) Differenced LFR filtered voiced speech segment (e) $\chi[n]$ depicting the intervals marked by negative cycles of $x_{1,F_0}[n]$ (f) $\Gamma[n]$ depicting the intervals marked by negative cycles of the LFR filtered voiced speech segment (g) $\vartheta[n]$ depicting only negative cycles of the differenced LFR filtered voiced speech segment (e) GCI detection signal $\tilde{\vartheta}[n]$ whose local minima are detected as GCI candidates (i) GCIs identified in solid line and the DEGG signal in dashed line.

rate of 19.98 kHz. The white noise signal was acquired by sampling a high quality analog noise generator. The source of the babble noise signal was 100 people speaking in a canteen. The noise signals have been resampled to 32 kHz before adding them to speech signals at various SNRs ranging from 0 dB to 20 dB.

Existing methods used for performance comparison: The proposed method is compared with two existing GCI identification methods which are shown to be robust to different noisy environments in [43] namely, ZFR based method [6] and SEDREAMS method [24].

(a) ZFR based method: In this method [6], the speech signal is passed through a cascade

of two ideal 0 Hz resonators after removing any DC bias which may be present in it. In order to remove the large mean value from the filtered output of the cascade of two resonators, the local mean is computed and subtracted from the filtered output, this process is called as 'trend removal' operation [6] and is repeated three times to extract the zero frequency filtered signal. The positive zero crossings of zero frequency filtered signal are indicative of GCIs. The length of the window used to compute the local mean of the filtered output of the cascade of two Hz resonators should be in the range of one to two times of the average pitch period. We have chosen a window length equal to 1.5 times of the average pitch period.

(b) SEDREAMS method: In this method [24], firstly, a mean based signal is computed to determine the intervals in the voiced speech signal where GCIs may be present. The precise GCI positions within the intervals are obtained by locating discontinuities in the LP residual. The length of the window used to compute the mean based signal should be one to two times of the average pitch period in order to achieve high identification rate. A length of window equal to 1.75 times of the average pitch period has been used while carrying out experiments.

Performance evaluation criteria: In order to perform a quantitative performance assessment of the proposed method for GCI identification and enable its performance comparison with existing methods, we have used the performance measures defined in [23]. In [23], the larynx cycle was defined as the range of samples: $\frac{n_{l-1}+n_l}{2} \leq n_l < \frac{n_l+n_{l+1}}{2}$, given a reference GCI at sample n_l with preceding and following reference GCIs at samples n_{l-1} and n_{l+1}. The following are the measures defined in [23] which have been used to evaluate the performance of GCI identification methods:

(a) Identification rate (IDR): It is the percentage of larynx cycles for which exactly one GCI are detected.

(b) Miss rate (MR): It is the percentage of larynx cycles for which no GCI is detected.

(c) False-alarm rate (FR): It is the percentage of larynx cycles for which more than one GCI are detected.

(d) Identification accuracy (IDA): It is the standard deviation of the timing error be-

tween the reference and detected GCI locations. The timing error is calculated for larynx cycles for which exactly one GCI is detected.

(e) Accuracy to ± 0.25 ms (AR): It is the percentage of GCI locations detected with a timing error of ± 0.25 ms.

Reference GCIs are located as local minima in the DEGG signal. The following subsections presents the comparison and discussion of results of the proposed and existing GCI identification methods in clean and noisy environments.

4.3.1 Clean environment

The performance evaluation of the proposed GCI identification method and two existing methods on male and female speech signals of the CMU-Arctic database are tabulated in Table 4.1 and Table 4.2. It can be deduced from Table 4.1 and Table 4.2 that the proposed method has performed better than the existing methods in terms of IDR, MR, FR and IDA for both male and female speech signals. For male speech signals, a marginal improvement of 0.66% and 1.47% has been achieved by the proposed method in the value of IDR in comparison to SEDREAMS and ZFR based methods respectively. A significant reduction of 69.39% and 86.61% has been obtained by the proposed method in the value of MR in comparison to SEDREAMS and ZFR based methods respectively, for male speech signals. The value of FR obtained for the proposed method is at least 40% lower than those obtained by existing methods for male speech signals. The best AR results for male speech signals has been obtained by the SEDREAMS method, achieving an improvement of nearly 7.67% and 13.37% in comparison to the proposed and ZFR based methods respectively.

A slight improvement of 0.63% and 1.28% has been obtained by the proposed method in the value of IDR, computed for female speech signals, in comparison to SEDREAMS and ZFR based methods respectively. The proposed method has achieved the lowest possible value of MR for female speech signals. The proposed method has obtained a considerable improvement of nearly 70.83% and 79.1% in the value of FR, computed for female speech signals, in comparison to the SEDREAMS and ZFR based methods. The best AR results for female speech signals has been obtained by the SEDREAMS method,

Table 4.1: Performance comparison of GCI identification methods on clean male speech signals

Method	IDR (%)	MR (%)	FR (%)	IDA (ms)	AR (%)
Proposed	99.40	0.15	0.45	0.28	73.77
ZFR	97.96	1.12	0.92	0.39	70.06
SEDREAMS	98.75	0.49	0.76	0.33	79.43

Table 4.2: Performance comparison of GCI identification methods on clean female speech signals

Method	IDR (%)	MR (%)	FR (%)	IDA (ms)	AR (%)
Proposed	99.86	0.00	0.14	0.22	76.65
ZFR	98.60	0.73	0.67	0.33	73.21
SEDREAMS	99.24	0.28	0.48	0.27	83.34

achieving an improvement of nearly 8.73% and 13.84% in comparison to the proposed and ZFR based methods respectively.

The histograms of the timing error between the reference and the identified GCI locations obtained by the proposed method and two existing GCI identification methods are depicted in Fig. 4.3. It can be inferred from Fig. 4.3 that the standard deviation of the timing error and hence the value of IDA obtained by the proposed method is less than those obtained by SEDREAMS and ZFR based methods, in conformance to the IDA results compiled in Table 4.1 and Table 4.2. The reason of the excellent performance obtained by the proposed method is the employment of the time-varying F_0 component to reliably provide coarse estimate of intervals where GCIs are likely to occur. The proposed method performed marginally better for clean female speech signals in comparison to clean male speech signals because owing to the lower F_0 range of male speech signals, the LFR filtered male speech signal contains more number of harmonic components in comparison to the LFR filtered female speech signal. This increases the difficulty in reliable extraction of the time-varying F_0 component of a male speech signal for weakly voiced segments and accounts for a lower performance of the proposed method on male speech signals than female speech signals.

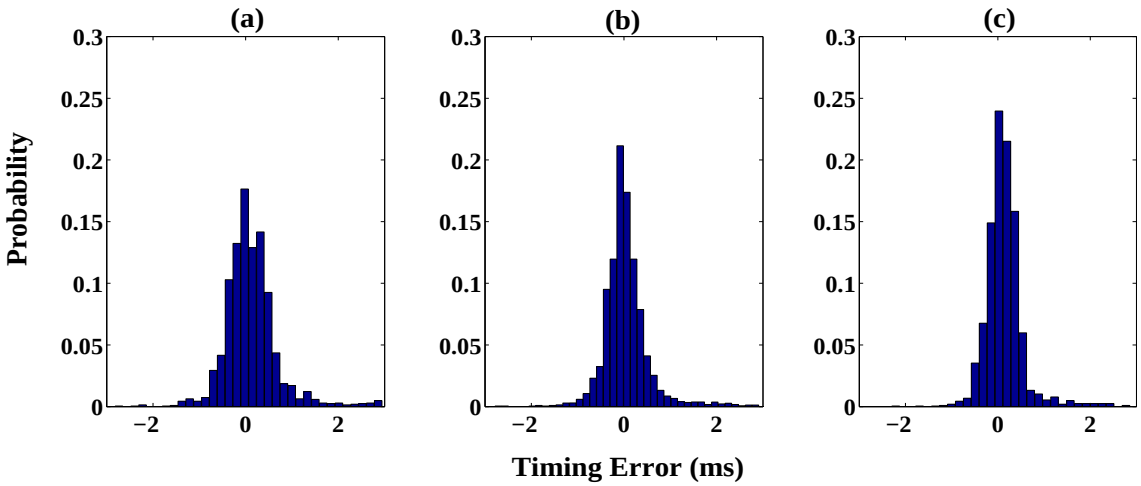

Figure 4.3: Histogram of the timing error between the identified and reference GCIs in a clean environment obtained by (a) ZFR based method (b) SEDREAMS method (c) Proposed method.

4.3.2 Noisy environment

The experimental results of the proposed GCI identification method obtained for a noisy male voiced speech signal at 0 dB SNR in white and babble noise environments are depicted in Fig. 4.4 and Fig. 4.5 respectively. The identification rate (IDR) of 100% and 90.19% and identification accuracy (IDA) of 0.28 ms and 0.45 ms are obtained in white and babble noise environments respectively. Please note in Fig. 4.5 that GCI identification errors mainly occurred in a weakly voiced segment of the voiced speech signal at around 0.8 s and 0.82 s because of the extraction of an erroneous component by the iterative algorithm of section 3.4.4 at 0 dB SNR in the babble noise environment.

The experimental results of the proposed GCI identification method obtained for a noisy female voiced speech signal at 0 dB SNR in white and babble noise environments are depicted in Fig. 4.6 and Fig. 4.7 respectively. The identification rate (IDR) of 100% and 100% and identification accuracy (IDA) of 0.22 ms and 0.55 ms are obtained in white and babble noise environments respectively. These results demonstrate the noise resilience of the proposed method. The proposed method performed fairly well even for severely degraded conditions.

The performance comparison of the proposed GCI identification method with two

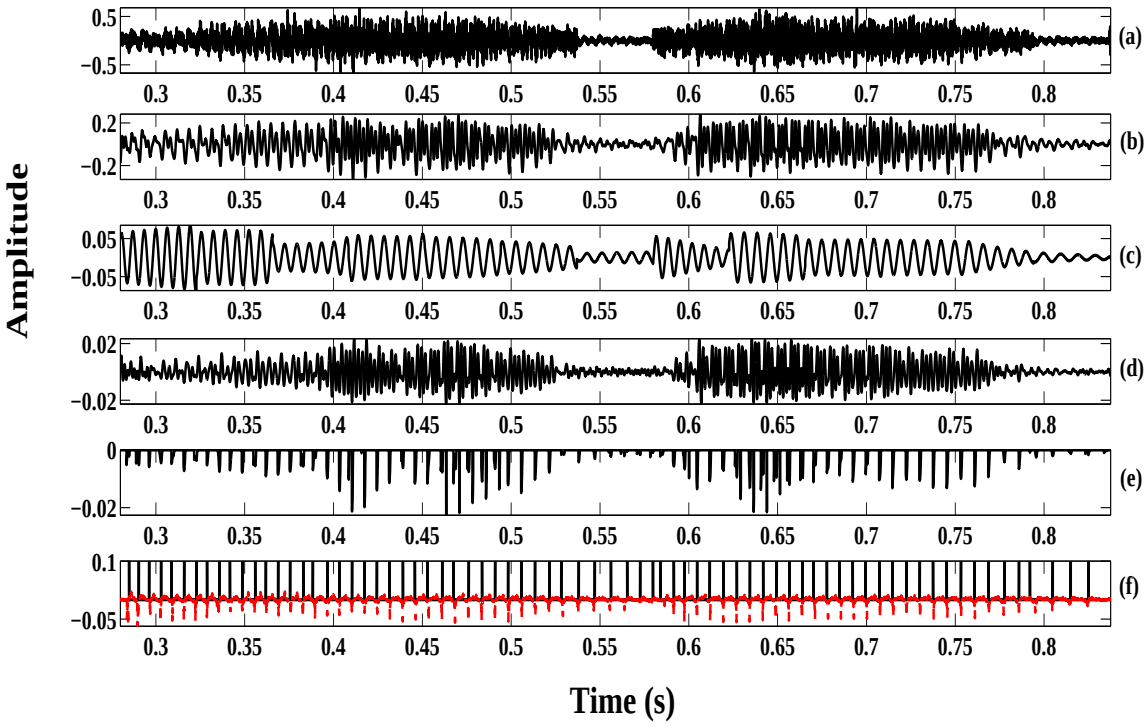

Figure 4.4: (a) Noisy male voiced speech signal at 0 dB SNR in a white noise environment (b) LFR filtered noisy voiced speech signal (c) Extracted time-varying F_0 component ($x_{F_0}[n]$) (d) Differenced LFR filtered noisy voiced speech signal (e) GCI detection signal $\tilde{\vartheta}[n]$ whose local minima are GCI candidates (f) GCIs identified in solid line and DEGG signal in dashed line.

existing methods on speech signals of the CMU-Arctic database in white and babble noise environments at a SNR range of 0 dB to 20 dB are shown in Fig. 4.8 and Fig. 4.9 respectively. In the white noise environment (Fig. 4.8), the proposed GCI identification method based on the extraction of the time-varying F_0 component of a voiced speech signal has performed noticeably better than the existing methods in terms of IDR, MR, FR and IDA. The proposed method remained robust to noise even at low SNRs and its performance did not degrade much with a reduction in the SNR, even when the SNR dropped to 0 dB value. The SEDREAMS method was able to detect a large percentage of GCIs with a higher accuracy than the proposed method and ZFR based methods, as reflected in the AR results obtained by the SEDREAMS method. Over the SNR range from 5 dB to 20 dB, the AR results obtained by the SEDREAMS method are better than the proposed and ZFR based methods. However, over the SNR range from 0 dB to 5 dB,

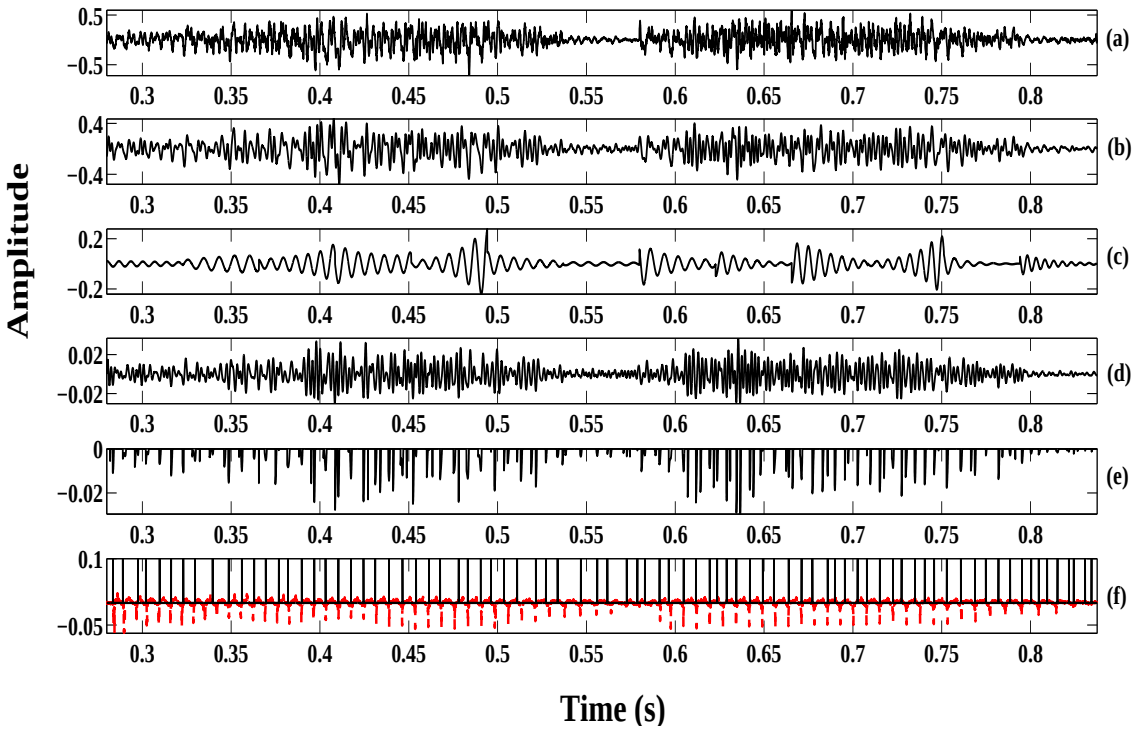

Figure 4.5: (a) Noisy male voiced speech signal at 0 dB SNR in a babble noise environment (b) LFR filtered noisy voiced speech signal (c) Extracted time-varying F_0 component ($x_{F_0}[n]$) (d) Differenced LFR filtered noisy voiced speech signal (e) GCI detection signal $\vartheta[n]$ whose local minima are GCI candidates (f) GCIs identified in solid line and DEGG signal in dashed line.

the proposed method has obtained better results for the AR than the SEDREAMS and ZFR based methods.

In the babble noise environment (Fig. 4.9), the values of the performance parameters: IDR, MR, FR obtained by the proposed method are better than the existing methods for the SNR range from 10 dB to 20 dB. However, over the SNR range from 0 dB to 5 dB, the proposed method performed similar to the existing methods in terms of IDR, MR, FR. The proposed method has provided the best MR results over the SNR range from 3 dB to 20 dB. Over the SNR range from 0 dB to 3 dB the SEDREAMS method has provided better MR results than the proposed and ZFR based methods. In terms of IDA, the proposed method performed better than the existing methods throughout the entire SNR range from 0 dB to 20 dB. The SEDREAMS method provided better AR results than the proposed and ZFR based methods over the SNR range from 10 dB to 20

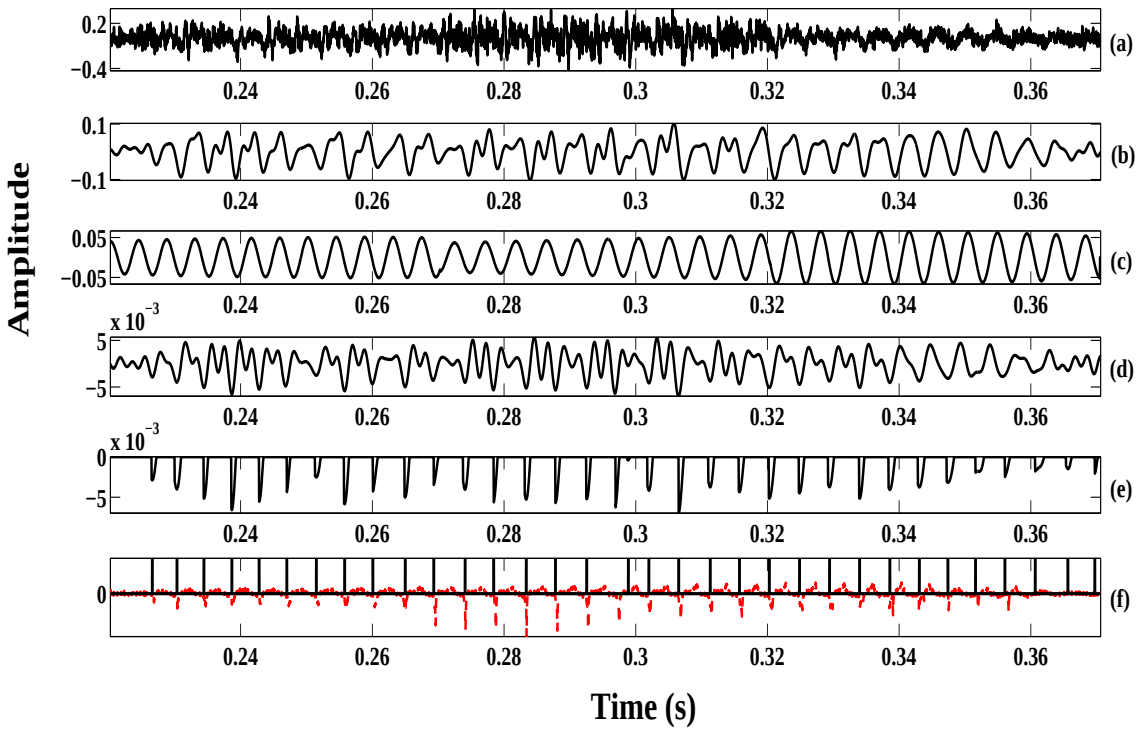

Figure 4.6: (a) Noisy female voiced speech signal at 0 dB SNR in a white noise environment (b) LFR filtered noisy voiced speech signal (c) Extracted time-varying F_0 component ($x_{F_0}[n]$) (d) Differenced LFR filtered noisy voiced speech signal (e) GCI detection signal $\tilde{\vartheta}[n]$ whose local minima are GCI candidates (f) GCIs identified in solid line and DEGG signal in dashed line.

dB. Over the SNR range from 0 dB to 10 dB, the proposed method obtained better AR results than the existing methods.

4.4 Conclusion

The proposed GCI identification method employs the extracted time-varying F_0 component to provide reliable coarse estimate of intervals where GCIS are likely to occur, which accounts for the high identification rate obtained by the proposed method. The high accuracy obtained by the proposed method is attributed to locating GCI candidates as local minima in the derivative of the falling edges of the isolated negative cycles of the LFR filtered voiced speech signal. The proposed method neither requires modeling or characterization of the vocal tract system nor assumes the voiced speech signal to be

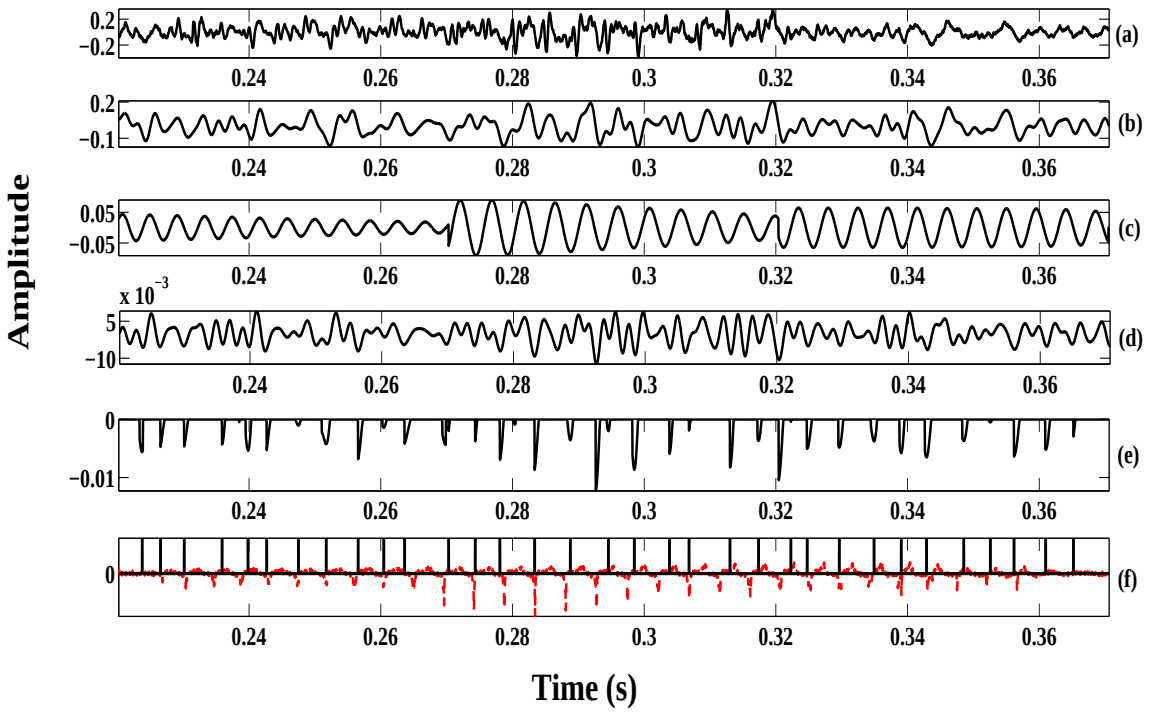

Figure 4.7: (a) Noisy female voiced speech signal at 0 dB SNR in a babble noise environment (b) LFR filtered noisy voiced speech signal (c) Extracted time-varying F_0 component ($x_{F_0}[n]$) (d) Differenced LFR filtered noisy voiced speech signal (e) GCI detection signal $\vartheta[n]$ whose local minima are GCI candidates (f) GCIs identified in solid line and DEGG signal in dashed line.

stationary for short duration.

It has been shown that the proposed GCI identification method is resilient to different noise environments and has provided high identification rate and accuracy, even when voiced speech signal is severely degraded by noise. The noise resilience of the proposed method is attributed to the use of a robust algorithm for extraction of the time-varying F_0 component of a voiced speech signal. The filtering of the voiced speech signal in the LFR eliminates the noise energy outside the LFR and also attenuates formant components. The employed selection criterion efficiently discards false GCI candidates and ensure that only one GCI is identified per negative cycle of the extracted time-varying F_0 component of the voiced speech signal.

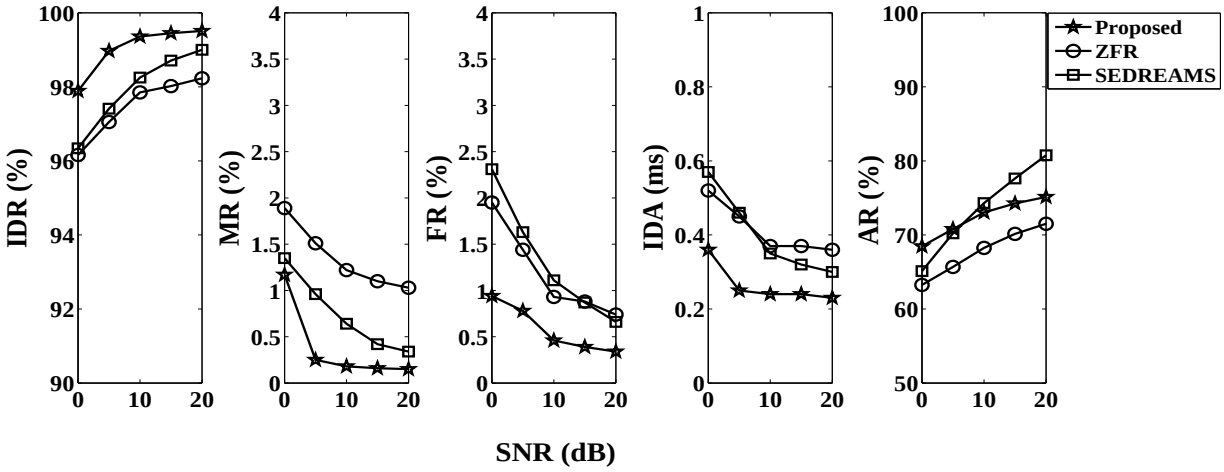

Figure 4.8: Performance comparison of GCI identification methods in white noise environment.

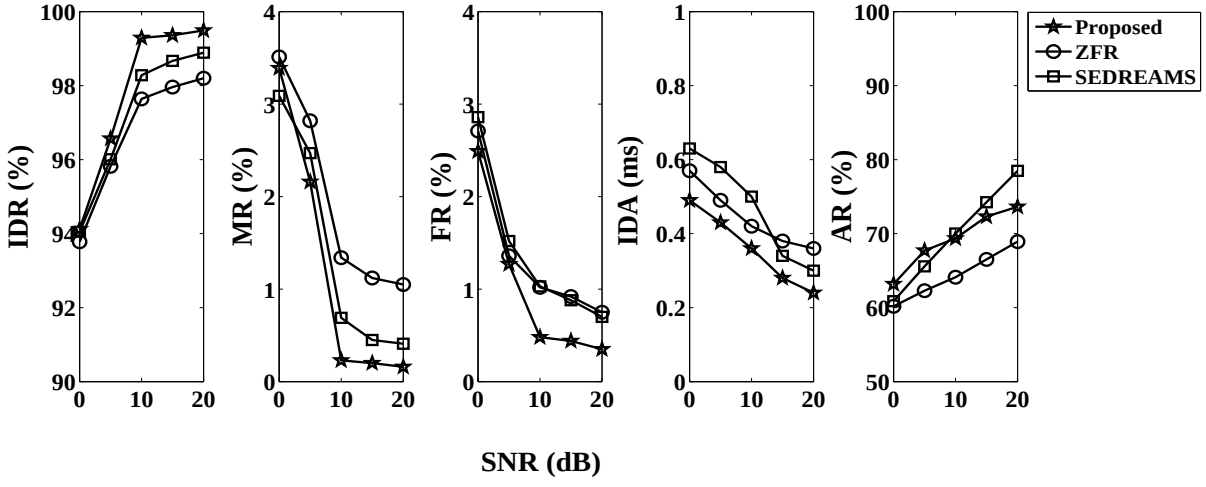

Figure 4.9: Performance comparison of GCI identification methods in babble noise environment.

Chapter 5

Estimation of Instantaneous Fundamental Frequency

This chapter presents a robust event-based method for estimation of the instantaneous fundamental frequency of a voiced speech signal. The time-varying F_0 component of a voiced speech signal is extracted by the robust algorithm of subsection 3.4.4 that iteratively performs eigen value decomposition (EVD) of the Hankel matrix, initially constructed from the samples of the LFR filtered voiced speech signal. The negative cycles of the extracted time-varying F_0 component provide reliable coarse estimate of the intervals where the glottal closure instants (GCIs) are likely to occur. The negative cycles of the LFR filtered voiced speech signal occurring within these intervals are isolated. GCIs are detected as local minima in the derivative of the falling edges of the isolated negative cycles of the LFR filtered voiced speech signal, followed by a selection criterion to discard false GCI candidates. The instantaneous F_0 is estimated as the inverse of the time interval between two consecutive GCIs. Experiments were performed on Keele and CSTR speech databases in white and babble noise environments at various levels of degradation to assess the performance of the proposed method. The proposed method substantially reduces the gross F_0 estimation errors in comparison to some state of the art methods.

5.1 Introduction

Voiced speech is a quasi-periodic signal produced by excitation of the vocal tract system by quasi-periodic puffs of air [1]. The excitation of the vocal tract system is maximum at the instant of closure of the glottis (GCI). The fundamental frequency (F_0) of the voiced

speech signal is one of the most important acoustic parameters which is defined as the rate of vibration of the vocal folds during the production of voiced speech. It is a time-varying quantity which also depends on the gender, emotion, language, accent, age and health condition of the speaker. Pitch is a subjective psychoacoustical attribute of voiced speech and its human perception corresponds closely to F_0 of the voiced speech signal. An accurate estimate of the instantaneous F_0 is required by various pitch synchronous speech signal processing algorithms used in speech compression, text-to-speech synthesis, voice conversion and expressive speech synthesis [12, 14–16]. Prosodic features derived from the variations in F_0 find use in applications such as speaker recognition and emotion recognition [17, 89]. Some speech signal processing applications like speech enhancement, speech recognition, emotion recognition, diagnosis of pathological voice disorders [11, 13, 20, 89, 90] require a reliable estimate of the instantaneous F_0 from noisy speech signals.

A number of F_0 estimation algorithms have been reported in the literature. They can be broadly categorized into three classes: block-based, instantaneous and event-based methods. An extensive review of these methods can be found in [91–93]. Block-based methods based on the short-time average magnitude difference function (AMDF), autocorrelation function (ACF), cepstrum, simplified inverse filter tracking (SIFT), modulation model or subharmonic summation [3, 94–98] divide the voiced speech signal into many segments, assuming it to be stationary within the segment duration. An estimate of F_0 is obtained for each segment. Only a few block-based methods based on weighted autocorrelation (WAC), dominant harmonic components, harmonic sinusoidal autocorrelation model (HSAC) and autocorrelation pitch detector [99–102] considered F_0 estimation from noisy speech signals. The main limitation of these block-based methods lie in their inability to track F_0 variations occurring within the segment, spanning a few glottal cycles. Instantaneous methods estimate the F_0 value at each sample instant of the voiced speech signal. Instantaneous F_0 methods based on the filtering of the speech signal using a bank of two band pass filters, modeling of the time-varying F_0 using B-spline expansion, the Hilbert-Huang transform (HHT) or ensemble empirical mode decomposition (EEMD) were proposed in [21, 22, 26, 44]. However, these methods are sensitive to noise.

On the other hand, event-based methods as proposed in [10, 45, 46, 103] mark the

occurrence of a characteristic event in each glottal cycle such as the GCI, and F_0 is computed as the inverse of the time interval between successive GCIs. The time resolution of a glottal cycle is sufficient for F_0 estimation in order to describe the continuous variation in F_0 [44]. The presence of spurious peaks in the GCI determination signal in the presence of white noise causes ambiguity in detecting the true GCIs [45]. The method in [46] was based on the concept that a local maximum occurs across various scales in the wavelet transform (WT) of the voiced speech signal around a GCI. The limitation of using a heuristic approach in [46] for finding local maxima in the WT was overcome in [103] by the use of an optimization scheme. However, performance of methods in [46, 103] was not evaluated on large speech databases that contain reference F_0 values determined using a simultaneously recorded EGG signal. In [10], the final F_0 contour was obtained from the positive zero crossings of the zero frequency resonator (ZFR) filtered signals derived from the voiced speech signal and its Hilbert envelope respectively. A noise resilient GCI detection method presented in [24] was based on the computation of a mean based signal from the voiced speech signal to extract intervals where GCIs may be present. GCIs were located within the extracted intervals by determining the instants of discontinuity in the linear prediction (LP) residual. The methods in [10, 24] were shown to be robust to noise.

In the last chapter of this thesis, a novel GCI identification method has been presented which provided better identification rate and accuracy than the existing methods [10, 24] in clean and noisy environments. It is based on the extraction of the time-varying F_0 component of a voiced speech signal using the iterative algorithm of subsection 3.4.4. The event-based instantaneous F_0 method proposed in this chapter estimates the instantaneous F_0 as the inverse of the time interval between two consecutive GCIs, identified using the GCI identification method of Section 4.2. This chapter is organized as follows: The proposed event-based method for estimating the instantaneous F_0 is presented in Section 5.2. The quantitative performance evaluation of the proposed method on Keele and CSTR speech databases in clean, white and babble noise environments is presented in Section 5.3. This section also presents a comparison of the experimental results obtained using the proposed method with those obtained by some state of the art methods. The chapter is concluded in Section 5.4.

5.2 Proposed Event Based Instantaneous Fundamental Frequency Method

The detection of voiced regions is a prerequisite for estimation of the instantaneous F_0 of voiced regions. The V/NV detection method proposed in the second chapter of this thesis is used to detect the voiced regions of a speech signal. The detected voiced regions are filtered in the LFR using the FB coefficients as described in subsection 3.4.1 to render their time-varying F_0 components discernible among other harmonic components and to remove the noise energy present outside the LFR. Let $x[n]$ denote a detected voiced region of a speech signal. Let $y[n] = x_{\mathrm{LF}}[n]$, where $x_{\mathrm{LF}}[n]$ represents the LFR filtered noisy voiced speech signal given by (3.31). $y[n]$ is divided into L equal size segments $y_l[n], l = 1, 2, ..., L$ as described in subsection 3.4.2. The time-varying F_0 component of $y[n]$ is extracted using the iterative algorithm of subsection 3.4.4 of this thesis. Let $y_{F_0}[n]$ and $y_{l,F_0}[n]$ represent the extracted time-varying F_0 component of $y[n]$ and $y_l[n]$ respectively. Only one GCI can occur in the interval marked by a negative cycle of $y_{F_0}[n]$. Thus, the negative cycles of $y_{F_0}[n]$ provide a reliable coarse estimate of intervals where GCIs may occur [39].

The GCI identification method based on the extracted time-varying F_0 component of a voiced region (detailed in Section 4.2) is used to identify GCIs in $x[n]$. The GCI identification method of Section 4.2 isolates the negative cycles of $y[n]$ that are occurring within the intervals marked by the negative cycles of $y_{F_0}[n]$. The GCI candidates are detected as local minima in the derivative of the falling edges of the isolated negative cycles of $y[n]$. However, sometimes more than one negative cycle of $y[n]$ are contained in the interval marked by a negative cycle of $y_{F_0}[n]$. Moreover, the shape of the falling edge of a isolated negative cycle of $y[n]$ sometimes give rise to more than one local minimum per negative cycle of $y_{F_0}[n]$. This results in detection of many false GCI candidates. Out of all GCI candidates detected in the interval marked by a negative cycle of $y_{l,F_0}[n]$, a selection criterion is used in the GCI identification method of Section 4.2 to select the GCI candidate corresponding to the peak negative cycle of $y_l[n]$ occurring in the interval. Thus, the selection criterion ensures that only one GCI candidate is identified per negative

cycle of $y_{F_0}[n]$.

The proposed event-based instantaneous F_0 estimation method computes the instantaneous F_0 as the inverse of the time interval between successive GCIs identified in $x[n]$. The interpolated F_0 contour is obtained at intervals of 0.1 ms by linearly interpolating the values of instantaneous F_0 computed at the time-instants of occurrence of respective GCIs. The F_0 contour value at 10 ms intervals is obtained by downsampling the interpolated F_0 contour every 10 ms. Please note that these operations are required only for comparison with the reference F_0 values provided in the Keele speech database [104–106] at 10 ms intervals. The next section quantitatively evaluates the performance of the proposed event-based instantaneous F_0 estimation method and presents a comparison of the performance of the proposed method with some state of the art methods.

5.3 Quantitative Performance Evaluation and Comparison

5.3.1 Speech signal databases

The speech signals of Keele and CSTR databases [104–106] have been used for the performance evaluations of the proposed method and some state of the art methods.

(a) *Keele Database*: It consists of phonetically balanced utterances of about 34 s duration each, spoken by five female and five male speakers in the English language, sampled at a rate of 20 kHz. The simultaneously recorded EGG signal of an utterance was divided into frames of 25.6 ms duration, and the reference N_0 value for a frame was obtained by determining the peak in the respective ACF, at a frame rate of 100 Hz. The database provides reference fundamental period values in samples (N_0) at 10 ms intervals. Positive, negative and zero reference N_0 values are indicated for certain voiced frames, uncertain frames and unvoiced frames respectively. We have computed reference F_0 values as F_s/N_0 for voiced frames.

(b) *CSTR Database*: It is approximately of 7 min duration and consists of 50 English sentences spoken by a male and a female speaker. The speech signals were sampled at

20 kHz. The reference F_0 values are provided in the database at the time-instants of occurrence of respective GCIs, identified in the first-order derivative of the simultaneously recorded EGG signal. The reference F_0 contour at 10 ms intervals by linearly interpolating the reference F_0 values at 0.1 ms intervals, followed by a downsampling operation at 10 ms intervals.

5.3.2 Noise database

The white and babble noise signals are taken from the NOISEX-92 database [71]. The white noise signal was acquired by sampling a high quality analog noise generator. The source of the babble noise signal was 100 people speaking in a canteen. These noise signals are available at F_s of 19.98 kHz and therefore, have been resampled to 20 kHz before adding them to speech signals.

5.3.3 Existing pitch frequency estimation methods

The performance of the proposed method in clean and noisy environments has been compared with the performance of four existing methods namely: ZFR based method [6], YIN [107], Praat's autocorrelation (AC) method [108] and the method based on the iterative algorithm of [4].

(a) *ZFR based method*: In this method, the final F_0 contour is derived from the positive zero crossings of ZFR filtered signals derived from the voiced speech signal and its Hilbert envelope respectively. During experiments, we have used a window length of 1.5 times of the average fundamental period ($\hat{T}_0$) for the trend removal operation [6].

(b) *YIN*: This method introduced a few modifications to the conventional ACF; e.g., a difference function formulation, normalization and parabolic interpolation for attenuation of the secondary peaks in the ACF at F_0 harmonics to reduce errors in the estimated fundamental period (T_0) [107]. We have downloaded the YIN software from the author$'$s homepage.

(c) *Praat's autocorrelation (AC) method*: In this method [108], the ACF of the windowed

voiced segment is divided by the ACF of the window function to reduce windowing artifacts. A sinc interpolation is carried out around the local maxima corresponding to T_0 to overcome the artifacts introduced because of sampling of the continuous-time speech signal.

(d) *Method based on the iterative algorithm of* [4]: This method is based on the identification of GCIs by employing the time-varying F_0 component of a voiced speech signal extracted using the iterative algorithm of [4]. The instantaneous F_0 is estimated as the inverse of the time interval between two consecutive GCIs.

While evaluating the performance, the F_0 search range for all F_0 estimation methods has been set to 50 Hz - 500 Hz.

5.3.4 Performance evaluation criteria

The following performance evaluation measures have been used to determine the efficacy of the F_0 estimation methods [10]:

(a) *Gross Error Percentage* (GEP): It is the percentage of voiced frames (10 ms duration) with estimated F_0 values deviating from the reference F_0 values by more than 20%.

(b) *Mean Error* (ME): It is the mean of the absolute value of the difference between the estimated and reference F_0 values in Hz.

(c) *Standard Deviation* (SD): It is the standard deviation of the absolute value of the difference between the estimated and the reference F_0 values in Hz.

Please note that the gross errors have not been considered in the calculation of ME and SD.

5.3.5 Clean environment

The experimental results on the clean voiced speech signal demonstrated in Fig. 5.1 shows that the proposed event-based method, method based on the iterative algorithm of [4] and the YIN method were excellent in estimating the value of F_0 for the entire duration of the voiced speech signal (Fig. 5.1 (a)) while the ZFR based and AC methods made an error in estimating the F_0 at 0.54 s and 0.99 s respectively. A more comprehensive

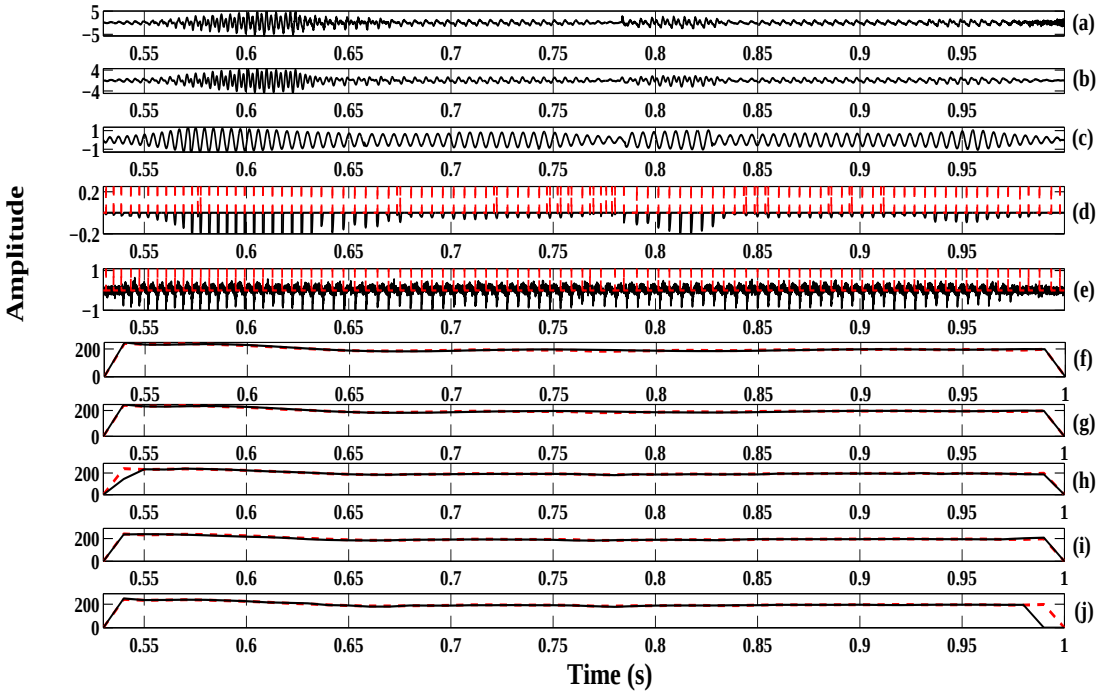

Figure 5.1: (a) Clean voiced speech signal (b) LFR filtered voiced speech signal (c) Extracted F_0 component using the proposed iterative algorithm of subsection 3.4.4 (d) GCI detection signal in solid line and the detected GCI candidates in dashed line (e) GCIs identified after applying the selection criterion in dashed line and the DEGG signal in solid line (please refer to Section 4.2). Estimated F_0 contour in solid line using (f) Proposed event-based method (g) Method based on iterative algorithm of [4] (h) ZFR based method (i) YIN method (j) Praat's AC method. Reference F_0 contour shown in dashed lines in (f)-(j).

and quantitative performance comparison of different methods on male and female speech signals of the two speech databases are presented in Table 5.1 and Table 5.2. It can be inferred from Table 5.1 and Table 5.2 that the proposed method provides the lowest GEP among all methods for both male and female speech signals. The proposed method has achieved at least 50% reduction in the GEP for both male and female speech signals in comparison to the ZFR based, YIN and Praat's AC methods. The substantial reduction in the GEP achieved by the proposed method is due to the employment of the extracted time-varying F_0 component of the voiced speech signal to reliably and accurately identify GCIs. The GEP obtained by the proposed method for female speech signals is nearly 23% less than the GEP obtained for male speech signals. The reason is that a LFR filtered male speech signal usually contains more number of harmonic components than

Table 5.1: Comparison of performance of different F_0 estimation methods on clean male speech signals

	KEELE SPEECH DATABASE					
	MALE			**FEMALE**		
METHOD	GE (%)	ME (Hz)	SD (Hz)	GE (%)	ME (Hz)	SD (Hz)
Proposed	1.634	3.225	3.971	1.224	3.313	4.421
Method Based on Iterative Algorithm of [32]	1.915	3.614	4.018	1.361	3.539	4.662
ZFR Based	3.026	3.159	4.124	2.251	3.418	4.763
YIN	3.522	2.718	4.215	2.744	3.276	5.093
AC	5.617	2.162	3.628	4.926	2.815	4.379

Table 5.2: Comparison of performance of different F_0 estimation methods on clean female speech signals

	CSTR SPEECH DATABASE					
	MALE			**FEMALE**		
METHOD	GE (%)	ME (Hz)	SD (Hz)	GE (%)	ME (Hz)	SD (Hz)
Proposed	1.033	3.857	4.345	0.817	4.028	4.813
Method Based on Iterative Algorithm of [32]	1.213	4.119	4.561	0.880	4.206	5.184
ZFR Based	2.372	5.011	5.923	1.536	5.548	7.297
YIN	3.681	4.217	4.639	2.772	4.632	5.784
AC	5.823	2.916	5.011	5.104	3.592	5.930

a LFR filtered female speech signal. This presents difficulty in reliably extracting the F_0 component of weakly voiced regions of a male speech signal. The values of ME and SD obtained by the proposed method are commensurate with those obtained by the ZFR based and YIN methods. The values of ME obtained by the proposed method are greater than those obtained by Praat's AC method, especially for male speech signals. The reason is that Praat's AC method resulted in gross errors while estimating the F_0 for weakly and non quasi-periodic voiced frames. These gross errors have not been included while calculating the ME, which resulted in low values of ME. On the other hand, the estimated F_0 values by the proposed method resulted in high absolute errors for some weakly voiced segments of male speech signals but they do not come under the category of gross errors; i.e., the magnitude of such errors does not exceed 20% of the reference F_0 values and therefore, contributed towards an increase in overall ME values.

The Kruskal-Wallis test [109] when applied to the absolute F_0 estimation errors ob-

tained using the proposed method and the method based on the iterative algorithm of [4] results in p-values of 0.1440 and 0.6109 for male and female speech signals respectively. At a level of significance, denoted α of 0.05, the results obtained by the two methods are statistically insignificant. Thus, it can be concluded from Table II and statistical results that in the clean environment, the proposed method achieves similar performance as the method based on the iterative algorithm of [32]. The main advantage of using the proposed method over the method based on the iterative algorithm of [32] becomes apparent in the next section discussing the results obtained in noisy environments.

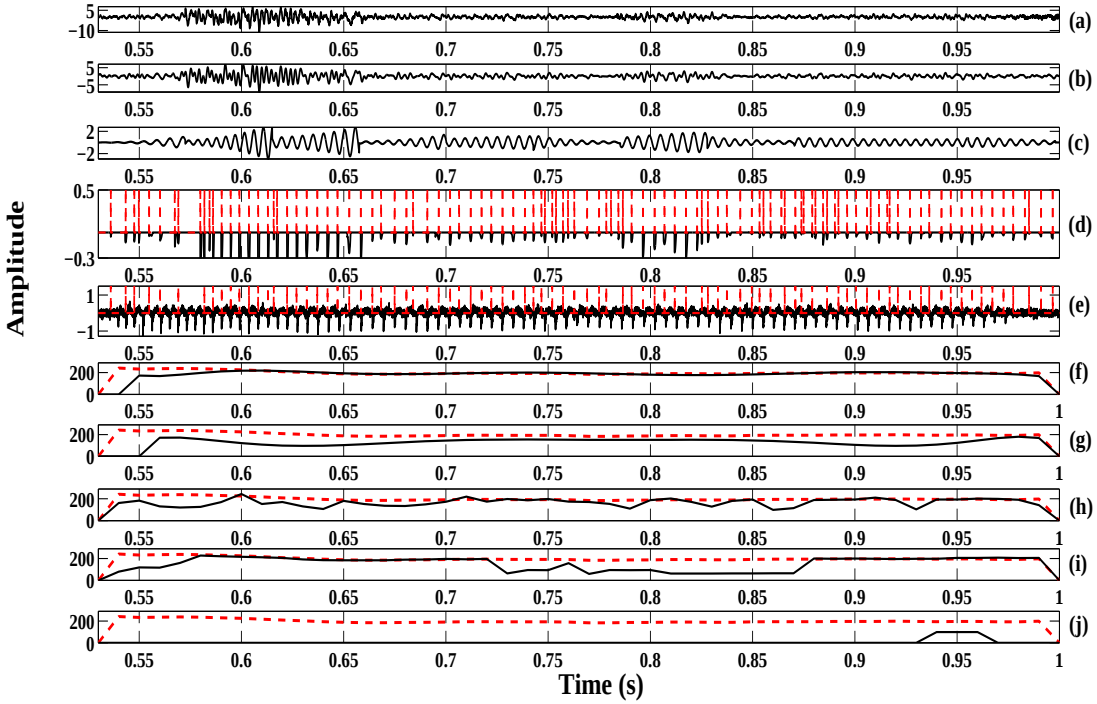

Figure 5.2: (a) Noisy voiced speech signal at 0 dB SNR in a babble noise environment (b) LFR filtered noisy voiced speech signal (c) Extracted F_0 component using the proposed iterative algorithm of subsection 3.4.4. (d) GCI detection signal in solid line and the detected GCI candidates in dashed line (e) GCIs identified after applying the selection criterion in dashed line and the DEGG signal in solid line (please refer to Section 4.2). Estimated F_0 contour in solid line using (f) Proposed event-based method (g) Method based on iterative algorithm of [4] (h) ZFR based method (i) YIN method (j) Praat's AC method. Reference F_0 contour in dashed lines in (f)-(j).

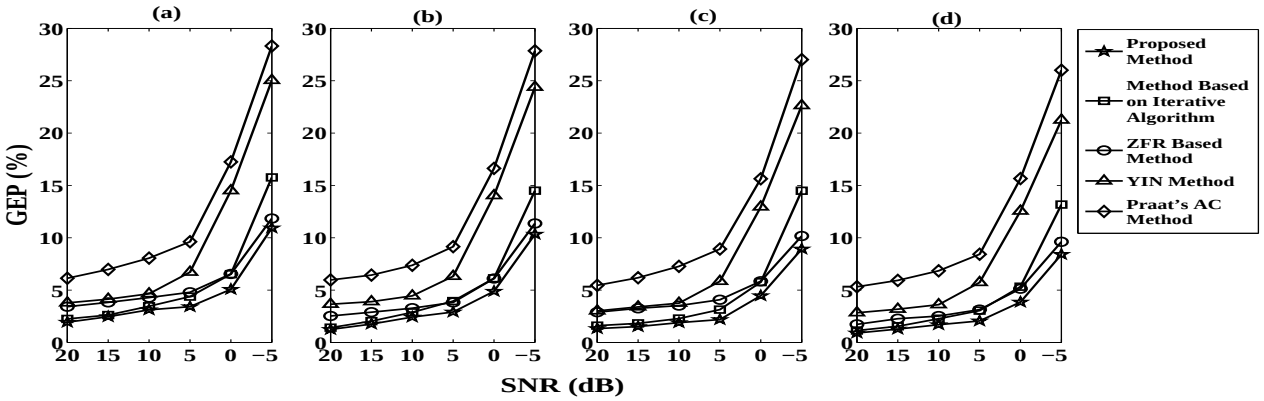

Figure 5.3: Performance comparison of different F_0 estimation methods in white noise environment in terms of GEP (%) on (a) Keele database male speech signals (b) CSTR database male speech signals (c) Keele database female speech signals (d) CSTR database female speech signals.

Table 5.3: p-values obtained by applying the Kruskal-Wallis test on absolute F_0 estimation errors obtained using the proposed event-based method and the method based on the iterative algorithm of [4] at different SNRs

	MALE		**FEMALE**	
NOISE ENVIRONMENT	**WHITE**	**BABBLE**	**WHITE**	**BABBLE**
SNR (dB)				
20	0.1123	0.5431	0.0731	0.0377
15	0.0130	0.0001	0.0562	0.0005
10	0.0027	0.0016	0.0252	0.0000
5	0.0000	0.0000	0.0047	0.0000
0	0.0001	0.0000	0.0001	0.0003
-5	0.0000	NA	0.0000	NA

5.3.6 Noisy environment

The experimental results of F_0 estimation at 0 dB SNR in a babble noise environment has been shown in Fig. 5.2. It is evident from the comparison of Fig. 5.2 (f) with Fig. 5.2 (h) and Fig. 5.2 (i) that the proposed method more reliably estimated the F_0 with the GEP value of 10.85% than the ZFR based method and the YIN method with the GEP values of 34.13% and 34.59% respectively. Praat's AC method and the method based on the iterative algorithm of [4] failed to estimate the F_0 in this scenario with the GEP values higher than 50%. The quantitative performance comparison of the proposed event-based method with existing methods in terms of the GEP obtained for male and female speech signals in white and babble noise environments at different levels of degradation are shown

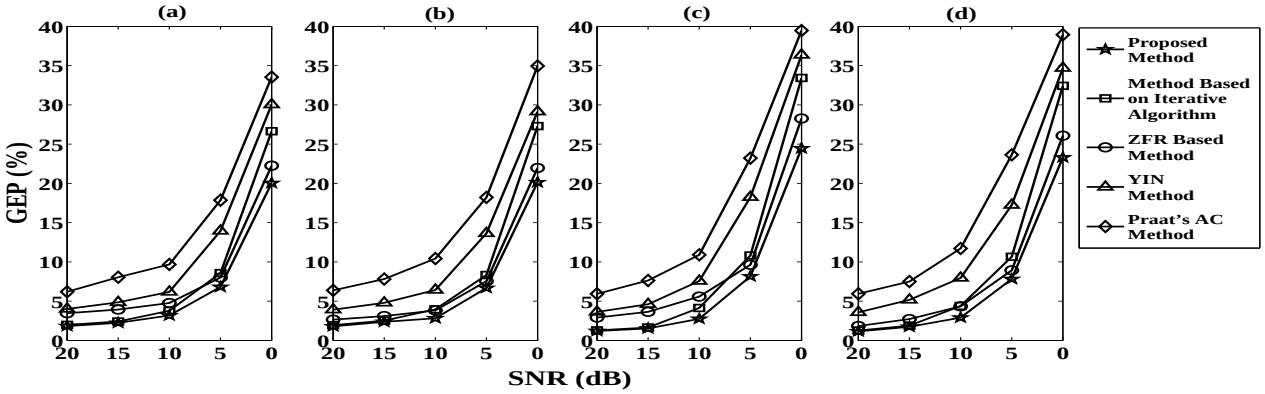

Figure 5.4: Performance comparison of different F_0 estimation methods in babble noise environment in terms of GEP (%) on (a) Keele database male speech signals (b) CSTR database male speech signals (c) Keele database female speech signals (d) CSTR database female speech signals.

in Fig. 5.3 and Fig. 5.4 respectively.

In the white noise environment, the proposed method has achieved a GEP reduction of nearly 28% and 36% on male and female speech signals respectively, on an average across the SNR range of -5 dB to 5 dB, in comparison to the ZFR based, YIN and Praat's AC methods. The proposed method has obtained a GEP reduction of approximately 12% and 19% on male and female speech signals respectively, in comparison to the method based on the iterative algorithm of [4], over the SNR range of 10 dB to 20 dB. A high GEP reduction of approximately 25% and 31% has been achieved by the proposed method on male and female speech signals respectively, in comparison to the method based on the iterative algorithm of [4], over the SNR range of -5 dB to 5 dB. The proposed method has obtained nearly 25% lower GEP values for female speech signals than male speech signals across the entire SNR range.

In the babble noise environment, a GEP reduction of nearly 37% and 46% on male and female speech signals respectively, has been achieved by the proposed method in comparison to the ZFR based, YIN and AC methods, over the SNR range of 10 dB to 20 dB. The proposed method has obtained roughly 12% GEP reduction on both male and female speech signals in comparison to the ZFR based, YIN and AC methods, over the SNR range of 0 dB to 5 dB. The proposed method has obtained a GEP reduction of around 7% for both male and female speech signals in comparison to the method based

on the iterative algorithm of [4], over the SNR range of 15 dB to 20 dB. A high GEP reduction of around 28% has been achieved by the proposed method for both male and female speech signals in comparison to the method based on the iterative algorithm of [4], over the SNR range of 0 dB to 10 dB.

The absolute F_0 estimation errors obtained using the proposed method and the method based on the iterative algorithm of [32] at different SNRs in white and babble noise environments for female and male speech signals are depicted in Fig. 5.5 and Fig. 5.6 respectively. The p-values obtained when the Kruskal-Wallis test [109] is applied on absolute F_0 estimation errors of the two methods are tabulated in Table 5.3. In Table 5.3, NA stands for not applicable and implies that the method described in subsection 2.3.2 is not able to efficiently detect voiced regions at the specified SNR. It can be inferred from Table 5.3 and Fig. 5.6 that at α equals to 0.05, the results obtained for male speech signals using the two methods are statistically significant over the SNR range of -5 dB to 15 dB in a white noise environment and 0 dB to 15 dB in a babble noise environment. It can be deduced from Table 5.3 and Fig. 5.5 that the results obtained by two methods for female speech signals are statistically significant over the SNR range of -5 dB to 10 dB in a white noise environment and 0 dB to 20 dB in a babble noise environment at α equals to 0.05.

The performance of the proposed method is worse in the babble noise environment than the white noise environment. The reason is that babble noise is correlated and introduces high energy components in the LFR, resulting in an increased probability of extraction of an erroneous component by the proposed iterative algorithm of subsection 3.4.4.

5.4 Conclusion

A robust event-based method for instantaneous F_0 estimation based on the GCI identification method of Section 4.2 has been presented in this chapter. The proposed method can efficiently track F_0 variations occurring at each glottal cycle and unlike the auto-correlation method, it does not require assumption of stationarity of the voiced speech

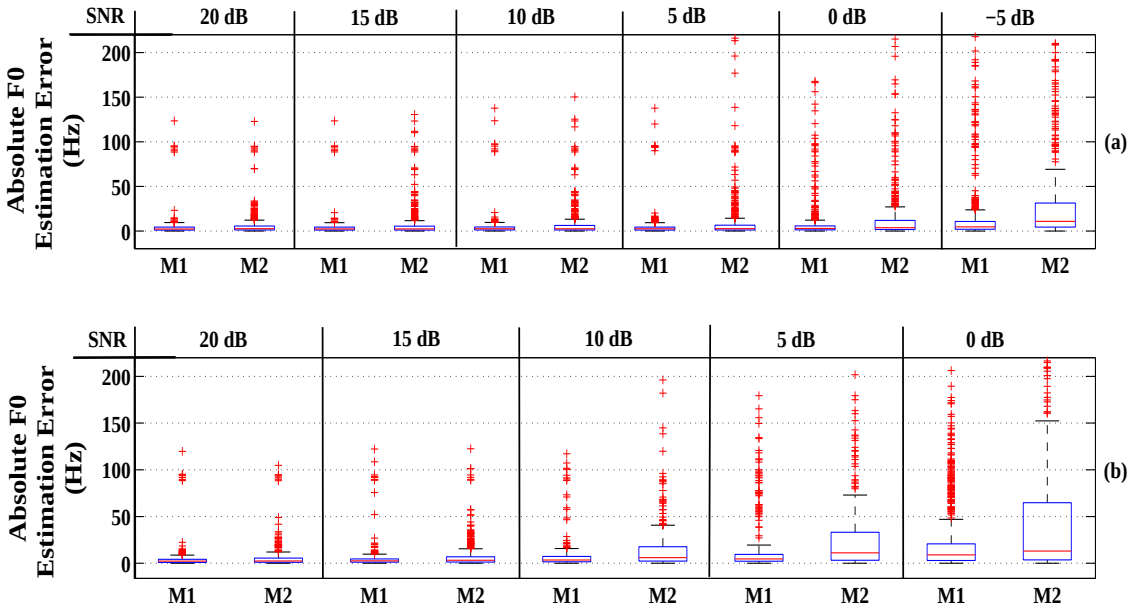

Figure 5.5: Absolute F_0 estimation errors for female speech signals obtained using the proposed event-based method (denoted by M1) and the method based on the iterative algorithm of [4] (denoted by M2) at different SNRs in a (a) white noise environment (b) babble noise environment.

signal over short duration. The proposed method achieves substantial reduction in the gross error percentage in comparison to some state of the art methods in different noise environments at various levels of degradation.

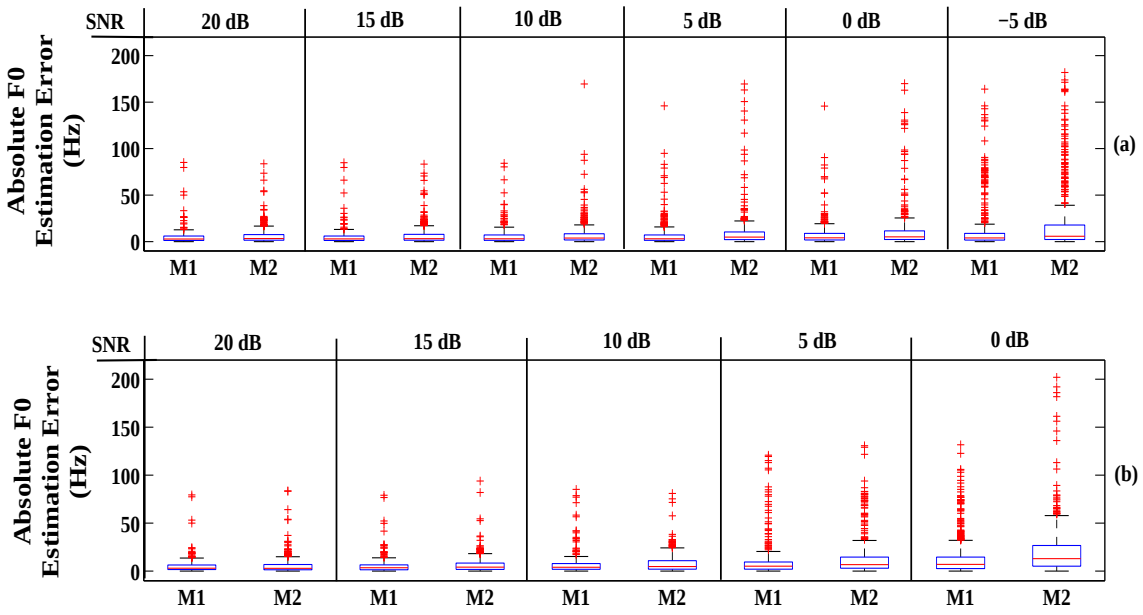

Figure 5.6: Absolute F_0 estimation errors for male speech signals obtained using the proposed event-based method (denoted by M1) and the method based on the iterative algorithm of [4] (denoted by M2) at different SNRs in a (a) white noise environment (b) babble noise environment.

Chapter 6

A Novel Iterative Approach for Decomposition and Analysis of Multi-component Non-stationary Signals

The decomposition of multi-component signals finds use in the signal analysis. This chapter presents a novel iterative approach to decompose a multi-component non-stationary signal into amplitude-frequency modulated (AM-FM) mono-component signals. The extracted AM-FM mono-component signals are narrowband signals whose instantaneous amplitudes and frequencies can be computed using DESA or Hilbert transform. The proposed iterative decomposition approach is based on repeatedly performing eigenvalue decomposition (EVD) of the Hankel matrix and extracting components corresponding to significant eigenvalue pairs of the Hankel matrix. The Hankel matrix is initially constructed from the samples of the multi-component non-stationary signal. The proposed iterative decomposition approach is adaptive and provides good frequency resolution over the entire frequency range. It has also been shown that unlike the EMD, the ability of the proposed iterative approach to separate constituent mono-component signals of a multi-component signal is neither affected by the ratio of their mean frequencies nor by their relative amplitudes.

6.1 Introduction

The extraction of mono-component signals from a multi-component signal is referred to as decomposition of the multi-component signal [110]. Many signals encountered in practice, such as speech signal, electroencephalogram (EEG) signal, seismic signal, phono-

cardiogram (PCG) signal, electrocardiogram (ECG) signal etc. are multi-component non-stationary signals. The decomposition of multi-component signal into mono-component signals finds use in the signal analysis [111–114], signal modeling and signal classification [115–117]. The Hilbert transform or discrete-energy separation algorithm (DESA) can be applied on mono-component signals extracted from a multi-component non-stationary signal to compute their instantaneous frequencies [12, 44, 118].

The frequency domain methods based on the Fourier transform are not suitable to analyze the time-varying nature of the non-stationary signal. Many time-frequency analysis techniques such as: short-time Fourier transform (STFT), continuous wavelet transform (CWT), Wigner-Ville distribution (WVD) facilitate the analysis of multi-component non-stationary signals. The STFT and CWT are linear transformations while the WVD is a quadratic transformation [119]. The limitation of the STFT is that it assumes piecewise stationarity of the non-stationary signal over the duration of the employed window function. The drawback of the STFT is that it suffers from fixed time-frequency resolution over the entire time-frequency plane. The time-frequency resolution is governed by the window function being used to compute the STFT. The Heisenberg's uncertainty principle states that infinitesimal time and frequency resolutions can not be simultaneously achieved by a window function [119]. The CWT offers more flexibility than the STFT and offers good time-resolution at high frequencies and good frequency resolution at low frequencies [120]. The optimal wavelet basis needs to be selected depending on the type of application and the signal under consideration [121, 122]. Once chosen, the type of the wavelet basis is kept fixed for the analysis of data. However, the CWT is unable to resolve closely or moderately spaced mono-component signals in the high frequency range. The WVD provides optimum time and frequency resolutions over the entire time-frequency plane. However, the WVD of a multi-component signal suffers from cross-terms occurring at mid-time and mid-frequency of auto-components of the WVD of the multi-component signal and can have significant magnitudes. These cross-terms arise from the quadratic nature of the WVD. These artifacts obscure the analysis of the multi-component signal [119]. Various methods based on exponential kernel, wavelet packet decomposition, time-order representation (TOR), image processing techniques have been suggested in

the literature to reduce cross terms in the WVD of the multi-component signal [123–126]. However, the suppression of cross-terms is achieved at the expense of sacrificing time and frequency resolutions obtained by the WVD.

The adaptive filter bank has been used to decompose a multi-component signal into mono-component signals [75]. The adaptive filters track the center frequencies of the strong mono-component signals contained in the multi-component signal. The determination of the number of filters required to track emerging and decaying mono-component signals in the time-varying multi-component signal, tracking of bandwidth and center frequencies of all the strong mono-components signals, attaining the desired frequency resolution, determination of the required adaptation rate of the filter parameters are key challenges of such methods. Moreover, an adaptive all-zero filter is required along with each adaptive resonance filter to attenuate leakages in the desired mono-component signal from the other mono-component signals contained in the multi-component signal [75].

An adaptive method, empirical mode decomposition (EMD) has been proposed in [78] to decompose multi-component non-stationary signals. It decomposes a multi-component signal into a set of intrinsic mode functions (IMFs) which are derived from the multi-component signal using the sifting process. The IMFs have nearly zero instantaneous mean and are amplitude-frequency modulated (AM-FM) signals. However, the limitation of the EMD is that its ability to separate any two constituent mono-component signals of a multi-component signal is affected by the ratio of their mean frequencies and relative amplitudes. EMD is not able to separate two mono-component signals if the ratio of the lower mean frequency to higher mean frequency of the two mono-component signals is approximately in the range of 0.5 - 2 [127]. Therefore, the frequency resolution of the EMD decreases in the high frequency range for a specific value of the sampling rate [128]. EMD also faces difficulty in separating the two constituent mono-component signals of a multi-component signal if the amplitude of the lower frequency component is higher than the amplitude of the higher frequency component [128, 129]. Moreover, the problem of mode-mixing among different IMFs exists in the EMD. Mode-mixing refers to the presence of oscillations of disparate frequency ranges occurring in an IMF during different durations and presence of oscillations of the same frequency range occurring in different IMFs during

different durations. The ensemble EMD (EEMD) proposed in [130] attempts to rectify the mode-mixing problem of EMD but with the disadvantage of very high computational complexity.

The parametric approach of modeling a multi-component non-stationary signal as a sum of AM-FM mono-component signals and estimating model parameters corresponding to each constituent mono-component signal has also been considered in [50, 131, 132]. These methods require prior information about the number of mono-component signals present in the multi-component signal. The determination of maximum likelihood estimate of model parameters is a non-linear process in [131]. In [50], the voiced speech signal is modeled as a multi-component amplitude modulated (AM) signal and the frequency modulation present in it due to the time-varying fundamental frequency (F_0) and its time-varying harmonics is neglected. The method proposed in [132] requires coarse estimates of the frequencies of the mono-component signals present in a multi-component signal in advance. The limitation of being able to track only slowly time-varying mono-component signals of [132] was overcome in [111] which was based on an adaptive quasi-harmonic model (QHM). However, the QHM is suitable to model only those signals which have harmonic structure like voiced speech signal.

There exists a need for a decomposition technique which is adaptive, free of cross-terms, complete and has good frequency resolution over the entire time-frequency plane. In this chapter, we present a new approach to decompose a multi-component non-stationary signal into AM-FM mono-component signals based on performing repeated eigenvalue decomposition (EVD) of the Hankel matrix, initially constructed from the samples of the multi-component non-stationary signal. The proposed approach is iterative in nature. The process of constructing the Hankel matrix, performing EVD of the Hankel matrix and extraction of components, which we refer to as '*Iteration*' is repeated till all extracted components satisfy the defined *Mono-component Signal Criteria* (MSC). This chapter is organized as follows: Section 6.2 derives the conditions on the Hankel matrix size to enable separation and extraction of constituent constant amplitude/frequency mono-component signals of a multi-component signal. Section 6.3 extends the theory and concepts developed in Section 6.2 and proposes an iterative approach for decomposition

of a multi-component non-stationary signal into AM-FM mono-component signals. Section 6.4 presents the decomposition results of the proposed iterative approach on different kinds of synthetic and natural multi-component non-stationary signals and their comparison with the decomposition results obtained using the DFT and EMD. Section 6.5 concludes the paper.

6.2 Extraction of Components from a Multi-component Signal consisting of constant amplitude/frequency mono-component signals

Let H_N^x represent the square Hankel matrix of size $N \times N$ consisting of $2N-1$ elements, constructed from a real signal $x[n]$ spanning Q samples as follows [77]:

$$H_N^x = \begin{bmatrix} x[0] & x[1] & . & . & . & x[N-1] \\ x[1] & x[2] & . & . & . & x[N] \\ . & . & . & . & . & . \\ x[N-1] & x[N] & . & . & . & x[2N-2] \end{bmatrix} \tag{6.1}$$

where $n = 0, 1, ..., Q-1$ and $Q \geq 2N-1$. N is an even number. The square Hankel matrix constructed from a real signal is a symmetric matrix; i.e., $H_N^x = (H_N^x)^T$, where T denotes the transpose operator. The EVD of the square matrix H_N^x can be expressed as [77]:

$$H_N^x = V_x \Lambda_x V_x^T \tag{6.2}$$

where Λ_x is a diagonal matrix with N real and scalar eigenvalues $\lambda_{x,i}, i = 1, 2, ..., N$. V_x is an orthogonal matrix; i.e., $V_x^{-1} = V_x^T$, consisting of real eigenvectors $\vec{v}_{x,i}, i = 1, 2, ..., N$ as its columns, each column consisting of N elements. Any two eigenvectors, $\vec{v}_{x,i}$ and $\vec{v}_{x,j}$ corresponding to different eigenvalues ($\lambda_{x,i} \neq \lambda_{x,j}$) are orthogonal [77].

Let $x[n]$ be a multi-component signal consisting of L constant amplitude/frequency

mono-components signals as:

$$x[n] = \sum_{l=1}^{L} x_l[n] = \sum_{l=1}^{L} A_l \cos\left(2\pi f_l n + \theta_l\right), \; n = 0, 1, .., Q-1 \tag{6.3}$$

such that $A_k \neq A_l$ for $k \neq l$; where $k, l = 1, 2, ..., L$. In (6.3), $f_l = \dfrac{1}{N_l} = \dfrac{F_l}{F_s}$. The frequency of $x_l[n]$ in Hz, the sampling frequency in Hz, the normalized frequency of $x_l[n]$ are denoted by $F_l, F_s, \; f_l$ respectively. The period of $x_l[n]$ in samples is denoted by N_l. A_l and θ_l represent the amplitude and phase of $x_l[n]$ respectively. The number of mono-component signals in $x[n]$ is represented by L. Let $F_l < F_{l+1}, l = 1, 2, ..., L-1$. In order to avoid aliasing, $F_s > 2F_L$. Using (6.1) and (6.3), the Hankel matrix of $x[n]$, H_N^x can be expressed as sum of the Hankel matrices of its mono-component signals $H_N^{x_l}$ as:

$$H_N^x = \sum_{l=1}^{L} H_N^{x_l} \quad , \quad \text{where} \;\; H_N^{x_l} = (H_N^{x_l})^T \tag{6.4}$$

The characteristic equation of H_N^x is given by [77]:

$$\det(H_N^x - \lambda I) = \lambda^N - \mathrm{Tr}(H_N^x)\lambda^{N-1} + ... + \det(H_N^x) = 0 \tag{6.5}$$

where Tr(.) and det(.) denote the trace and determinant of the matrix respectively. Irrespective of the value of N, the ranks and number of non-zero eigenvalues of H_N^x and $H_N^{x_l}$ cannot be greater than twice of the number of constant amplitude/frequency components contained in them. The trace of the matrix can be expressed in terms of its non-zero eigenvalues as follows [77]:

$$\begin{aligned} \mathrm{Tr}(H_N^x) &= \sum_{i=1}^{2L} \lambda_{x,i} \\ \mathrm{Tr}(H_N^{x_l}) &= \sum_{i=1}^{2} \lambda_{x_l,i} \quad , \quad N \geq 2L \end{aligned} \tag{6.6}$$

We now derive the conditions on the Hankel matrix N to enable separation of mono-component signals of $x[n]$ (given by (6.3)) using EVD of H_N^x. We have considered two different cases based on the value of N.

6.2.1 Case (i): when the Hankel matrix size is an integer multiple of the LCM of the fundamental periods contained in the multi-component signal

In this case, $\mathbf{N} = \sigma \mathbf{N}_{\mathrm{LCM}}$. The symbols σ and N_{LCM} denote a positive integer and the least common multiple (LCM) of $N_l, l = 1, 2, ..., L$ respectively. Using (6.1), (6.3) and (6.4), $\mathrm{Tr}(H^{x_l}_{\sigma N_{\mathrm{LCM}}})$ and $\mathrm{Tr}(H^{x}_{\sigma N_{\mathrm{LCM}}})$ are given by:

$$\begin{aligned}
\mathrm{Tr}(H^{x_l}_{\sigma N_{\mathrm{LCM}}}) &= A_l \sum_{n=0}^{\sigma N_{\mathrm{LCM}}-1} \cos(2\pi f_l 2n + \theta_l) \\
&= A_l \, \Re\left(e^{j\theta_l} \sum_{n=0}^{\sigma N_{\mathrm{LCM}}-1} e^{j2\pi f_l 2n} \right) \\
&= \ 0 \ , \quad \forall l \\
\mathrm{Tr}(H^{x}_{\sigma N_{\mathrm{LCM}}}) &= \sum_{l=1}^{L} \mathrm{Tr}(H^{x_l}_{\sigma N_{\mathrm{LCM}}}) = 0
\end{aligned} \tag{6.7}$$

The inner product of i^{th} row/column of $H^{x_l}_{\sigma N_{\mathrm{LCM}}}$ and j^{th} row/ column of $H^{x_k}_{\sigma N_{\mathrm{LCM}}}$ denoted by $\left\langle H^{x_l}_{\sigma N_{\mathrm{LCM}}}, H^{x_k}_{\sigma N_{\mathrm{LCM}}} \right\rangle_{i,j}$ is given by:

$$\begin{aligned}
&\left\langle H^{x_l}_{\sigma N_{\mathrm{LCM}}}, H^{x_k}_{\sigma N_{\mathrm{LCM}}} \right\rangle_{i,j} \\
&= A_l A_k \sum_{n=0}^{\sigma N_{\mathrm{LCM}}-1} \Big(\cos(2\pi f_l (n+i-1) + \theta_l) \\
&\qquad\qquad \times \cos(2\pi f_k (n+j-1) + \theta_k) \Big) \\
&= \frac{A_l A_k}{2} \, \Re\Big(e^{j(2\pi(m_1+m_2)+\theta_l+\theta_k)} \sum_{n=0}^{\sigma N_{\mathrm{LCM}}-1} e^{j2\pi(f_l+f_k)n} \\
&\qquad + e^{j(2\pi(m_1-m_2)+\theta_l-\theta_k)} \sum_{n=0}^{\sigma N_{\mathrm{LCM}}-1} e^{j2\pi(f_l-f_k)n} \Big) \\
&= 0, \qquad\qquad i, j = 1, 2, ..., \sigma N_{\mathrm{LCM}} \ \text{ and } \ k \neq l
\end{aligned} \tag{6.8}$$

where $m_1 = f_l(i-1)$ and $m_2 = f_k(j-1)$. It can be deduced from (6.8) that rows and columns of $H^{x_l}_{\sigma N_{\mathrm{LCM}}}$ and $H^{x_k}_{\sigma N_{\mathrm{LCM}}}$ for $l \neq k$ are orthogonal to each other, where k, l can take values from 1, 2, ..., L. In such scenario, the $2L$ non-zero eigenvalues and corre-

sponding eigenvectors of $H^{x}_{\sigma N_{\mathrm{LCM}}}$ are equal to the set consisting of non-zero eigenvalues and corresponding eigenvectors of $H^{x_l}_{\sigma N_{\mathrm{LCM}}}$ as follows:

$$\begin{aligned} \lambda_{x,(2l+j-2)} &= \lambda_{x_l,j} \\ \vec{v}_{x,(2l+j-2)} &= \vec{v}_{x_l,j}, \qquad l = 1, 2, ..., L \text{ and } j = 1, 2 \end{aligned} \tag{6.9}$$

Moreover, using (6.6) and (6.7), it can be deduced that the two non-zero eigenvalues of $H^{x_l}_{\sigma N_{\mathrm{LCM}}}$ constituting a pair are equal and opposite in sign (EOS) as follows:

$$\lambda_{x_l,1} = -\lambda_{x_l,2} \;\; \forall l \tag{6.10}$$

It can be inferred from (6.9) and (6.10) that the k^{th} mono-component signal of $x[n]$ can be extracted by creating a modified eigenvalue diagonal matrix $\tilde{\Lambda}_{x_k}$ which preserves only the k^{th} non-zero eigenvalue pair of Λ_x as follows:

$$\begin{aligned} \tilde{\Lambda}_{x_k} &= \mathrm{diag}(0, ..., 0, \lambda_{x,2k-1}, \lambda_{x,2k}, 0, ..., 0) \\ &= \mathrm{diag}(0, ..., 0, \lambda_{x_k,1}, -\lambda_{x_k,1}, 0, ..., 0) \end{aligned} \tag{6.11}$$

where diag(.) denotes diagonal matrix. Construct $\tilde{H}^{x_k}_N$ as follows:

$$\tilde{H}^{x_k}_N = V_x \tilde{\Lambda}_{x_k} V_x^T \tag{6.12}$$

where $N = \sigma N_{\mathrm{LCM}}$, $\tilde{H}^{x_k}_N = H^{x_k}_N$. The k^{th} mono-component signal of $x[n]$, denoted $\tilde{x}_k[n]$ is extracted as the average of the skew diagonal elements of $\tilde{H}^{x_k}_N$. Please note that $\tilde{x}_k[n]$ = $x_k[n]$. Thus, in this case the extracted mono-component signals of $x[n]$ denoted by $\tilde{x}_k[n]$, $\forall k$ are same as the original mono-component signals of $x[n]$ denoted by $x_k[n]$, $\forall k$. Here's an example:

$$\text{Example 1:} \quad x[n] = \sum_{l=1}^{L} x_l[n] = \sum_{l=1}^{L} A_l \cos\left(\frac{2\pi F_l n}{F_s} + \theta_l\right) \qquad n = 0, 1, ..., Q-1 \tag{6.13}$$

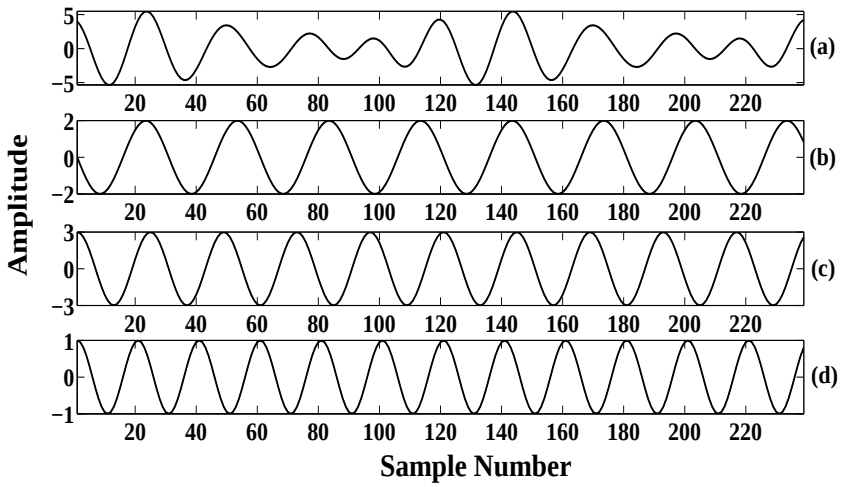

Figure 6.1: (a) Multi-component signal $x[n]$ given by (6.13) (b) Mono-component signal $x_1[n]$ (c) Mono-component signal $x_2[n]$ (d) Mono-component signal $x_3[n]$. Hankel matrix size $N = N_{\text{LCM}} = 120$. Please note that $x_k[n] = \tilde{x}_k[n],\ \forall k$.

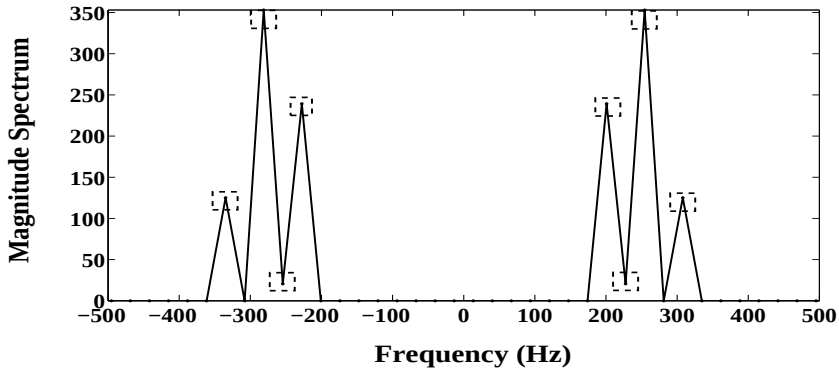

Figure 6.2: Magnitude spectrum of the multi-component signal $x[n]$ (given by (6.13)). Significant transform coefficients are marked by rectangles in dashed lines. Length of the DFT: 239.

where $Q = 239, F_s = 6400$ Hz, $L = 3, A_1 = 2, A_2 = 3, A_3 = 1, \theta_1 = \pi/2, \theta_2 = 0, \theta_3 = 0, F_1 = \dfrac{640}{3}$ Hz, $F_2 = \dfrac{800}{3}$ Hz and $F_3 = 320$ Hz. The value of N is chosen to be equal to $N_{\text{LCM}} = 120$. The non-zero eigenvalue pairs corresponding to the three mono-component signals of $x[n]$ contained in H^x_{120} found using MATLAB are {(-120, 120), (-180, 180), (-60, 60)}. Please note that $|\lambda_{x_l,1}| = \dfrac{NA_l}{2}\ \forall l$, which implies that the magnitude of the eigenvalues constituting a pair is directly proportional to the amplitude of the mono-component signal corresponding to it [4]. The extracted mono-component signals of $x[n]$ using (6.1), (6.2), (6.11) and (6.12) are shown in Fig. 6.1. For comparison, we computed the Q-point ($Q = 239$) discrete Fourier transform (DFT) [62] of $x[n]$, which resulted in 8 significant

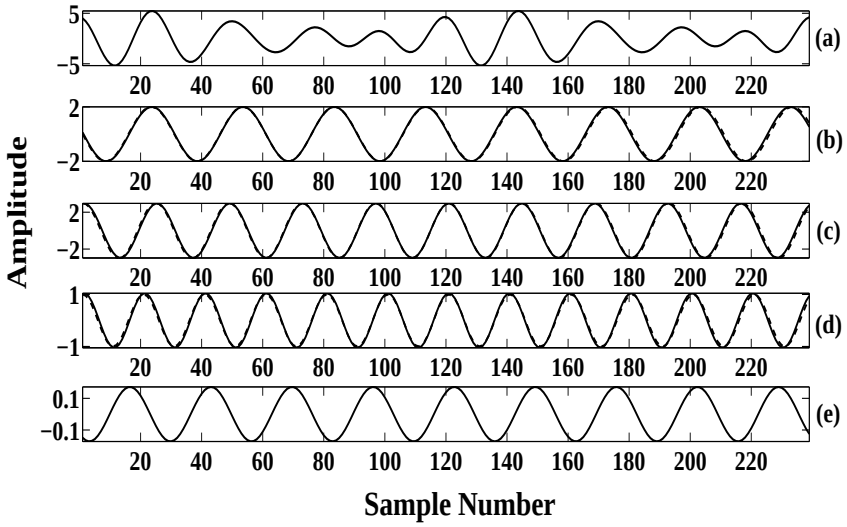

Figure 6.3: (a) Multi-component signal $x[n]$ given by (6.13). Mono-component signals $x_l[n], l = 1, 2, 3$ and mono-component signals extracted by computing the Q-point inverse DFT of each of the three significant transform coefficient pairs are shown in (b), (c) and (d) in dashed and solid lines respectively. (e) Mono-component signal extracted using the Q-point inverse DFT of the fourth significant transform coefficient pair is shown in solid line.

transform coefficients (4 significant transform coefficient pairs, one coefficient corresponding to the positive frequency and the other corresponding to the negative frequency) as shown in Fig. 6.2. We have considered the DFT coefficients significant if their magnitude is greater than 5% of the maximum magnitude in the Q-point DFT of $x[n]$. Three original mono-component signals of $x[n]$ and four mono-component signals extracted by computing the Q-point inverse DFT [62] of each of the significant transform coefficient pairs are depicted in dashed and solid lines respectively in Fig. 6.3. The time limited nature of $x[n]$ gives rise to side leakages around the local maxima in the DFT of $x[n]$, which resulted in one additional component being extracted from $x[n]$ as shown in Fig. 6.3 (e). The local maxima in the DFT of $x[n]$ occur at around the fundamental frequencies of $x_l[n]$, $\forall l$.

6.2.2 Case (ii): when the Hankel matrix size is not an integer multiple of the LCM of the fundamental periods contained in the multi-component signal

In this case, $N \neq N_{\text{LCM}}$. Let us consider the same signal $x[n]$ given by (6.13), but now spanning $(0, 1, ..., 2N-2)$ samples. In practical scenarios, N_{LCM} is not known in advance.

In such cases, the relations in (6.7), (6.8), (6.9), (6.10) no longer hold. Let the eigenvalues of H_N^x be now arranged in an ascending order; i.e., $\lambda_{x,i+1} \geq \lambda_{x,i}$, $i = 1, 2, ..., N-1$. In this case, the modified eigenvalue diagonal matrix preserving the k^{th} non-zero eigenvalue pair of Λ_x is given by:

$$\tilde{\Lambda}_{x_k} = \text{diag}(0, ..., 0, \lambda_{x,k}, 0, ..., 0, \lambda_{x,N-k+1}, 0, ..., 0) \tag{6.14}$$

The Hankel matrix formed by preserving the k^{th} eigenvalue pair of H_N^x denoted by $\tilde{H}_N^{x_k}$ is computed using $\tilde{\Lambda}_{x_k}$ as follows:

$$\begin{aligned} \tilde{H}_N^{x_k} &= V_x \tilde{\Lambda}_{x_k} V_x^T \\ &= \lambda_{x,k} \vec{v}_{x,k} \vec{v}_{x,k}^T + \lambda_{x,N-k+1} \vec{v}_{x,N-k+1} \vec{v}_{x,N-k+1}^T \end{aligned} \tag{6.15}$$

The k^{th} mono-component signal of $x[n]$ is extracted by taking the mean of the skew diagonal elements of $\tilde{H}_N^{x_k}$. Let the k^{th} original component of $x[n]$ and the k^{th} extracted component of $x[n]$ be denoted by $x_k[n]$ and $\tilde{x}_k[n]$ respectively. In order to objectively measure the difference between $x_k[n]$ and $\tilde{x}_k[n]$, we define a quantity, error to signal ratio for $x_k[n]$, denoted ESR_N^k as follows:

$$ESR_N^k = \frac{\sum_{n=0}^{2N-2} (x_k[n] - \tilde{x}_k[n])^2}{\sum_{n=0}^{2N-2} (x_k[n])^2} \tag{6.16}$$

where k can take values from 1, 2, ..., L. A small value of $ESR_N^k, \forall k$ ensures separation of mono-component signals of $x[n]$ with good accuracy. Therefore, we now study the variation of ESR_N^k with respect to N. It is apparent in (6.15), that $\tilde{x}_k[n]$ is a function of the two eigenvectors $(\vec{v}_{x,k}, \vec{v}_{x,N-k+1})$, corresponding to the k^{th} eigenvalue pair of H_N^x. It is not feasible to derive mathematical equations for the eigenvectors of Hankel matrix of arbitrary size N. Therefore, it is very difficult to analytically derive the relation between ESR_N^k and the Hankel matrix size N. Hence, we resort to an empirical study of the variation of ESR_N^k with respect to N. As depicted in Fig. 6.4, the value of

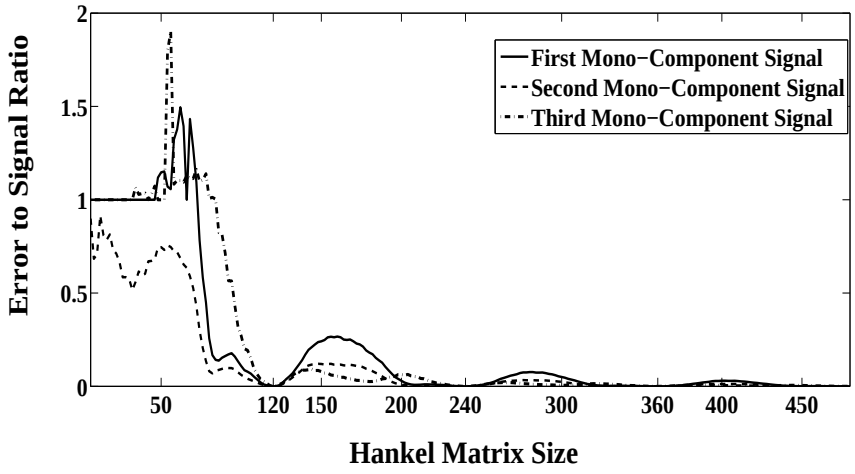

Figure 6.4: Error to signal ratio for the three mono-component signals of the multi-component signal $x[n]$ (given by (6.13)) with respect to the Hankel matrix size (N), computed after the first *Iteration*.

$ESR_N^k, k = 1, 2, ..., L$ does not vary monotonically with respect to N. However, the successive local maxima of $ESR_N^k, \forall k$ reduces with an increase in the value of N. It can be observed in Fig. 6.4 that the value of $ESR_{\sigma N_{\text{LCM}}}^k = 0$, $\forall k$, $\sigma = 1$, 2, 3, in accordance to the mathematically derived result of the last subsection.

In order to comprehend the variation of ESR_N^k with respect to N, it is necessary to understand the variation of the combined magnitude spectrum of the eigen vectors $(\vec{v}_{x,k}, \vec{v}_{x,N-k+1})$ with respect to $N, \forall k$. The combined magnitude spectrums of the eigenvectors corresponding to significant eigenvalue pairs of H_N^x over the positive frequency range are shown in Fig. 6.5, Fig. 6.6, Fig. 6.7, Fig. 6.8 for $N = 60, 90, 130, 270$. We have considered an eigenvalue pair to be significant if the magnitude of one of its eigenvalues is equal to or greater than 10% of the maximum eigen value of H_N^x. It can be deduced from Fig. 6.5, Fig. 6.6, Fig. 6.7, Fig. 6.8 that the eigenvectors $(\vec{v}_{x,k}, \vec{v}_{x,N-k+1}), \forall k$ are not sinusoidal signals. It can be easily computed using (6.13) that the minimum frequency separation denoted by $\Delta F_{x,\min}$ between the components of $x[n]$ is 53.33 Hz. The inverse of $\Delta F_{x,\min}$ is equal to 120 samples at $F_s = 6400$ Hz. It can be inferred from Fig. 6.5, Fig. 6.6, Fig. 6.7, Fig. 6.8 that the combined magnitude spectrum of $(\vec{v}_{x,k}, \vec{v}_{x,N-k+1}), \forall k$ attains the maximum value at the fundamental frequency of $x_k[n], \forall k$ only when $N > \dfrac{F_s}{\Delta F_{x,\min}}$, where $k = 1, 2, ..., L$. It can be deduced from Fig. 6.5, Fig.

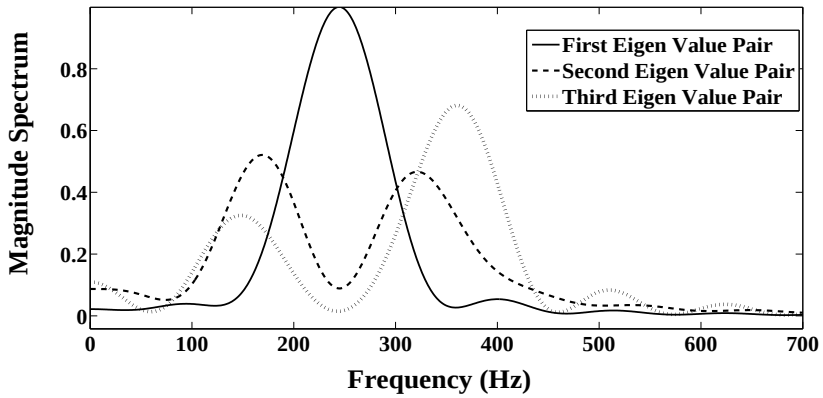

Figure 6.5: Combined magnitude spectrum of the eigenvectors corresponding to significant eigenvalue pairs of H_N^x after the first *Iteration*. $N = 60$.

6.6, Fig. 6.7 and Fig. 6.8 that the frequency range over which the combined magnitude spectrum of $(\vec{v}_{x,k}, \vec{v}_{x,N-k+1})$ has significant value, gradually reduces with an increase in the value of N, where $k = 1, 2, ..., L$. However, it can be inferred from the comparison of Fig. 6.7 with Fig. 6.9 that such a reduction is not monotonic in nature with respect to N, which accounts for the non-monotonic reduction of $ESR^k, \forall k$ with respect to N, as shown in Fig. 6.4. It can be observed in Fig. 6.4 that the value of $ESR_N^k \approx 0, \forall k$ for $N >> \frac{F_s}{\Delta F_{x,\min}}$. The reason is that the frequency range over which the combined magnitude spectrum of $(\vec{v}_{x,k}, \vec{v}_{x,N-k+1})$ has significant value reduces substantially and it takes the form of narrowband pulse at around the fundamental frequency of $x_k[n]$ for $N >> \frac{F_s}{\Delta F_{x,\min}}, \forall k$, which results in $\tilde{x}_k[n], \forall k$ to approach the original sinusoidal functions contained in $x[n]$. We have obtained similar variation of $ESR_N^k, \forall k$ with respect to N for 500 different combinations of the values of $L, A_l, \theta_l, F_l, F_s$. Hence, we conclude from this empirical study that the frequency resolution that can be achieved by performing EVD of H_N^x increases non-monotonically with an increase in the value of the Hankel matrix size N.

Multiple Iterations and Mono-component Signal Criteria

It can be inferred from Fig. 6.9 that even when $N > \frac{F_s}{\Delta F_{x,\min}}$, the combined magnitude spectrum of $(\vec{v}_{x,k}, \vec{v}_{x,N-k+1})$ may have side lobes around the fundamental frequencies contained in $x[n]$, other than the fundamental frequency at which it attains the maximum,

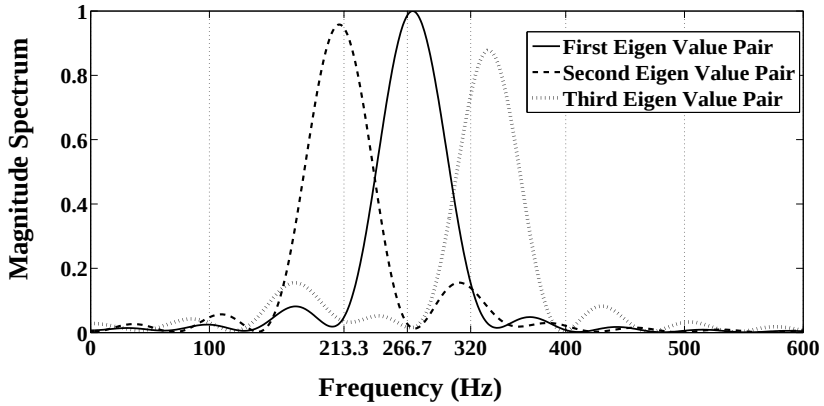

Figure 6.6: Combined magnitude spectrum of the eigenvectors corresponding to significant eigenvalue pairs of H_N^x after the first *Iteration*. $N = 90$.

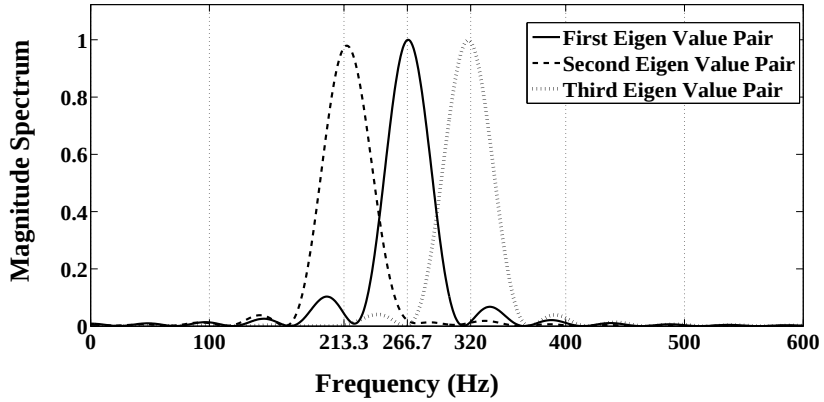

Figure 6.7: Combined magnitude spectrum of the eigenvectors corresponding to significant eigenvalue pairs of H_N^x after the first *Iteration*. $N = 130$.

where k can take values from $1, 2, ..., L$. The extracted components corresponding to such eigen vectors contain significant contribution from two or more mono-component signals of $x[n]$ and it may be possible to further decompose them. Therefore, each extracted component $\tilde{x}_k[n], \forall k$ is checked for the *Mono-Component Signal Criteria* (MSC) defined as follows:

(1) The magnitude of difference between the number of zero crossings and number of extrema (local minima and local maxima) of the extracted component denoted by D_n is equal to zero or one.

(2) The number of significant eigen values pairs obtained by performing EVD of the Hankel matrix constructed from the samples of the extracted component denoted by D_r is

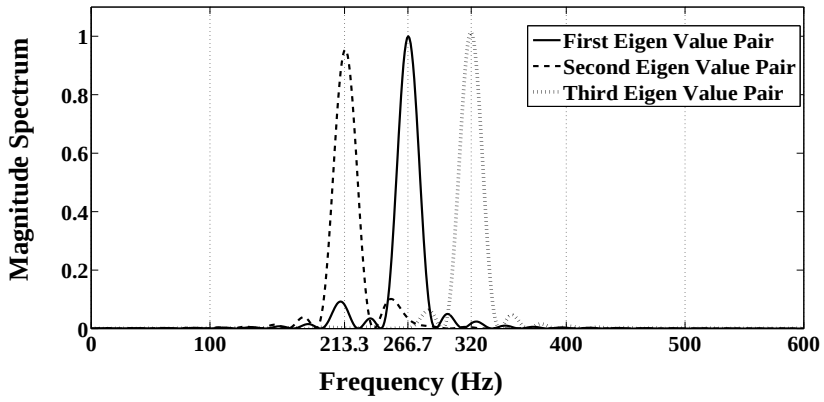

Figure 6.8: Combined magnitude spectrum of the eigenvectors corresponding to significant eigenvalue pairs of H_N^x after the first *Iteration*. $N = 270$.

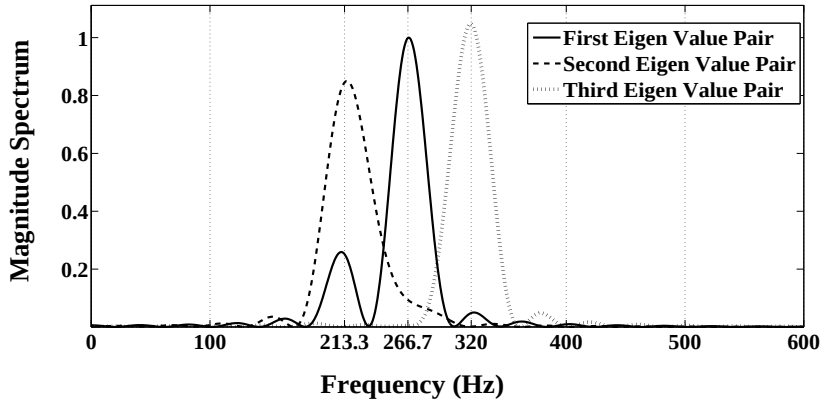

Figure 6.9: Combined magnitude spectrum of the eigenvectors corresponding to significant eigenvalue pairs of H_N^x after the first *Iteration*. $N = 160$.

equal to one.

If an extracted component of $x[n]$ does not satisfy the MSC, then the process of EVD and component extraction using (6.14) and (6.15) which we refer to as '*Iteration*' is repeated by treating the extracted component as multi-component signal for the next *Iteration*. The *Iterations* are repeated till all the extracted components of $x[n]$ satisfy the MSC. Please note that the first part of the MSC is first stated and used in [78] to extract the intrinsic mode functions (IMFs) from a multi-component signal using EMD. The combined magnitude spectrum of the eigenvectors corresponding to different mono-component signals of $x[n]$ extracted after the second *Iteration* for $N = 160$ are depicted in Fig. 6.10. Please observe in Fig. 6.10 that none of the extracted components have

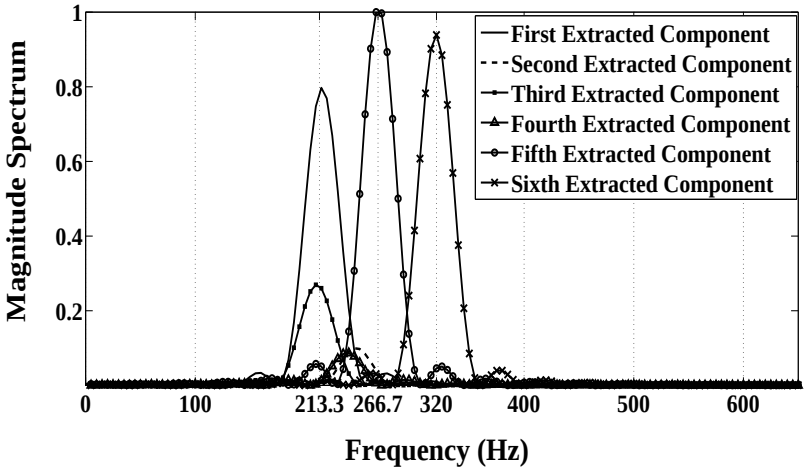

Figure 6.10: Combined magnitude spectrum of the eigenvectors corresponding to different extracted mono-component signals of $x[n]$ (given by (6.13)) after the second *Iteration*. At the second *Iteration* level, EVD is performed on the Hankel matrices constructed from the samples of the extracted components obtained after the first *Iteration* that do not satisfy the MSC. $N = 160$.

significant side lobes. Thus, the extraction of mono-component signals from $x[n]$ (given by (6.13)) became possible for $N = 160$ after the second *Iteration*. However, it can be observed in Fig. 6.10 that some of the extracted mono-component signals contain oscillations belonging to the same frequency range. Therefore, we require the next step as explained below.

Merging of extracted AM-FM mono-component signals with overlapping 1-dB bandwidth

At the last *Iteration* level, the extracted mono-component signals of $x[n]$ that have overlapping 1-dB bandwidth are added to each other. Let the P mono-component signals of $x[n]$ obtained at the last *Iteration* level be denoted by $y_p[n], p = 1, 2, ..., P$. The squared magnitude spectrum of $y_p[n]$ in dB denoted by $E_p(f_r)$ is given by:

$$E_p(f_r) = 10\log_{10}(|Y_p(f_r)|^2), \quad f_r = \frac{r}{R} \tag{6.17}$$

where $Y_p(f_r)$ denotes the R-point DFT of $y_p[n]$. The value of R must be chosen large to compute the DFT at a good frequency resolution. The 1-dB bandwidth of $y_p[n]$ is defined

as the range of frequencies over which the value of $E_p(f_r)$ is not below more than 1-dB of the maximum value of $E_p(f_r)$, where $p = 1, 2, ..., P$. Let the AM-FM mono-component signals of $x[n]$ (given by 6.13) obtained after merging of the extracted components at the last *Iteration* level with overlapping 1-dB bandwidth be denoted by $\bar{y}_p[n], p = 1, 2, ..., S$, where $S \leq P$. The variation of $ESR_N^k, \forall k$ with respect to N computed after the second *Iteration* is depicted in Fig. 6.11, where $x[n]$ is given by (6.13). For each original mono-component signal $x_k[n]$, the extracted mono-component signal $\bar{y}_p[n], p = 1, 2, ..., S$ that provides the minimum value of the error to signal ratio is chosen. We have obtained similar variation of ESR_N^k with respect to N for 500 different combinations of the values of $L, A_l, \theta_l, F_l, F_s$. Hence we conclude from this empirical study that $ESR_N^k, \forall k$ reduces substantially for $N > \dfrac{F_s}{\Delta F_{x,\min}}$, when *Iterations* are repeated till all the extracted components of $x[n]$ (given by (6.3)) satisfy the MSC. In practice, *Iterations* are terminated after the fourth level. Please note that multiple *Iterations* are introduced to separate components of $x[n]$ (given by (6.3)) belonging to disparate frequency ranges and hence, improve the frequency resolution of the decomposition process for a given value of the Hankel matrix size N.

The AM-FM mono-component signals of $x[n]$ (given by (6.13)) obtained in the second *Iteration* after the merging step are depicted in Fig. 6.12 and Fig. 6.13 for the two different values of N. It is evident from Fig. 6.12 that the iterative EVD of the Hankel matrix, initially constructed from the samples of $x[n]$ was not able to separate the contributions of three original mono-component signals contained in $x[n]$ for $N = 60$ because $60 < \dfrac{F_s}{\Delta F_{x,\min}}$, where $\dfrac{F_s}{\Delta F_{x,\min}} = 120$ (using (6.13)). When $N = 220$, the separation of the three components of $x[n]$ was achieved by performing the iterative EVD of the Hankel matrix, initially constructed from the samples of $x[n]$. The following inferences are drawn from this empirical study:

(1) In order to separate all the components contained in $x[n]$, the Hankel matrix size N must be greater than $\dfrac{F_s}{\Delta F_{x,\min}}$, where $\Delta F_{x,\min}$ is the minimum frequency separation between the components of $x[n]$ (given by (6.3)).

(2) The components extracted by EVD of H_N^x approach original sinusoidal mono-component signals contained in $x[n]$ when $N = \sigma N_{\text{LCM}}$ or when $N >> \dfrac{F_s}{\Delta F_{x,\min}}$.

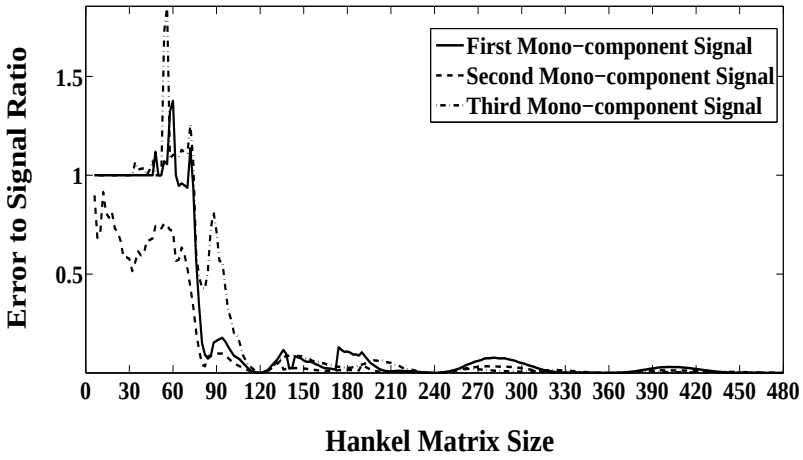

Figure 6.11: Error to signal ratio for the three mono-component signals of the multi-component signal $x[n]$ (given by (6.13)) with respect to the Hankel matrix Size (N) computed after the second *Iteration*.

In order to compare the results obtained in Fig. 6.12 and Fig. 6.13 with those obtained by the DFT, we computed the DFT of $2N-1$ samples of $x[n]$ (given by (6.13)) for the two values of N: 60 and 220, which resulted in 3 and 10 significant transform coefficient pairs respectively. It clearly indicates that the DFT results in representation of the time-limited multi-component signal $x[n]$ using either equal or significantly more number of sinusoidal functions than the number of original sinusoidal components contained in $x[n]$. On the other hand, brevity is achieved by representing $x[n]$ as sum of AM-FM mono-component signals extracted by performing iterative EVD of the Hankel matrix, initially constructed from the samples of $x[n]$. We now extend the concepts and theory developed in this section to devise an approach for the extraction of AM-FM mono-components signals from a multi-component non-stationary signal.

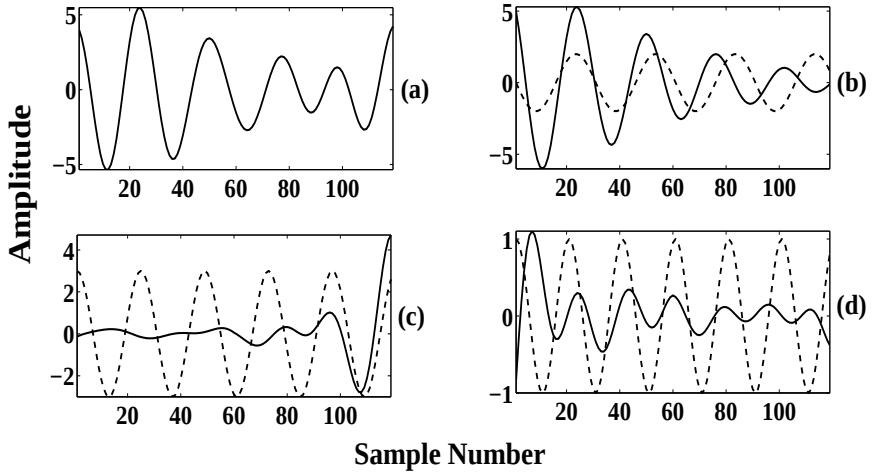

Figure 6.12: (a) Multi-component signal $x[n]$ given by (6.13). Extracted AM-FM mono-component signals $\bar{y}_p[n], p = 1, 2, 3$ in (b), (c) and (d) after the second *Iteration*. $N = 60$.

6.3 Extraction of AM-FM Mono-Component Signals from a Multi-Component Non-stationary Signal using Eigen-value Decomposition of Hankel Matrix

Let $x[n]$ be a multi-component non-stationary signal given by:

$$x[n] = \sum_{l=1}^{L} x_l[n], \qquad n = 0, 1, ..., Q-1 \tag{6.18}$$

where $x_l[n]$ is the l^{th} AM-FM mono-component signal of $x[n]$. Thus, $x_l[n]$ can be represented as:

$$x_l[n] = A_l[n] \cos\left(2\pi f_l[n] n + \theta_l[n]\right), \quad f_l[n] = \frac{F_l[n]}{F_s} \tag{6.19}$$

where $A_l[n], f_l[n], \theta_l[n]$ denote the time-varying amplitude, normalized frequency and phase of $x_l[n]$. The time-varying frequency of $x_l[n]$ in Hz is represented by $F_l[n]$.

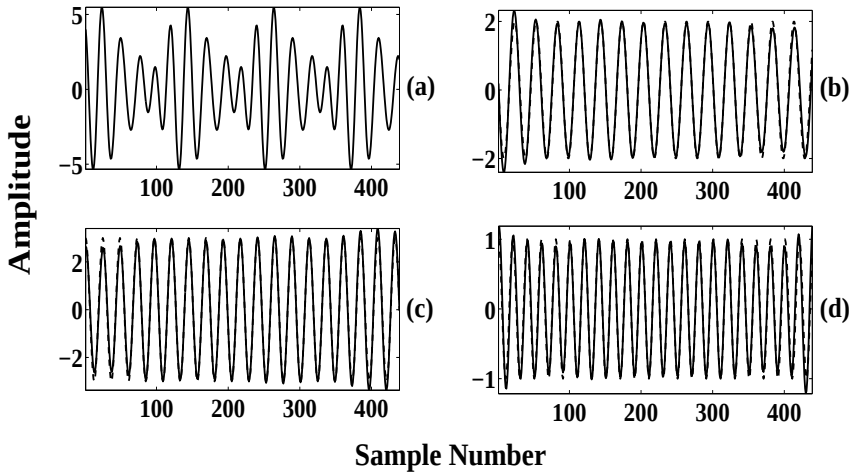

Figure 6.13: (a) Multi-component signal $x[n]$ given by (6.13). Extracted AM-FM mono-component signals $\bar{y}_p[n], p = 1, 2, 3$ in (b), (c) and (d) after the second *Iteration*. $N = 220$.

6.3.1 Tradeoff between frequency resolution and brevity of representation

The rank and number of non-zero eigenvalues of H_N^x are greater than $2L$, when $x_l[n]$ have time-varying amplitudes. H_N^x is a full rank matrix if $F_l[n]$ are time-varying in nature, where l can take values from 1, 2, ..., L. However, if the frequencies of $x_l[n], \forall l$ do not change significantly over N samples, the rank of H_N^x is close to $2L$ [84]. Therefore, in order to restrict the rank and number of significant eigenvalues of H_N^x close to $2L$, the value of N must be chosen such that the frequencies $F_l[n], \forall l$ do not change significantly over the N data samples. This condition on the value of N facilitates in achieving brevity by representing $x[n]$ in terms of the extracted components corresponding to a small number of significant eigenvalue pairs. This condition poses an upper limit on the value of N. An implication of this condition is that the rank and number of significant eigenvalue pairs of $H_N^{x_l}$, $\forall l$ is restricted to be close to two and one respectively. It implies that the time-varying component $x_l[n]$ can be very closely approximated using a few significant eigenvalue pairs and the corresponding eigenvectors of $H_N^{x_l}$, where $l = 1, 2, ..., L$.

On the other side, it has been deduced in the previous section that the value of N must be greater than $\dfrac{F_s}{\Delta F_{\text{des}}}$, to enable separation of components of $x[n]$ separated in frequency domain by equal to or greater than ΔF_{des}, using repeated EVD of Hankel matrix. The

desired frequency resolution in Hz is represented by ΔF_{des}. This condition poses a lower limit on the value of N. Thus, the two conditions on the value of N presents a tradeoff between the frequency resolution and the brevity of representing $x[n]$ in terms of the extracted AM-FM mono-component signals, that can be achieved by performing iterative EVD of Hankel matrix, initially constructed from the samples of $x[n]$. The value of $2N-1$ can be chosen as the smallest integer which divides $x[n]$ spanning Q samples into M equal size segments denoted by $\breve{x}_m[n], m = 1, 2, ..., M$ of length $2N-1$ samples, subject to the constraint that $N > \dfrac{F_s}{\Delta F_{\text{des}}}$. Please note that the assumption of short-term stationarity of $x[n]$ is not required while dividing the multi-component non-stationary signal $x[n]$ into segments.

6.3.2 Proposed iterative approach for decomposition of a multi-component non-stationary signal

Each segment of $x[n]$ is decomposed separately by iteratively performing EVD of $H_N^{\breve{x}_m}\ \forall m$. The EVD of Hankel matrix of $H_N^{\breve{x}_m}$ is given by:

$$H_N^{\breve{x}_m} = V_{\breve{x}_m} \Lambda_{\breve{x}_m} V_{\breve{x}_m}^T \tag{6.20}$$

It can be inferred from (6.15), that the Hankel matrix formed by preserving the k^{th} non-zero eigenvalue pair of $H_N^{\breve{x}_m}$ is given by:

$$\begin{aligned} \tilde{H}_N^{\breve{x}_{m,k}} &= V_{\breve{x}_m} \tilde{\Lambda}_{\breve{x}_{m,k}} V_{\breve{x}_m}^T \\ &= \lambda_{\breve{x}_{m,k}} \vec{v}_{\breve{x}_{m,k}} \vec{v}_{\breve{x}_{m,k}}^T + \lambda_{\breve{x}_{m,N-k+1}} \vec{v}_{\breve{x}_{m,N-k+1}} \vec{v}_{\breve{x}_{m,N-k+1}}^T \end{aligned} \tag{6.21}$$

The k^{th} extracted component of $\breve{x}_m[n]$ denoted by $\tilde{x}_{m,k}[n]$ is computed by taking average of the skew diagonal elements of $\tilde{H}_N^{\breve{x}_{m,k}}$, where k can take values from 1, 2, ..., $N/2$. The extracted component $\tilde{x}_{m,k}[n]$ may contain significant contributions from two or more mono-component signals contained in $\breve{x}_m[n]$. Therefore, each extracted component of the multi-component non-stationary signal segment $\breve{x}_m[n]$ is checked for the *Mono-Component Signal Criteria* (MSC) defined in subsection 6.2.2. If an extracted component of $\breve{x}_m[n]$

does not satisfy the MSC, then the *Iteration* comprising of the process of EVD of Hankel matrix, component extraction using (6.20) and (6.21), is repeated by treating the extracted component as the multi-component non-stationary signal segment for the next *Iteration*. The *Iterations* continue till all the extracted components of the considered segment satisfy the MSC. In practice, *Iterations* are terminated after the fourth level as stated in subsection 6.2.2.

The block diagram of the proposed iterative approach for decomposing a multi-component non-stationary signal $x[n]$ into AM-FM mono-component signals is depicted in Fig. 6.14. The multi-component non-stationary signal is divided into equal size segments $\breve{x}_m[n], m = 1, 2, ..., M$ as stated in the subsection 6.3.1. The components corresponding to different eigenvalue pairs of the considered segment are extracted using (6.20) and (6.21). Fig. 6.14 shows the extraction of components from the first segment of $x[n]$. The components of other segments of $x[n]$ are extracted similarly. Each extracted component of the considered segment is categorized as a multi-component or mono-component signal depending on whether it satisfies the MSC or not. All the extracted components that do not satisfy the MSC go through the next level of *Iteration*. The last step is to merge the extracted mono-component signals obtained at the last *Iteration* level with overlapping 1-dB bandwidth as explained in subsection 6.2.2. Let the extracted mono-component signals of $\breve{x}_m[n]$ obtained at the last *Iteration* level be denoted by $y_{m,p}[n], p = 1, 2, ..., P$. Let the AM-FM mono-component signals of $\breve{x}_m[n]$ obtained after performing the merging of $y_{m,p}[n], \forall p$ with overlapping 1-dB bandwidth be denoted by $\bar{y}_{m,p}[n]$, $p = 1, 2, ..., S$, where $S \leq P$. The reconstructed signal denoted by $\hat{x}_m[n]$ is computed by adding the extracted AM-FM mono-component signals as follows:

$$\hat{x}_m[n] = \sum_{p=1}^{S} \bar{y}_{m,p}[n] \tag{6.22}$$

The signal to reconstruction error ratio in dB denoted by SRE_{dB} for $\breve{x}_m[n]$ is computed

as:

$$SRE_{\text{dB}} = 10\log_{10}\left(\frac{\sum_{n=0}^{2N-2}(\breve{x}_m[n])^2}{\sum_{n=0}^{2N-2}(\breve{x}_m[n]-\hat{x}_m[n])^2}\right) \tag{6.23}$$

Adaptive

The proposed iterative decomposition approach is adaptive. Using (6.21), it can be deduced that the two eigenvectors $(\vec{v}_{\breve{x}_{m,k}}, \vec{v}_{\breve{x}_{m,N-k+1}})$ corresponding to the k^{th} eigenvalue pair of $H_N^{\breve{x}_m}$ act as basis functions for the extracted component $\tilde{x}_{m,k}[n]$. The process of EVD is data dependent. It can be inferred from Fig. 6.5, Fig. 6.6, Fig. 6.7, Fig. 6.8, Fig. 6.9, Fig. 6.12, Fig. 6.13 that the time and frequency domain characteristics of the eigenvectors (acting as basis) corresponding to the different eigenvalue pairs of Hankel matrix depend on the data used to construct the Hankel matrix and the Hankel matrix size N.

Completeness

The proposed iterative approach for decomposition of a multi-component non-stationary signal is complete by the virtue of (6.20) and (6.21) because it is based on the extraction of components corresponding to different eigenvalue pairs of the Hankel matrix, initially constructed from the samples of the multi-component non-stationary signal segment. More the number of eigenvalue pairs used to extract components corresponding to them, lesser is the value of the SRE_{dB}. It has been experimentally proven in subsection 6.2.1 that the magnitude of the eigenvalues constituting a pair is directly proportional to the amplitude of the component corresponding to it. Thus, eigenvalues constituting a pair having significant magnitudes correspond to the high energy components of $\breve{x}_m[n]$. Hence, it is sufficient to extract components corresponding to only significant eigenvalue pairs of $H_N^{\breve{x}_m}$. The criterion for determining whether an eigenvalue pair is significant or not is as follows: An eigenvalue pair is considered to be significant if the magnitude of one of the eigenvalues constituting a pair is greater than the significant threshold percentage (STP) of the maximum eigenvalue of the Hankel matrix, constructed from the samples of the multi-component signal segment under consideration. The value of STP

is a design issue and solely depends on the type of the application for which the proposed iterative approach for decomposition of a multi-component non-stationary signal is being used. If weak components of the multi-component non-stationary signal are required to be extracted, then the value of STP must be kept small.

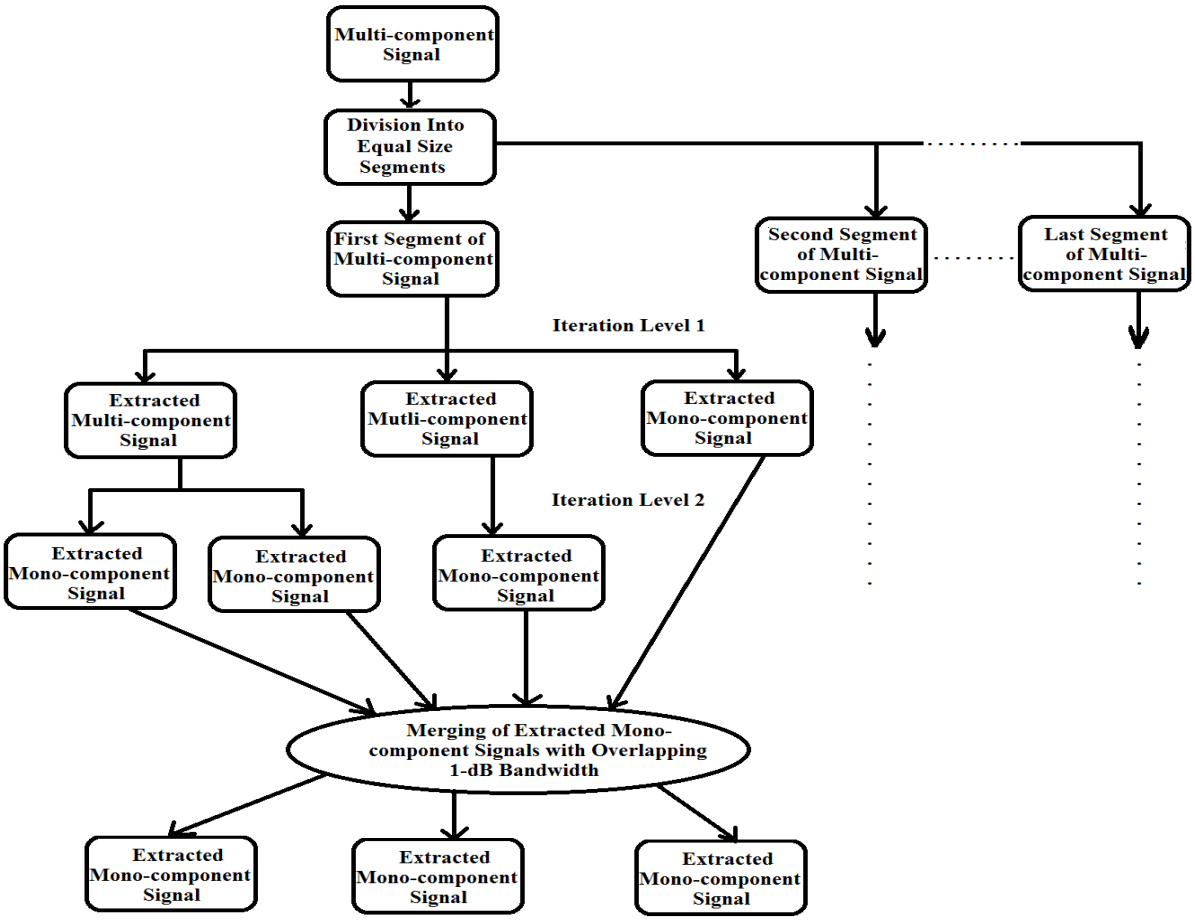

Figure 6.14: Block diagram of the proposed iterative approach for decomposing a multi-component non-stationary signal. The *Iterations* get terminated when all the extracted components are mono-component signals. The decomposition is performed on each segment of the multi-component non-stationary signal.

6.4 Experimental Results

In this section we present the decomposition results obtained by the proposed iterative approach on various synthetic and natural multi-component non-stationary signals (voiced

speech and unvoiced speech signal). The decomposition results obtained by the DFT and EMD are also shown for comparison.

6.4.1 Multi-component non-stationary signal consisting of only amplitude modulated mono-component signals

Let $x[n]$ be a multi-component non-stationary signal with only AM components given by:

$$x[n] = \sum_{l=1}^{L} x_l[n] = \sum_{l=1}^{L} A_l(1 + \alpha_l n)\cos(2\pi f_l n + \theta_l) \tag{6.24}$$

where $n = 0, 1, ..., Q-1$ and $f_l = \dfrac{F_l}{F_s}$. The factor α_l controls the rate of variation of $A_l[n], \forall l$. The values of various parameters used in (6.24) are: $L = 5, Q = 4800, F_s =$ 6400 Hz, $A_1 = 2, A_2 = 1, A_3 = 0.9, A_4 = 3, A_5 = 2.5, \alpha_1 = 20/Q, \alpha_2 = \alpha_5 = 18/Q, \alpha_3 = 14/Q, \alpha_4 = 16/Q, F_1 = 100$ Hz, $F_2 = 140$ Hz, $F_3 = 210$ Hz, $F_4 = 320$ Hz, $F_5 = 500$ Hz, $\theta_1 = \theta_4 = 0, \theta_2 = \pi, \theta_3 = \pi/2, \theta_5 = \pi/3$. $\theta_l, \forall l$ is specified in radians. We have chosen $\Delta F_{\text{des}} = \Delta F_{x,\min} = 40$ Hz. The 4800 samples of $x[n]$ are divided into equal size segments of length $2N-1$ samples, subject to the constraint that $N > \dfrac{F_s}{\Delta F_{x,\min}}$. The decomposition results obtained by the proposed iterative approach on the first segment of $x[n]$ denoted by $\breve{x}_1[n]$ for $N = 184$ and $STP = 10\%$ are depicted in Fig. 6.15. The *Iterations* got terminated after the second level. A very high value of $SRE_{\text{dB}} = 36.2$ dB has been obtained by extracting the components corresponding to only significant eigenvalue pairs.

The 367-point DFT of $\breve{x}_1[n]$ resulted in 32 significant transform coefficient pairs. Thus, the number of significant DFT coefficient pairs substantially increases for a multi-component signal consisting of AM modulated mono-component signals. The reason is that the DFT uses a fixed set of sinusoidal basis functions and presence of amplitude modulation in $x_l[n]$ increases the bandwidth of $x_l[n], \forall l$ [119]. The results obtained by the EMD of $\breve{x}_1[n]$ are shown in Fig. 6.16. The implementation of the EMD using the MATLAB programming language is available at [133]. It is apparent in Fig. 6.16 that the EMD was not able to separate the constituent AM mono-component signals of

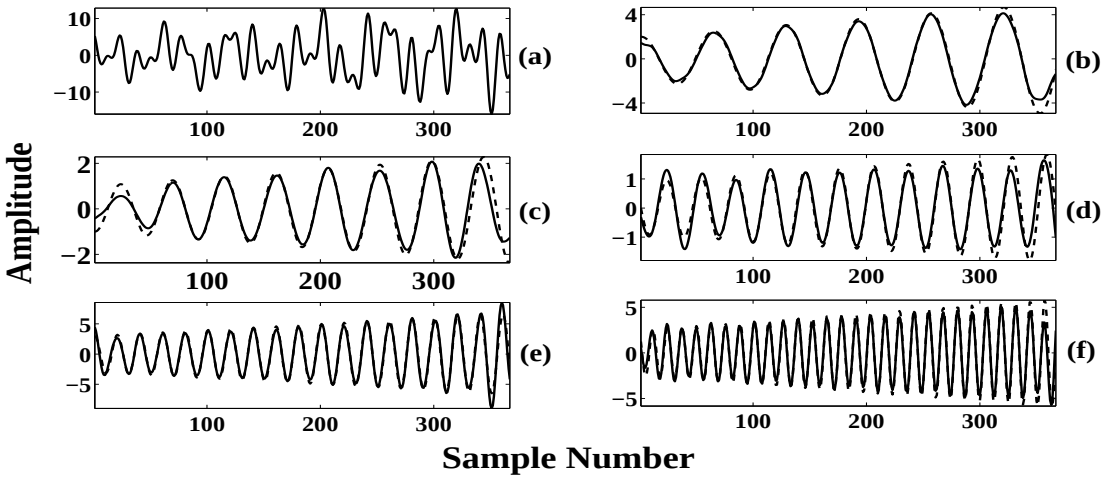

Figure 6.15: (a) Multi-component signal segment $\breve{x}_1[n]$. Extracted AM-FM mono-component signals $\bar{y}_{1,p}[n], p = 1, 2, ..., 5$ obtained using the proposed iterative decomposition approach are shown in solid lines in (b), (c), (d), (e), (f). Original mono-component signals of $\breve{x}_1[n]$ are depicted in dashed lines in (b), (c), (d), (e), (f). $N = 184$. $x[n]$ is given by (6.24).

$\breve{x}_1[n]$. The reason is that the EMD suffers from the mode-mixing problem. The ability of the EMD to separate mono-component signals is also adversely affected when the amplitude of the lower frequency mono-component signal is higher than the amplitude of the higher frequency mono-component signal [129]. Please note that in this example, $A_4 > A_5 > A_1 > A_2 > A_3$. It is apparent from Fig. 6.15 that the advantage of the proposed iterative approach lies in decomposing $\breve{x}_1[n]$ into its constituent AM mono-component signals with a good accuracy and achieving brevity of representation by expressing $\breve{x}_1[n]$ as sum of five AM-FM mono-component signals. This example also demonstrated that the ability of the proposed iterative decomposition approach to resolve components is not affected by the relative amplitudes of the constituent mono-component signals.

6.4.2 Multi-component non-stationary signal consisting of only frequency modulated mono-component signals

Let $x[n]$ be a multi-component non-stationary signal with only frequency modulated (FM) components given by:

$$x[n] = \sum_{l=1}^{L} x_l[n] = \sum_{l=1}^{L} A_l \cos(2\pi f_l(1 + \beta_l n)n + \theta_l) \tag{6.25}$$

where $n = 0, 1, ..., Q-1$. The parameter β_l controls the rate of variation of the instantaneous frequency of $x_l[n], \forall l$.

(1) Single FM component: The value of $L = 1$ in this case. The values of various parameters in (6.25) are chosen as: $F_s = 6400$ Hz, $Q = 4800$, $A_1 = 2$, $F_1 = 210$ Hz, $\theta_1 = 0$, $\beta_1 = 3/Q$. Using (6.25) and the values of β_1 and Q, the value of $F_1[Q-1]$ comes out to be 839.87 Hz. It implies that there is a change of nearly 300% in $F_1[n]$ over 750 ms. In this case, the value of $\Delta F_{x,\min}$ is theoretically infinite because $x[n]$ consists of a single mono-component signal. Therefore, we can choose N to be very small (such as 20) in order to restrict the rank of $H_N^{\breve{x}_1}$ to be close to 2 and to achieve a very high value of SRE_{dB} for a given value of STP. For the sake of demonstrating the outcome of the proposed iterative decomposition approach for relatively large value of N, we have chosen $N = 200$. The decomposition results obtained by proposed iterative approach by terminating the *Iterations* after the fourth level for the two different values of STP: 10% and 5% are depicted in Fig. 6.17. The SRE_{dB} values obtained for the two different values of STP: 10% and 5% are 11.19 dB and 14.23 dB respectively.

The 399-point DFT of $\breve{x}_1[n]$ resulted in 26 significant transform coefficient pairs. Thus, the number of significant DFT coefficient pairs substantially increases for a single FM modulated component. The reason is that the presence of frequency modulation in $\breve{x}_1[n]$ increases the bandwidth of $\breve{x}_1[n]$ [119]. The results obtained by the EMD of $\breve{x}_1[n]$ are shown in Fig. 6.18. As $\breve{x}_1[n]$ satisfies the two conditions to be considered as an IMF [78], the EMD results in only one IMF, which is same as $\breve{x}_1[n]$. The proposed iterative decomposition approach is able to represent $x[n]$ as a single component and offered advantage

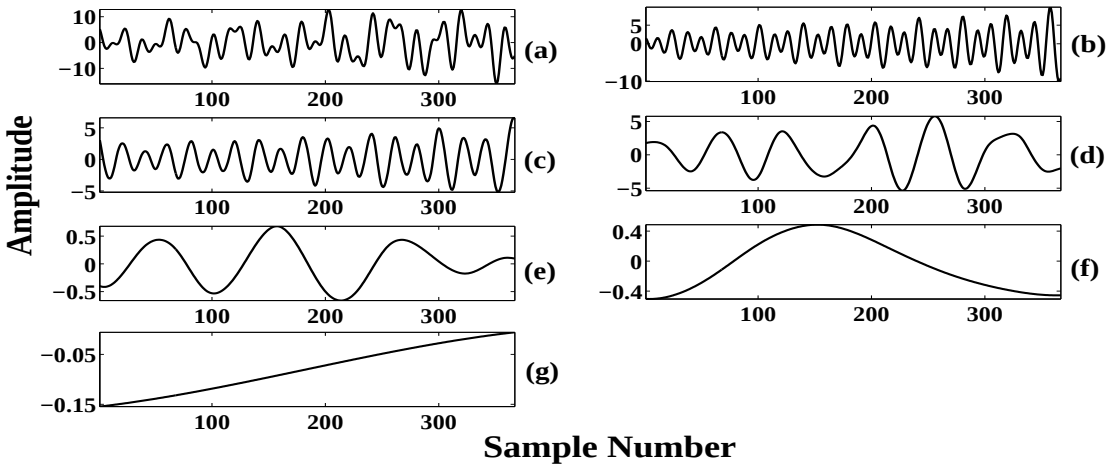

Figure 6.16: (a) Multi-component signal segment $\breve{x}_1[n]$. Extracted IMFs using the EMD of $\breve{x}_1[n]$ are shown in (b), (c), (d), (e), (f), (g). $x[n]$ is given by (6.24).

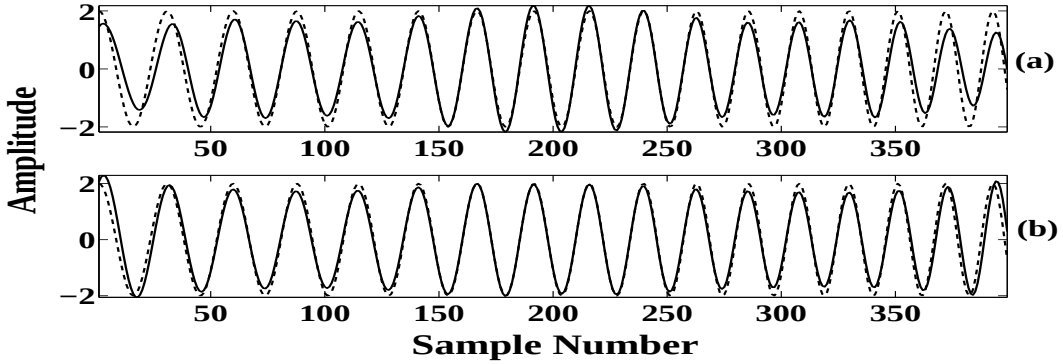

Figure 6.17: Signal segment $\breve{x}_1[n]$ in dashed line and the extracted AM-FM mono-component signal using the proposed iterative decomposition approach in solid line are shown for (a) $STP = 10\%$ (b) $STP = 5\%$. $x[n]$ is given by (6.25) and contains single FM component.

over the DFT which requires 26 sinusoidal functions to represent $x[n]$. The EMD is better than the proposed iterative decomposition approach in this scenario because it was able to represent $x[n]$ as a single component with absolute accuracy.

(2) Multiple FM components: Let $x[n]$ contain two FM components. The values of various parameters in (6.25) are chosen as: $L = 2$, $F_s = 6400$ Hz, $Q = 1190$, $A_1 = 3, A_2 = 2$, $F_1 = 500$ Hz, $F_2 = 700$ Hz, $\theta_1 = 0$, $\theta_2 = \pi$, $\beta_1 = 0.744/Q$, $\beta_2 = -0.620/Q$. Using (6.25), the value of $F_1[119] = 537.19$ Hz, $F_2[119] = 656.62$ Hz. The value of $\Delta F_{x,\min}$ for 119 samples of $x[n] = 119.43$ Hz. Therefore, we have chosen a value of $N = 60$, which is greater

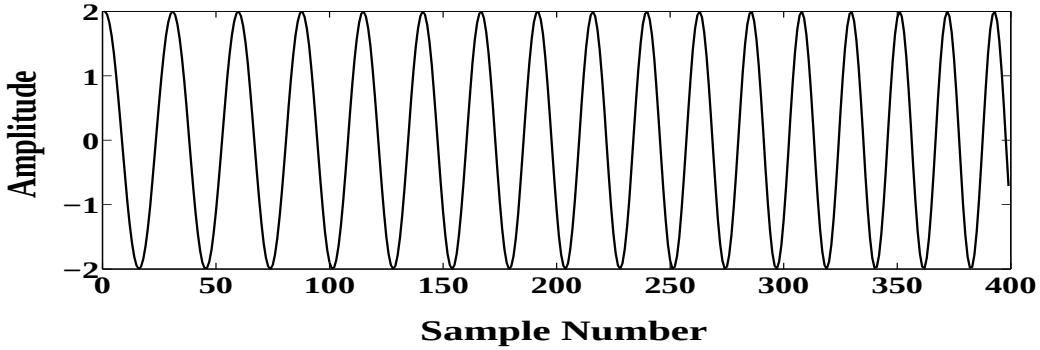

Figure 6.18: Signal segment $\breve{x}_1[n]$ in dashed line and the extracted AM-FM mono-component signal using the EMD in solid line. $x[n]$ is given by (6.25) and contains single FM component.

than $\dfrac{F_s}{\Delta F_{x,\min}}$. The decomposition results obtained by the proposed iterative approach on $\breve{x}_1[n]$ for $STP = 10\%$ are shown in Fig. 6.19. The *Iterations* got terminated after the second level. A high value of SRE_{dB} of 24.81 dB has been obtained by extracting the components corresponding to only significant eigenvalue pairs.

The 119-point DFT of $\breve{x}_1[n]$ resulted in 8 significant transform coefficient pairs. Thus, the number of significant DFT coefficient pairs substantially increases for a multi-component signal consisting of FM modulated mono-component signals. The reason is that the presence of frequency modulation in $x_l[n]$ increases the bandwidth of $x_l[n], \forall l$ [119]. The results obtained by the EMD of $\breve{x}_1[n]$ are shown in Fig. 6.20. It is apparent in Fig. 6.20 that the EMD was not able to separate the constituent FM mono-component signals of $\breve{x}_1[n]$. It can be inferred by observing Fig. 6.20 carefully, that the first IMF (Fig. 6.20 (b)) extracted using the EMD of $\breve{x}_1[n]$ closely matches $\breve{x}_1[n]$. The reason is that the EMD is not able to separate two mono-component signals if the ratio of the lower mean frequency to higher mean frequency of the two mono-component signals is approximately in the range of 0.5 - 2 [127, 128]. The ability of the EMD to resolve components also got adversely affected because $A_1 > A_2$ in this example [129]. It is apparent from Fig. 6.19 that the advantage of the proposed iterative approach lies in decomposing $\breve{x}_1[n]$ into its constituent FM mono-component signals with a good accuracy and achieving brevity of representation by expressing $\breve{x}_1[n]$ as sum of three AM-FM mono-component signals. Unlike the EMD, the proposed iterative decomposition approach is neither affected by

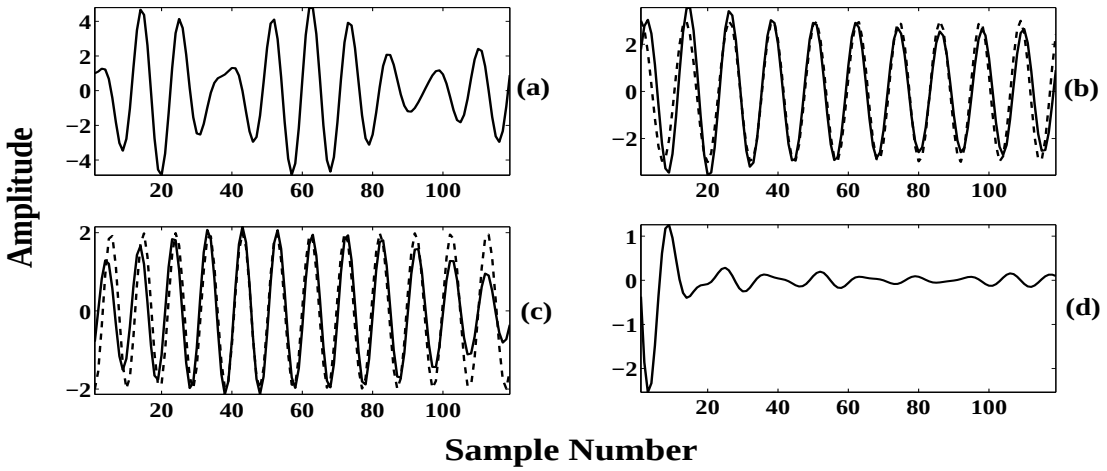

Figure 6.19: Multi-component signal segment $\breve{x}_1[n]$. Extracted AM-FM mono-component signals $\bar{y}_p[n], p = 1, 2, 3$ obtained using the proposed iterative decomposition approach are shown in solid lines in (b), (c), (d). The original FM mono-component signals of $\breve{x}_1[n]$ are shown in dashed lines in (b) and (c). $x[n]$ is given by (6.25) and contains two FM components.

the ratio of the mean frequencies nor by the relative amplitudes of the mono-component signals contained in a multi-component signal.

6.4.3 Multi-component non-stationary signal consisting of amplitude-frequency modulated mono-component signals

Let $x[n]$ be a multi-component non-stationary signal with AM-FM components given by:

$$x[n] = \sum_{l=1}^{L} x_l[n] = \sum_{l=1}^{L} A_l(1 + \alpha_l[n]) \cos(2\pi f_l(1 + \beta_l n)n + \theta_l) \tag{6.26}$$

where $n = 0, 1, ..., Q-1$. The values of various parameters used in (6.26) are chosen as: $L = 4$, $F_s = 10000$ Hz, $Q = 1190$, $A_1 = 3, A_2 = 2, A_3 = 0.9, A_4 = 1.5$, $\alpha_1 = 7.44/Q, \alpha_2 = \alpha_3 = 6.69/Q, \alpha_4 = 5.95/Q$, $F_1 = 500$ Hz, $F_2 = 700$ Hz, $F_3 = 1100$ Hz, $F_4 = 1400$ Hz, $\theta_1 = \theta_4 = 0$, $\theta_2 = \pi$, $\theta_3 = \pi/2$, $\beta_1 = \beta_3 = 0.744/Q$, $\beta_2 = -0.620/Q$, $\beta_4 = -0.496/Q$. Using (6.26), the value of $F_1[119] = 537.19$ Hz, $F_2[119] = 656.62$ Hz, $F_3[119] = 1181.80$ Hz, $F_4[119] = 1330.60$ Hz. The value of $\Delta F_{x,\min}$ for 119 samples of $x[n] = 119.43$ Hz. We have chosen $\Delta F_{x,\text{des}} = \Delta F_{x,\min}$. Therefore, we have kept the value of $N = 60$, which is greater

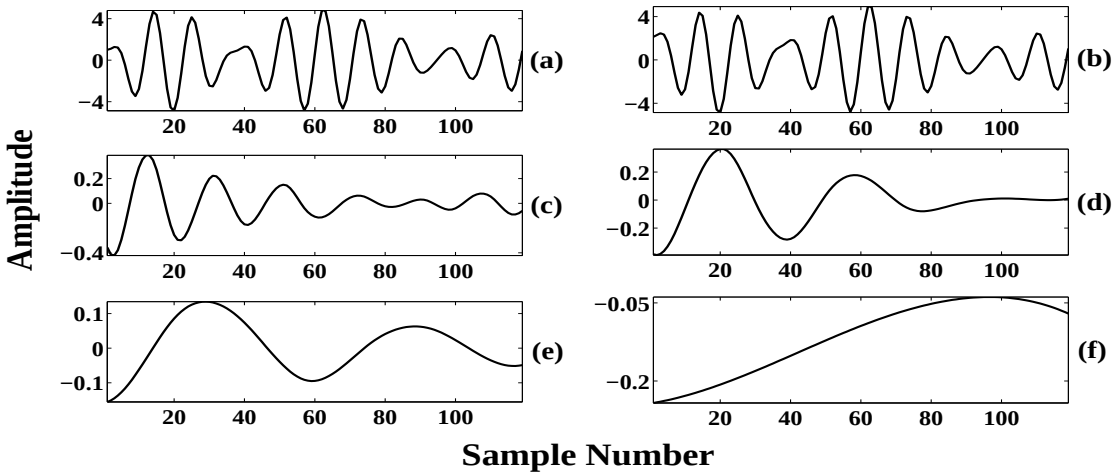

Figure 6.20: Multi-component signal segment $\breve{x}_1[n]$. Extracted IMFs using the EMD are shown in solid lines in (b), (c), (d), (e), (f). $x[n]$ is given by (6.25) and contains two FM components.

than $\dfrac{F_s}{\Delta F_{x,\min}}$. The decomposition results obtained by the proposed iterative approach on $\breve{x}_1[n]$ for $STP = 10\%$ are depicted in Fig. 6.21. The *Iterations* got terminated after the second level. A high value of SRE_{dB} of 22.29 dB has been obtained by extracting the components corresponding to only significant eigenvalue pairs.

The 119-point DFT of $\breve{x}_1[n]$ resulted in 20 significant transform coefficient pairs. Thus, the number of significant DFT coefficient pairs substantially increases for a multi-component signal consisting of AM-FM modulated component signals. The reason is that the DFT employs sinusoids as basis functions and the presence of amplitude and frequency modulation in $x_l[n]$ increases the bandwidth of $x_l[n], \forall l$ [119]. The results obtained by the EMD of $\breve{x}_1[n]$ are shown in Fig. 6.22. It is apparent in Fig. 6.22 that the EMD was not able to separate the constituent AM-FM mono-component signals of $\breve{x}_1[n]$. The mode-mixing problem of the EMD is evident in the first IMF (Fig. 6.22 (b)) that contains oscillations belonging to disparate frequency ranges. The ratio of mean frequencies and relative amplitudes of the mono-component signals contained in a multi-component signal affect the ability of the EMD to separate components. It is apparent from Fig. 6.21 that the AM-FM mono-component signals extracted by the proposed iterative approach do not exactly follow the instantaneous frequency and amplitude of the original AM-FM mono-component signals over the entire duration of $\breve{x}_1[n]$. The benefit of the proposed

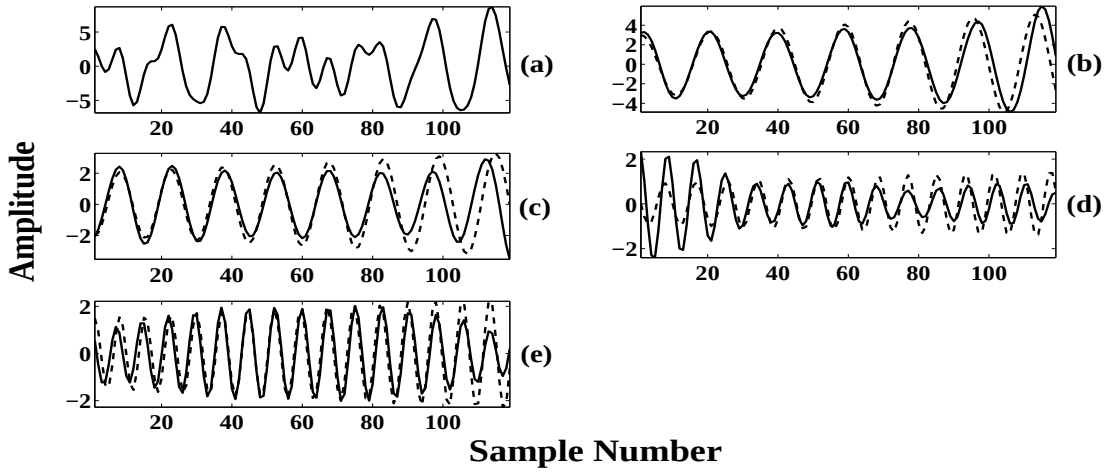

Figure 6.21: Multi-component signal segment $\breve{x}_1[n]$. Extracted AM-FM mono-component signals $\bar{y}_p[n], p = 1, 2, 3, 4$ obtained using the proposed iterative decomposition approach are shown in solid lines in (b), (c), (d), (e). The original AM-FM mono-component signals of $\breve{x}_1[n]$ are shown in dashed lines in (b), (c), (d), (e). $x[n]$ is given by (6.26).

iterative decomposition approach is that it reasonably separated the oscillations lying in disparate frequency bands and achieved brevity by representing $\breve{x}_1[n]$ as sum of four AM-FM mono-component signals.

6.4.4 Voiced speech signal

The vocal folds vibrate during the production of voiced speech like vowels, sonorants. It causes the excitation to the vocal tract system take the form of quasi periodic puffs of air, resulting in a quasi-periodic output signal. The rate of vibration of vocal folds is comprehended as the fundamental frequency (F_0) of voiced speech signal. The AM-FM signal model of voiced speech signal proposed in [66] states that the voiced speech signal is a multi-component signal containing significant energy only around the time-varying F_0 component and its time-varying harmonics.

We have determined the boundaries of voiced regions of a speech signal of the CMU-Arctic database [67,68] using the V/NV detection method proposed in the second chapter of this thesis. The speech signals are available at a F_s of 32 kHz. Voiced speech signal has significant energy in the frequency range of 0 Hz - 3400 Hz [69]. Therefore, the voiced speech signal under consideration is passed through a low pass filter (LPF) with cut off

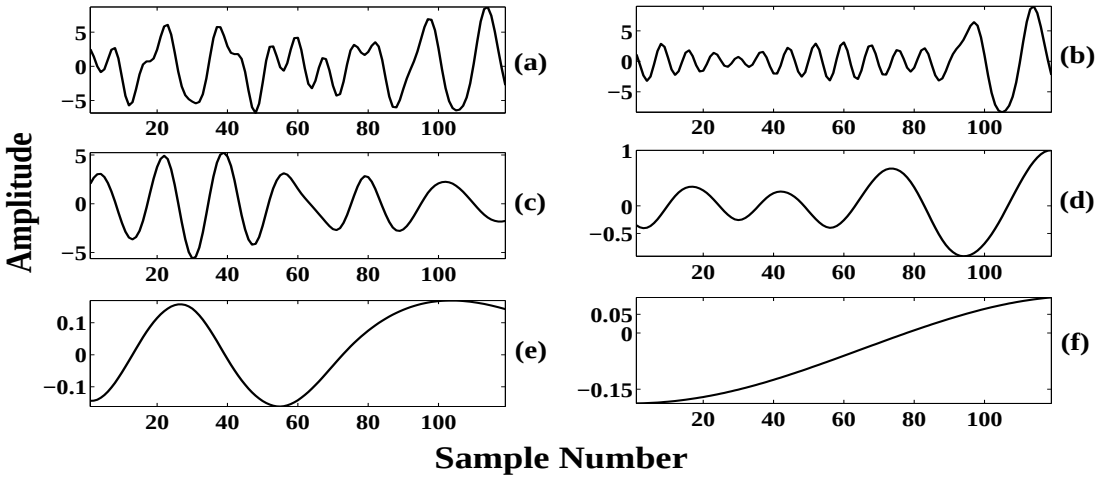

Figure 6.22: Multi-component signal segment $\breve{x}_1[n]$. Extracted IMFs using the EMD are shown in solid lines in (b), (c), (d), (e), (f). $x[n]$ is given by (6.26).

frequency at 3400 Hz, 80 dB attenuation in the stopband and 0.5 dB passband ripple. The F_0 range of voiced speech is 50 Hz - 500 Hz [1]. Let $x[n]$ denote the low pass filtered voiced region (spanning 3849 samples) under consideration for decomposition. According to the AM-FM signal model of voiced speech signal [66], the smallest possible value of $\Delta F_{x,\min}$ is equal to the lowest possible F_0 value of 50 Hz. We have chosen $\Delta F_{x,\text{des}} = \Delta F_{x,\min}$. Therefore, we have kept the value $N = 642$, which is greater than $\dfrac{F_s}{\Delta F_{x,\min}}$ and divides $x[n]$ into equal size segments. Thus, the length of each segment ($2N - 1$ samples) comes out to be 1283 samples. The decomposition results obtained by the proposed iterative approach on the first segment of the low pass filtered voiced speech signal denoted by $\breve{x}_1[n]$ for $STP = 7\%$ are depicted in Fig. 6.23. A moderate value of $SRE_{\text{dB}} = 14.40$ dB has been obtained by extracting components corresponding to only significant eigenvalue pairs. The *Iterations* got terminated after the second level. We have computed the following two parameters for AM-FM mono-component signals extracted from a multi-component non-stationary signal:

(a) Dominant Frequency: The positive frequency at which the squared magnitude spectrum of the extracted AM-FM mono-component signal attains the peak value.

(b) Energy: The squared sample values of the extracted AM-FM mono-component signal.

The extracted AM-FM mono-component signals are arranged in the increasing order of dominant frequencies. The dominant frequencies and energies of the AM-FM mono-

Table 6.1: Extraction of AM-FM mono-component signals from a voiced speech segment using the proposed iterative decomposition approach. $N = 642$.

Segment Number	Last *Iteration* Level	Dominant Frequency of Extracted Mono-Component Signals	Energy of Extracted Mono-Component Signals
1	2	127.1	0.30
		256.9	0.30
		510.7	3.80
		642.6	18.50
		765.5	0.50
		894.8	3.10
		1017.8	0.30
		3173.5	0.20
		3292.7	0.20

component signals $\bar{y}_p[n], p = 1, 2, ..., 9$ extracted from $\breve{x}_1[n]$ are compiled in Table 6.1. It can be inferred from Table 6.1 and Fig. 6.23 that the proposed iterative decomposition approach has been able to extract AM-FM mono-component signals corresponding to the time-varying F_0 component at around 127 Hz and its time-varying higher harmonic components ($2^{nd}, 4^{rd}, 5^{th}, 6^{th}, 7^{th}, 8^{th}, 25^{th}, 26^{th}$ harmonic). The extraction of the time-varying F_0 component of voiced speech signal finds applications in identifying GCIs and estimating the instantaneous F_0 [4, 26]. It can be deduced from Table 6.1 and Fig. 6.23 that the 4th extracted AM-FM mono-component signal $\bar{y}_4[n]$ corresponding to the 5^{th} harmonic of the time-varying F_0 is the formant component of $\breve{x}_1[n]$ because it has significantly more energy than the other extracted components.

The 1283-point DFT of $\breve{x}_1[n]$ resulted in 38 significant transform coefficient pairs. The disadvantage of the DFT is that it requires a large number of sinusoids to represent $\breve{x}_1[n]$. The results obtained by decomposing $\breve{x}_1[n]$ using EMD are shown in Fig. 6.24. The mode-mixing problem of EMD is apparent in the first and second IMFs (Fig. 6.24 (b) and Fig. 6.24 (c)). It can be deduced from Fig. 6.24 that EMD was not able to resolve the time-varying F_0 component and its higher harmonic components contained in $\breve{x}_1[n]$. Therefore, EMD of $\breve{x}_1[n]$ was not able to facilitate the understanding of the AM-FM signal model of voiced speech signal [66] and the underlying process of voiced speech production. The disadvantage of EMD is that its ability to separate components

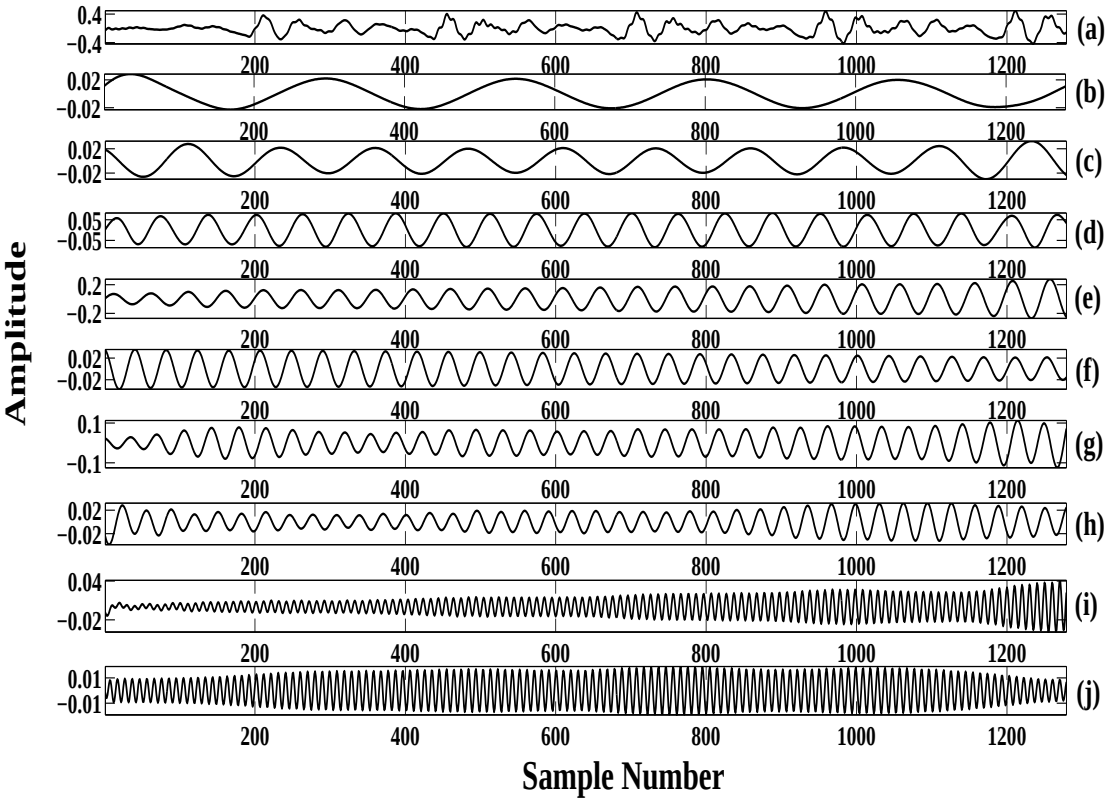

Figure 6.23: (a) Low pass filtered voiced speech segment $\breve{x}_1[n]$. Extracted AM-FM mono-component signals $\bar{y}_p[n], p = 1, 2, ..., 9$, obtained using the proposed iterative decomposition approach are shown in (b), (c), (d), (e), (f), (g), (h), (i), (j).

(belonging to disparate frequency ranges) from a multi-component signal is affected by the relative amplitudes of the mono-component signals contained in it [129]. EMD also suffers from poor frequency resolution over the high frequency range [128]. The proposed iterative approach offers flexibility. The frequency resolution ΔF_{des} that can be achieved by the proposed iterative decomposition approach can be altered by changing the Hankel matrix size N. The frequency resolution remains same over the entire frequency range once the value of N is fixed.

6.4.5 Unvoiced speech signal

Unvoiced speech is produced by the passage of air through a narrow constriction in the windpipe. The unvoiced speech includes fricatives, plosives etc. The unvoiced region has been taken from a speech signal of the CMU-Arctic database [67, 68] available at a F_s

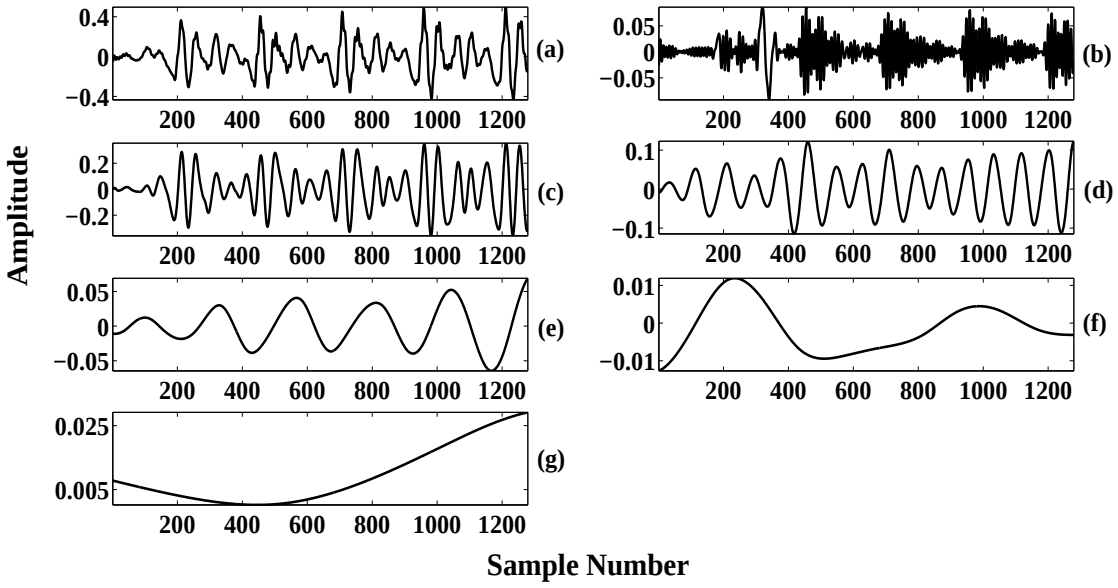

Figure 6.24: (a) Low pass filtered voiced speech segment $\breve{x}_1[n]$. Extracted IMFs obtained using the EMD are shown in solid lines in (b), (c), (d), (e), (f), (g).

of 32 kHz. Unvoiced speech signal has significant energy at high frequencies up to 8000 Hz [69]. Therefore, the unvoiced speech signal under consideration is passed through a LPF with cut-off frequency at 8000 Hz, 80 dB stop band attenuation and 0.5 dB passband ripple. The unvoiced speech signal energy is scattered over the frequency range of 0 - 8000 Hz because of its noise-like random nature, with more energy in the high frequency components than in the low frequency components. We choose the value of ΔF_{des} to be 400 Hz. We have chosen a value N equal to 94 which is greater than $\dfrac{F_s}{\Delta F_{\text{des}}}$ and divides $x[n]$ (spanning 4489 samples) into equal size segments. The decomposition results obtained by the proposed iterative approach on the first segment of the low pass filtered unvoiced speech signal $\breve{x}_1[n]$ after the fourth *Iteration* level for $STP = 1\%$ are depicted in Fig. 6.25 and Fig. 6.26. A high value of SRE_{dB} of 16.94 dB has been obtained in this case, when components corresponding to only significant eigenvalue pairs were extracted. The dominant frequencies and energies of the extracted AM-FM mono-component signals are tabulated in Table 6.2.

The 187-point DFT of $\breve{x}_1[n]$ resulted in 56 significant transform coefficient pairs. The proposed iterative approach achieves brevity in comparison to the DFT while representing $\breve{x}_1[n]$ as sum of mono-component signals. The IMFs obtained by performing EMD of $\breve{x}_1[n]$

Table 6.2: Extraction of AM-FM mono-component signals from an unvoiced speech segment using the proposed iterative decomposition approach. $N = 94$.

Segment Number	Last *Iteration* Level	Dominant Frequency of Extracted Mono-Component Signals	Energy of Extracted Mono-Component Signals
1	4	296.9	0.2650×10^{-3}
		972.7	0.1718×10^{-3}
		1885.6	0.0276×10^{-3}
		2235.4	0.0002×10^{-3}
		2538.9	0.0001×10^{-3}
		3090.0	0.0109×10^{-3}
		3515.8	0.0483×10^{-3}
		3813.8	0.0026×10^{-3}
		4051.2	0.0001×10^{-3}
		4345.1	0.0046×10^{-3}
		4679.3	0.0013×10^{-3}
		4984.4	0.0180×10^{-3}
		5369.9	0.0002×10^{-3}
		5520.6	0.0008×10^{-3}
		5912.1	0.0062×10^{-3}
		6483.3	0.0115×10^{-3}
		6785.2	0.0134×10^{-3}
		7376.1	0.0748×10^{-3}
		7641.4	0.0013×10^{-3}
		7981.0	0.0466×10^{-3}

are shown in Fig. 6.27. The severe mode-mixing problem is apparent in the three IMFs depicted in Fig. 6.27 (b), Fig. 6.27 (c). Fig. 6.27 (d). It can be inferred by carefully observing Fig. 6.27, that there are more number of IMFs with oscillations in the low frequency range than IMFs with oscillations in the high frequency range, even when noise-like unvoiced speech signal is known to have more energy scattered in the high frequency range than in the low frequency range [1, 69]. It is evident from this example that the frequency resolution of the EMD degrades in the high frequency range [128]. It can be inferred from Table 6.2, Fig. 6.25, Fig. 6.26 that the advantage of the proposed iterative approach is that it was able to resolve mono-component signals in the low frequency range as well as in the high frequency range at the desired frequency resolution ΔF_{des}.

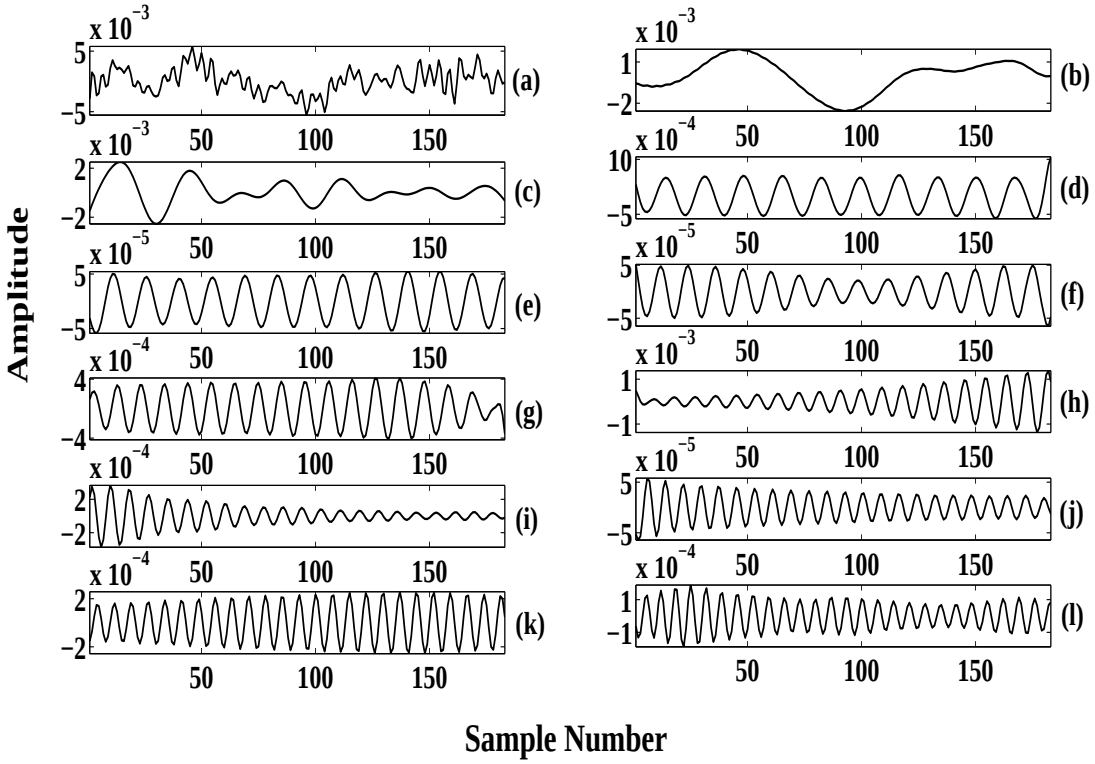

Figure 6.25: (a) Low pass filtered unvoiced speech segment $\breve{x}_1[n]$. Extracted AM-FM mono-component signals using the proposed iterative decomposition approach $\bar{y}_p[n], p = 1, 2, ..., 11$ are shown in (b), (c), (d), (e), (f), (g), (h), (i), (j), (k), (l). In total 20 AM-FM mono-component signals are extracted, 11 of which are depicted in this figure and the rest 9 extracted AM-FM mono-component signals are depicted in the next figure (Fig. 6.26).

6.4.6 Formant analysis

Formants are frequencies at which local peaks occur in the magnitude spectrum of the speech signal. Thus, formants correspond closely to the resonant frequencies of the vocal tract system. The voiced speech signal can be efficiently parameterized using formants. Formant parameters are used in speech synthesizers to produce high quality speech. Noise resilient formant analysis finds use in speech recognition. The method used for formant analysis should result in detection of a few missed or extra formants and must have the ability to track variations in formants [134].

The proposed iterative approach for decomposition of a multi-component signal can be used along with DESA for formant analysis of the speech signal using an appropriate value of STP. DESA employs Teager's non-linear energy operator [134] to compute

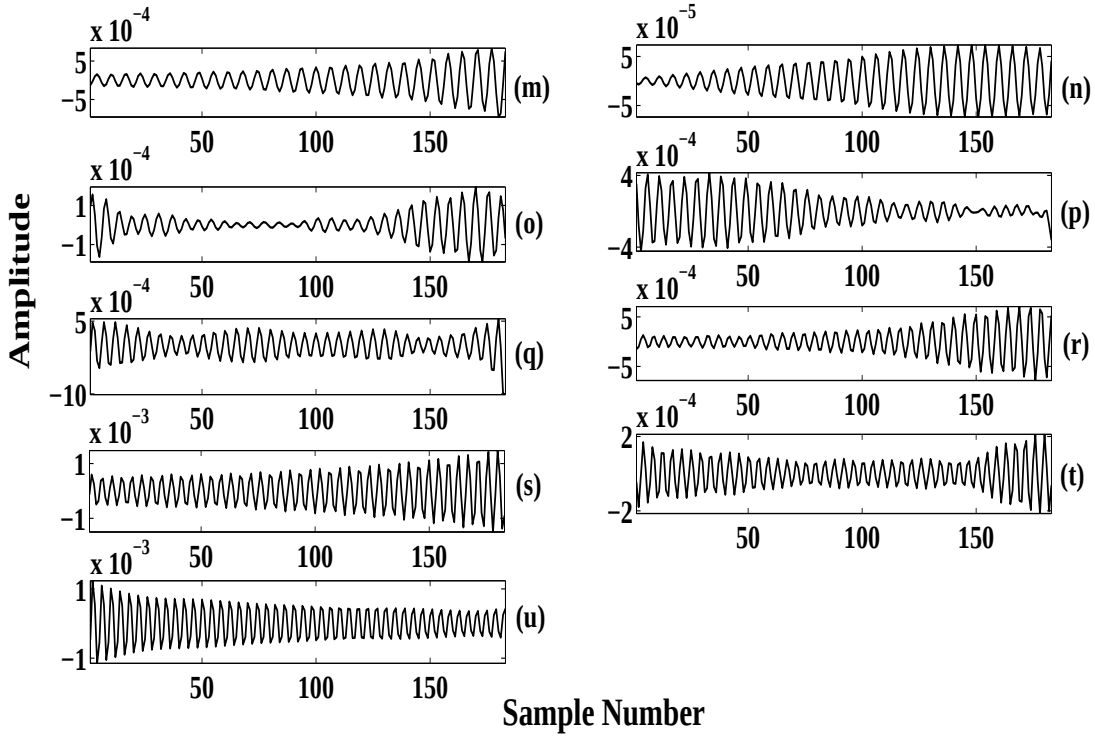

Figure 6.26: Remaining 9 out of the 20 AM-FM mono-component signals $\bar{y}_p[n], p = 12, 13, ..., 20$ extracted from the low pass filtered unvoiced speech segment (shown in the Fig. 6.25 (a)) using the proposed iterative decomposition approach are shown in (m), (n), (o), (p), (q), (r), (s), (t), (u).

instantaneous amplitudes and frequencies of extracted mono-component signals. The Teager-energy of a discrete-time mono-component signal $y[n]$, represented using the operator $\psi(.)$ as $\psi(y[n])$ is computed as [135]:

$$\psi(y[n]) = y^2[n] - y[n-1]y[n+1] \tag{6.27}$$

Teager energies of $y[n]$ and its first order derivative $r[n] = y[n] - y[n-1]$ are used by DESA to compute the instantaneous amplitude and instantaneous angular frequency in

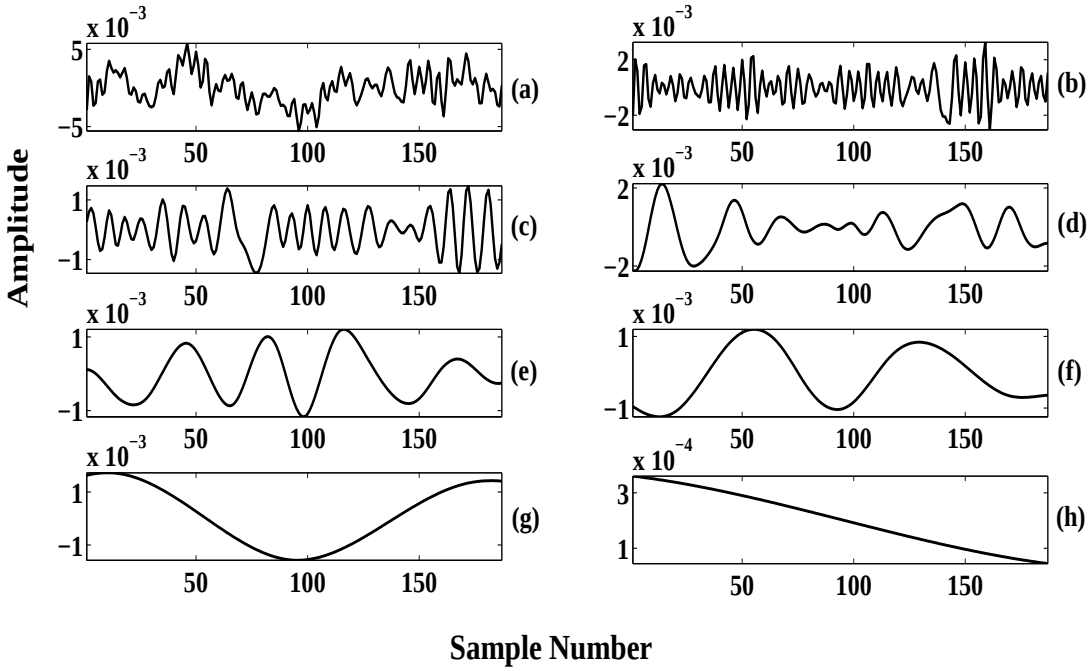

Figure 6.27: (a) Low pass filtered unvoiced speech segment (b) Extracted IMFs using the EMD are shown in (b), (c), (d), (e), (f), (g), (h).

radians/sec of $y[n]$, denoted $A[n]$ and $\omega[n]$ respectively as follows [63, 136]:

$$\begin{aligned} A[n] &\approx \sqrt{\frac{\psi(y[n])}{1-\left(1-\frac{\psi(r[n])+\psi(r[n+1])}{4\psi(y[n])}\right)^2}} \\ \omega[n] &\approx \cos^{-1}\left(1-\frac{\psi(r[n])+\psi(r[n+1])}{4\psi(y[n])}\right) \end{aligned} \tag{6.28}$$

A filter is applied to perform smoothing of the estimated instantaneous amplitude and instantaneous angular frequency to reduce estimation errors [63]. Let us consider an example, a multi-component non-stationary signal $x[n]$ consisting of harmonically related AM-FM mono-component signals given by:

$$x[n] = \sum_l x_l[n] = \sum_l A_l(1+\alpha_l[n])\cos(2\pi l f_0(1+\beta n)n + \theta_l) \tag{6.29}$$

where $n = 0, 1, ..., Q-1$. The values of various parameters used in (6.29) are chosen as: $l = 1, 2, 4, 5, 6, 7, 8, 24, 25$, $F_s = 20000$ Hz, $Q = 3600$, $A_1 = 0.075, A_2 = 0.2, A_4 = 0.3, A_5 =$

$0.7, A_6 = 0.05, A_7 = 0.04, A_8 = 0.25, A_{24} = 0.06, A_{25} = 0.09$, $\alpha_1 = 5/Q, \alpha_2 = \alpha_6 = \alpha_7 = \alpha_{25} = 4.5/Q, \alpha_4 = \alpha_8 = 3.5/Q, \alpha_5 = \alpha_{24} = 4.0/Q$, $F_0 = 120$ Hz, $\theta_1 = \theta_5 = \theta_{24} = 0, \theta_2 = \theta_7 = \pi, \theta_4 = \theta_8 = \pi/2, \theta_6 = \theta_{25} = \pi/3$, $\beta = 0.75/Q$. The value of $\Delta F_{x,\min} = F_0 = 120$ Hz. The value of $\Delta F_{x,\text{des}}$ is chosen to be equal to $\Delta F_{x,\min}$. Therefore, the value of N is chosen to be 360, which is greater than $\dfrac{F_s}{\Delta F_{x,\min}}$. The dominant frequencies and energies of extracted AM-FM mono-component signals of $x[n]$ (given by 6.29) obtained by the proposed iterative decomposition approach in a clean environment and at 5 dB SNR in a white noise environment are compiled in Table 6.3 and Table 6.4 respectively. The energy of white noise is distributed over the entire time-frequency plane. As explained in the third chapter of this thesis that in the presence of noise in $x[n]$, components corresponding to insignificant eigenvalue pairs of H_N^x either correspond to noise contained in $x[n]$ or weak components of $x[n]$. Therefore, the value of STP is chosen relatively higher than previous examples because for formant analysis, only strong mono-component signals of $x[n]$ need to be extracted.

Table 6.3: Extracted AM-FM mono-component signals of $x[n]$ (given by (6.29)) obtained using the proposed iterative decomposition approach in a clean environment. $N = 360, STP = 15\%$.

Segment Number	Last *Iteration* Level	Dominant Frequency of Extracted Mono-Component Signals	Energy of Extracted Mono-Component Signals
1	2	268.0	25.30
		531.0	41.90
		675.0	315.90
		1064.0	24.10

Let $E_{\max}$ denote the maximum energy of extracted AM-FM mono-component signals. The instantaneous frequencies are computed by applying DESA on extracted AM-FM mono-component signals with energy equal to or greater than 5% of $E_{\max}$. The reference instantaneous frequencies of original four strongest mono-component signals of $x[n]$ obtained using (6.29) and instantaneous frequencies of extracted AM-FM mono-component signals computed using DESA in a clean environment and at 5 dB SNR in a white noise environment are depicted in Fig. 6.28. It is evident from Fig. 6.28 that the proposed

Table 6.4: Extracted AM-FM mono-component signals of $x[n]$ (given by (6.29)) using the proposed iterative decomposition approach at 5 dB SNR in a white noise environment. $N = 360, STP = 15\%$.

Segment Number	Last *Iteration* Level	Dominant Frequency of Extracted Mono-Component Signals	Energy of Extracted Mono-Component Signals
1	2	269.0	26.70
		533.0	33.05
		643.0	239.50
		717.0	6.70
		1070.0	23.80

iterative approach along with DESA is able to track instantaneous frequencies of strong mono-component signals contained in $x[n]$ in both clean and noisy environments.

The formant analysis results obtained using the proposed iterative decomposition approach along with DESA on a male voiced segment of 40 m duration in a clean environment and at 5 dB SNR in a white noise environment are depicted in Fig. 6.29. The sampling rate of the male voiced segment is 32 kHz. The value of N is chosen to be 640, which is greater than $F_s/\Delta F_{x,\min} = F_s/F_{0,\min}$, where $F_{0,\min}$ represents the minimum value of the fundamental frequency of the speech signal (F_0) is equal to 50 Hz. The dominant frequencies and energies of extracted AM-FM mono-component signals of the male voiced segment using the proposed iterative decomposition approach in a clean environment and at 5 dB SNR in a white noise environment are tabulated in Table 6.5 and Table 6.6 respectively. It is evident from Table 6.5 and Table 6.6 that the value of $E_{\max}$ is 18.47 and 17.38 in the clean and white noise environment respectively. Please note in Table 6.5 and Table 6.6 that there are two extracted AM-FM mono-component with energies less than 5% of $E_{\max}$. Instantaneous frequencies of only those extracted AM-FM mono-component signals with energies equal to or greater than 5% of $E_{\max}$ are computed using DESA. The extracted AM-FM mono-component signals with instantaneous frequencies spanning a frequency range of 500 Hz or above over the voiced segment duration (40 ms) are also rejected, because such components are assumed to contribute to the overall spectral shape and not the vocal-tract resonances [134]. It is apparent from Fig. 6.29 that the proposed

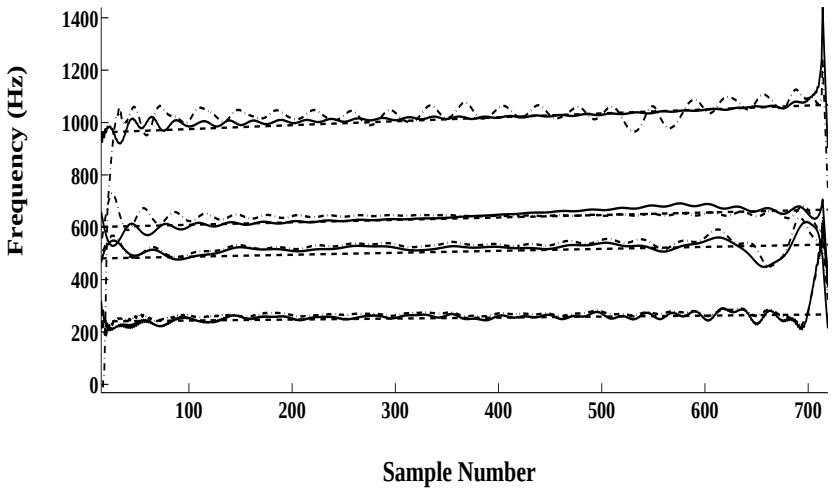

Figure 6.28: Reference instantaneous frequencies of four strongest mono-component signals of $x[n]$ (given by (6.29)) in dashed lines. Instantaneous frequencies of extracted AM-FM mono-component signals of $x[n]$ (given by (6.29)) in a clean environment and at 5 dB SNR in a white noise environment in solid and dash-dotted lines respectively. AM-FM mono-component signals of $x[n]$ (given by (6.29)) are extracted using the proposed iterative decomposition approach. Instantaneous frequencies of extracted mono-component signals are computed using DESA.

iterative decomposition approach has satisfactorily extracted the formant components of the male voiced segment in both clean and noisy environments and instantaneous frequencies of formant components can be tracked by applying DESA on extracted formant components.

Table 6.5: Extracted AM-FM mono-component signals of a clean male voiced segment using the proposed iterative decomposition approach. $N = 640, STP = 15\%$.

Segment Number	Last *Iteration* Level	Dominant Frequency of Extracted Mono-Component Signals	Energy of Extracted Mono-Component Signals
1	2	125.0	0.30
		255.0	1.08
		382.0	0.35
		511.0	4.03
		643.0	18.47
		765.0	1.09
		895.0	3.00

Table 6.6: Extracted AM-FM mono-component signals of a male voiced speech segment at 5 dB SNR in a white noise environment using the proposed iterative decomposition approach. $N = 640, STP = 15\%$.

Segment Number	Last *Iteration* Level	Dominant Frequency of Extracted Mono-Component Signals	Energy of Extracted Mono-Component Signals
1	3	123.0	0.71
		255.0	0.87
		381.0	0.34
		511.0	5.43
		642.0	17.38
		764.0	0.92
		895.0	2.77

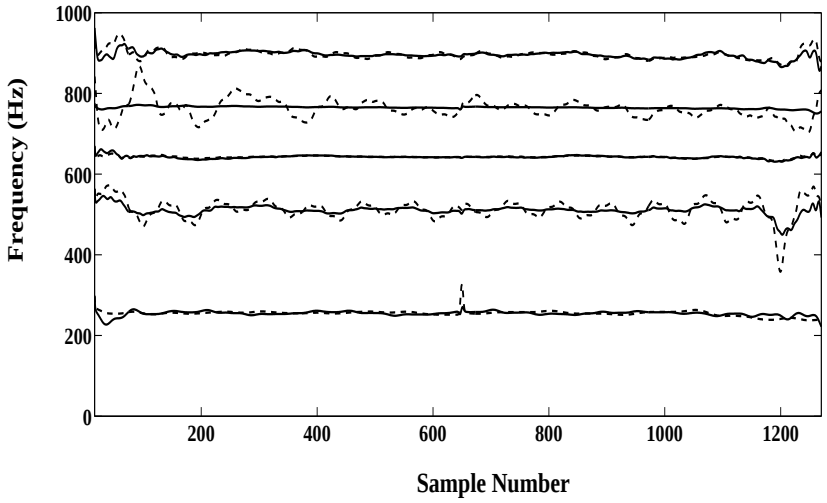

Figure 6.29: Instantaneous frequencies of formant components extracted from a male voiced segment in a clean environment and at 5 dB SNR in a white noise environment in solid and dashed lines respectively. Formant components are extracted using the proposed iterative decomposition approach. Instantaneous frequencies of extracted formant components are computed using DESA.

The formant analysis results obtained using the proposed iterative decomposition approach along with DESA on a female voiced segment of 40 ms duration in a clean environment and at 5 dB SNR in a white noise environment are shown in Fig. 6.30. The sampling rate of the female speech segment is 32 kHz. The value of N is chosen to be 640, which is greater than $F_s/\Delta F_{x,\min} = F_s/F_{0,\min}$, where $F_{0,\min}$ is equal to 50 Hz. The

Table 6.7: Extracted AM-FM mono-component signals of a female voiced speech segment in a clean environment using the proposed iterative decomposition approach. $N = 640, STP = 15\%$.

Segment Number	Last *Iteration* Level	Dominant Frequency of Extracted Mono-Component Signals	Energy of Extracted Mono-Component Signals
1	3	238.0	1.1
		470.0	0.2
		573.0	0.05
		704.0	1.4
		747.0	0.06
		842.0	0.32
		936.0	0.72
		1117.0	0.50

Table 6.8: Extracted AM-FM mono-component signals of a female voiced speech segment at 5 dB SNR in a white noise environment using the proposed iterative decomposition approach. $N = 640, STP = 15\%$.

Segment Number	Last *Iteration* Level	Dominant Frequency of Extracted Mono-Component Signals	Energy of Extracted Mono-Component Signals
1	3	237.0	1.2
		473.0	0.2
		573.0	0.11
		622.0	0.15
		704.0	1.6
		743.0	0.07
		806.0	0.03
		852.0	0.38
		936.0	0.77
		1181.0	0.43
		1416.0	0.02

dominant frequencies and energies of extracted AM-FM mono-component signals of the female voiced segment using the proposed iterative decomposition approach in a clean environment and at 5 dB SNR in a white noise environment are tabulated in Table 6.7 and Table 6.8 respectively. It is evident from Table 6.7 and Table 6.8 that the value of $E_{\max}$ is 1.40 and 1.60 in the clean and noisy environment respectively. Please note that in Table

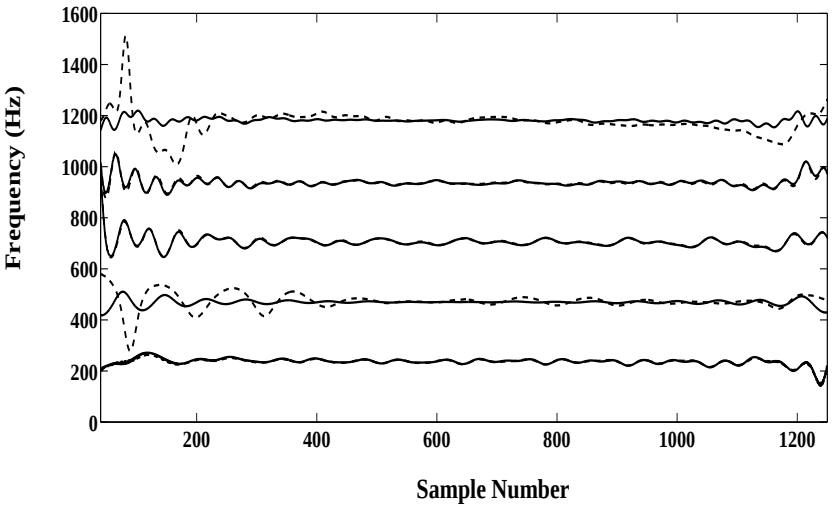

Figure 6.30: Instantaneous frequencies of formant components extracted from a female voiced segment in a clean environment and at 5 dB SNR in a white noise environment in solid and dashed lines respectively. Formant components are extracted using the proposed iterative decomposition approach. Instantaneous frequencies of extracted formant components are computed using DESA.

6.7 and Table 6.8 that there are two or more extracted AM-FM mono-component signals with energies less than 5% of $E_{\max}$. Instantaneous frequencies of only those extracted AM-FM mono-component signals with energies equal to or greater than 5% of $E_{\max}$ are computed using DESA. The extracted AM-FM mono-component signals with instantaneous frequencies spanning a frequency range of 500 Hz or above over the voiced segment duration (40 ms) duration are also rejected, because such components are assumed to contribute to the overall spectral shape and not the vocal-tract resonances [134]. It is apparent from Fig. 6.30 that the proposed iterative decomposition approach has satisfactorily extracted the formant components of the female voiced segment in both clean and noisy environments and instantaneous frequencies of formant components can be tracked by applying DESA on extracted formant components.

6.5 Conclusion

A new iterative approach for decomposition of a multi-component non-stationary signal into AM-FM mono-component signals has been proposed in this chapter. The proposed it-

erative decomposition approach is based on performing repeated eigenvalue decomposition (EVD) of the Hankel matrix, initially constructed from the samples of multi-component non-stationary signal segment. The AM-FM mono-component signals extracted from the multi-component non-stationary signal using the proposed iterative approach are narrow band signals whose instantaneous frequencies can be obtained using the Hilbert transform or DESA.

The proposed iterative approach is adaptive in the sense that the eigenvectors of the Hankel matrix acting as bases for the extracted AM-FM mono-component signals are determined using EVD, which is a data-dependent process. The proposed iterative approach offers flexibility in defining the criterion for significant eigenvalue pair. Therefore, depending upon the type of application either only the strong mono-components signals or strong cum weak mono-component signals of a multi-component signal can be extracted. The proposed decomposition iterative approach is suitably employed for robust formant analysis of the voiced speech signal.

The proposed iterative approach does not require prior information about the number of mono-component signals present in the multi-component non-stationary signal. Unlike the EMD, the proposed approach does not suffer from mode-mixing problem or degradation of the frequency resolution in the high frequency range. The frequency resolution that can be achieved by the proposed iterative decomposition approach can be altered by changing the Hankel matrix size. Unlike the EMD, the ability of the proposed iterative approach to resolve components of a multi-component signal is neither affected by the ratio of their mean frequencies nor by their relative amplitudes.

Chapter 7

Summary

The previous chapters stated the objectives considered in this thesis and described the evolution of algorithms developed to accomplish the identified objectives. This chapter provide the concluding remarks on the obtained analytical results and experimental studies carried out in the previous chapters.

It has been derived and demonstrated that voiced speech signal has significant energy only around the time-varying F_0 and its harmonics in the LFR. The smoothed MEDT over the LFR computed using the PWVD coefficients of the analytic speech signal has been shown to provide an excellent discrimination between the voiced and non-voiced regions of the speech signal. The PWVD provides good time resolution in the LFR and hence, enables instantaneous V/NV detection. It is one of the rare applications of the PWVD where the cross-terms in the PWVD of analytic speech signal have been found to be facilitating the discrimination of the voiced and non-voiced regions of the speech signal by increasing the value of the MEDT over the LFR during the voiced regions of the speech signal.

The analysis of the speech signal in the LFR has been demonstrated to provide a better distinction between voiced and non-voiced regions of the speech signal and enhanced robustness against the high frequency and white noise environments than the analysis of the speech signal in the FFR by eliminating noise energy lying outside the LFR. The filtering of the speech signal in the LFR suppresses formants and renders the time-varying

F_0 component of the voiced speech signal distinguishable among its harmonic components.

The time-varying F_0 component has been extracted from the LFR filtered voiced speech signal without designing any time-varying filter and without using time-frequency analysis techniques. The proposed iterative algorithm has been demonstrated to reliably and efficiently extract the time-varying F_0 component of voiced speech signal even at low SNRs. The filtering of voiced speech signal in the LFR, rejection of non-significant eigenvalue pairs, *Distance Metric* based F_0 range estimation have enhanced the noise robustness of the proposed iterative algorithm. The proposed GCI identification method has been demonstrated to accurately and reliably identify the GCIs in the voiced speech signal by employing the negative cycles of the extracted time-varying F_0 component to provide coarse estimate of intervals where GCIs may occur. The performance of all the algorithms for V/NV detection, GCI identification and F_0 estimation have been found to be better in the white noise environment than the babble noise environment because of the babble noise environment containing higher energy in the LFR.

The proposed iterative decomposition approach to extract AM-FM mono-component signals from a multi-component signal has been shown to be adaptive because the eigenvectors acting as bases functions are data-dependent. The proposed decomposition approach offers flexibility and only the strong or strong-cum weak components of the multi-component signal can be extracted. It has been analytically cum experimentally shown that the frequency resolution that can be achieved by the proposed decomposition approach can be changed by varying the Hankel matrix size. The proposed iterative decomposition approach is suitably employed to perform formant analysis of the voiced speech signal.

REFERENCES

[1] Deller J.R., Hansen J.H.L., Proakis J.G. (2011), Discrete-Time Processing of Speech Signals, Wiley-India, New Delhi.

[2] Drugman T., Bozkurt B., Dutoit T. (January 2012), A comparative study of glottal source estimation techniques, Computer, Speech and Lang, 26(1), 20–34.

[3] Rabiner L. (February 1977), On the use of autocorrelation analysis for pitch detection, IEEE Trans Acoust, Speech and Signal Process, 25(1), 24–33.

[4] Jain P., Pachori R.B. (September 2013), GCI identification from voiced speech using the eigen value decomposition of Hankel matrix, In: Proceedings of IEEE Intl Symposium on Image and Signal Process and Analysis, Treiste, Italy, pp. 371–376.

[5] Bachu R., Kopparthi S., Adapa B., Barkana B.D. (December 2008), Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy, In: Proceedings of Intl Conf Systems, Computing Sci and Software Eng, Bridgeport, USA, pp. 279–282.

[6] Murty K.S.R., Yegnanarayana B. (November 2008), Epoch extraction from speech signals, IEEE Trans Audio, Speech and Lang Process, 16(8), 1602–1613.

[7] Jinachitra P. (September 2006), Glottal closure and opening detection for flexible parametric voice coding, In: Proceedings of Interspeech, Pittsburgh, USA.

[8] Plumpe M.D., Quatieri T.F., Reynolds D.A. (September 1999), Modeling of the glottal flow derivative waveform with application to speaker identification, IEEE Trans Speech and Audio Process, 7(5), 569–586.

[9] Neocleous A., Naylor P.A. (September 1998), Voice source parameters for speaker verification, In: Proceedings of European Signal Process Conf, Rhodes, Greece, pp. 697–700.

[10] Yegnanarayana B., Murty K.S.R. (May 2009), Event-based instantaneous fundamental frequency estimation from speech signals, IEEE Trans Audio, Speech and Lang Process, 17(4), 614–624.

[11] Manfredi C., D'Aniello M., Bruscaglioni P., Ismaelli A. (March 2000), A comparative analysis of fundamental frequency estimation methods with application to pathological voices, Medical Eng and Phy, 22(2), 135–147.

[12] Taori R., Sluijter R.J., Kathmann E. (May 1995), Speech compression using pitch synchronous interpolation, In: Proceedings of IEEE Intl Conf Acoust, Speech and Signal Process, volume 1, Detroit, USA, pp. 512–515.

[13] Kuroiwa Y., Shimamura T. (July 1999), An improvement of LPC based on noise reduction using pitch synchronous addition, In: Proceedings of IEEE Int Symp Circuits and Systems, volume 3, Orlando, USA, pp. 122–125.

[14] Moulines E., Charpentier F. (December 1990), Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Comm, 9(5), 453–467.

[15] Rao K.S. (July 2010), Voice conversion by mapping the speaker-specific features using pitch synchronous approach, Computer, Speech and Lang, 24(3), 474–494.

[16] Kang Y., Tao J., Xu B. (May 2006), Applying pitch target model to convert F0 contour for expressive mandarin speech synthesis, In: Proceedings of IEEE Intl Conf Acoust, Speech and Signal Process, Toulouse, France, pp. 733–736.

[17] Shriberg E., Ferrer L., Kajarekar S., Venkataraman A., Stolcke A. (July 2005), Modeling prosodic feature sequences for speaker recognition, Speech Comm, 46(3-4), 455–472.

[18] Tawari A., Trivedi M.M. (October 2010), Speech emotion analysis: Exploring the role of context, IEEE Trans Multimedia, 12(6), 502–509.

[19] Koolagudi S.G., Reddy R., Rao K.S. (July 2010), Emotion recognition from speech signal using epoch parameters, In: Proceedings of Intl Conf Signal Process and Comm, Bangalore, India, pp. 1–5.

[20] Tawari A., Trivedi M. (August 2010), Speech emotion analysis in noisy real-world environment, In: Proceedings of Intl Conf Pattern Recognition, Istanbul, Turkey, pp. 4605–4608.

[21] Resch B., Nilsson M., Ekman A., Kleijn W.B. (March 2007), Estimation of the instantaneous pitch of speech, IEEE Trans Audio, Speech and Lang Process, 15(3), 813–822.

[22] Schlotthauer G., Torres M.E., Rufiner H.L. (August 2009), A new algorithm for instantaneous F0 speech extraction based on ensemble empirical mode decomposition, In: Proceedings of European Signal Process Conf, Glasgow, Scotland, pp. 2347–2351.

[23] Naylor P.A., Kounoudes A., Gudnason J., Brookes M. (January 2007), Estimation of glottal closure instants in voiced speech using the DYPSA algorithm, IEEE Trans Audio, Speech and Lang Process, 15(1), 34–43.

[24] Drugman T., Dutoit T. (September 2009), Glottal closure and opening instant detection from speech signals, In: Proceedings of Interspeech, Brighton, UK, pp. 2891–2894.

[25] Dhananjaya N., Yegnanarayana B. (March 2010), Voiced/nonvoiced detection based on robustness of voiced epochs, IEEE Signal Process Letters, 17(3), 273–276.

[26] Qiu L., Yang H., Koh S.N. (June 1995), Fundamental frequency determination based on instantaneous frequency estimation, Signal Process, 44(2), 233–241.

[27] Atal B., Rabiner L. (June 1976), A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition, IEEE Trans Acoust, Speech and Signal Process, 24(3), 201–212.

[28] Qi Y., Hunt B.R. (April 1993), Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier, IEEE Trans Speech and Audio Process, 1(2), 250–255.

[29] Kaushik L., O'Shaughnessy D. (September 2008), Voice activity detection using modified Wigner-Ville distribution, In: Proceedings of Interspeech, Brisbane, Australia, pp. 2574–2577.

[30] Davis A., Nordholm S., Togneri R. (March 2006), Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold, IEEE Trans Audio, Speech and Lang Process, 14(2), 412–424.

[31] Tahmasbi R., Rezaei S. (July 2008), Change point detection in GARCH models for voice activity detection, IEEE Trans Audio, Speech and Lang Process, 16(5), 1038–1046.

[32] Shahnaz C., Zhu W.P., Ahmad M.O. (May 2007), An approach for voiced/unvoiced decision of colored noise-corrupted speech, In: Proceedings of IEEE Intl Symp Circuits and Systems, New Orleans, USA, pp. 3944–3947.

[33] Arifianto D. (April 2007), Dual parameters for voiced-unvoiced speech signal determination, In: Proceedings of IEEE Intl Conf Acoust, Speech and Signal Process, Honolulu, Hawaii, pp. 749–752.

[34] Murty K.S.R., Yegnanarayana B., Joseph M.A. (June 2009), Characterization of glottal activity from speech signals, IEEE Signal Process Letters, 16(6), 469–472.

[35] Grenier Y. (August 1983), Time-dependent ARMA modeling of nonstationary signals, IEEE Trans Acoust, Speech and Signal Process, 31(4), 899–911.

[36] Strube H.W. (November 1974), Determination of the instant of glottal closure from the speech wave, J Acoust Soc Am, 56(5), 1625–1629.

[37] Ananthapadmanabha T., Yegnanarayana B. (August 1979), Epoch extraction from linear prediction residual for identification of closed glottis interval, IEEE Trans Acoust, Speech and Signal Process, 27(4), 309–319.

[38] Ma C., Kamp Y., Willems L. (April 1994), A Frobenius norm approach to glottal closure detection from the speech signal, IEEE Trans Speech and Audio Process, 2(2), 258–265.

[39] Hu H., Hsu S., Yu C. (September 2003), Determination of glottal closure instants by harmonic superposition, Signal Process, 83(9), 1985–1995.

[40] Sturmel N., d'Alessandro C., Rigaud F. (April 2009), Glottal closure instant detection using lines of maximum amplitudes (LOMA) of the wavelet transform, In: Proceedings of IEEE Intl Conf Acoust, Speech and Signal Process, Taipei, Taiwan, pp. 4517–4520.

[41] Pachori R.B., Gangashetty S.V. (May 2010), AM-FM model based approach for detection of glottal closure instants, In: Proceedings of Intl Conf Inf Sci Signal Process and their Appli, Kualalumpur, Malaysia, pp. 266–269.

[42] Thomas M.R.P., Gudnason J., Naylor P.A. (January 2012), Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm, IEEE Trans Audio, Speech and Lang Process, 20(1), 82–91.

[43] Drugman T., Thomas M.R.P., Gudnason J., Naylor P.A., Dutoit T. (March 2012), Detection of glottal closure instants from speech signals: A quantitative review, IEEE Trans Audio, Speech, Lang Process, 20(3), 994–1006.

[44] Huang H., Pan J. (April 2006), Speech pitch determination based on Hilbert-Huang transform, Signal Process, 86(4), 792–803.

[45] Cheng Y.M., O'Shaughnessy D. (December 1989), Automatic and reliable estimation of glottal closure instant and period, IEEE Trans Acoust, Speech and Signal Process, 37(12), 1805–1815.

[46] Kadambe S., Boudreaux-Bartels G.F. (March 1992), Application of the wavelet transform for pitch detection of speech signals, IEEE Trans Inf Theory, 38(2), 917–924.

[47] Yin B., Ambikairajah E., Chen F. (April 2009), Voiced/unvoiced pattern-based duration modeling for language identification, In: Proceedings of IEEE Intl Conf on Acoust, Speech and Signal Process, Taipei, Taiwan, pp. 4341–4344.

[48] Paksoy E., Martin J.C., McCree A., Gerlach C.G., Anandakumar A., Lai W.M., Viswanathan V. (March 1999), An adaptive multi-rate speech coder for digital cellular telephony, In: Proceedings of IEEE Intl Conf on Acoust, Speech and Signal Process, Phoenix, USA, pp. 193–196.

[49] Kondoz A.M. (2004), Digital Speech: Coding for Low Bit Rate Communication Systems, Wiley, England.

[50] Sircar P., Saini R.K. (November 2007), Parametric modeling of speech by complex AM and FM signals, Digital Signal Process, 17(6), 1055–1064.

[51] Joho D., Bennewitz M., Behnke S. (April 2007), Pitch estimation using models of voiced speech on three levels, In: Proceedings of IEEE Intl Conf on Acoust, Speech and Signal Process, Honolulu, USA, pp. 1077–1080.

[52] Jancovic P., Kokuler M. (May 2006), Voicing-character estimation of speech spectra: application to noise robust speech recognition, In: Proceedings of IEEE Intl Conf on Acoust, Speech and Signal Process, Toulouse, France, pp. 257–260.

[53] Jang S., Choi S., Kim H., Choi H., Yoon Y. (August 2007), Evaluation of performance of several established pitch detection algorithms in pathological voices, In: Proceedings of IEEE Intl Conf on Eng in Medicine and Biology Society, Lyon, France, pp. 620–623.

[54] Shahnaz C., Zhu W.P., Ahmad M.O. (August 2007), A bifeature voiced/unvoiced discrimination algorithm for speech signals in the presence of noise, In: Proceedings of IEEE Northeast Workshop on Circuits and Systems, Montreal, Canada, pp. 89–92.

[55] Lobo A.P., Loizou P.C. (April 2003), Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition, In: Proceedings of IEEE Intl Conf on Acoust, Speech and Signal Process, HongKong, pp. 820–823.

[56] Fisher E., Tabrikian J., Dubnov S. (March 2006), Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model, IEEE Trans Audio, Speech and Lang Process, 14(2), 502–510.

[57] Classen T.A.C.M., Mecklenbrauker W.F.G. (March 1980), The Wigner distribution: a tool for time-frequency signal analysis - Part 2: Discrete-time signals, Philips J of Research, 35, 276–300.

[58] Kadambe S., Boudreaux-Bartels G.F. (October 1992), A comparision of the existence of cross terms in the Wigner distribution and the squared magnitude of the

wavelet transform and the short-time Fourier transform, IEEE Trans Signal Process, 40(10), 2498–2517.

[59] Pachori R.B., Sircar P. (March 2007), A new technique to reduce cross terms in the Wigner distribution, Digital Signal Process, 17(2), 466–474.

[60] Ferguson B.G., Quinn B.G. (January 1994), Application of the short-time Fourier transform and the Wigner-Ville distribution to the acoustic localization of aircraft, J Acoustical Society of Amer, 96(2), 821–827.

[61] Hlawatsch F., Auger F. (2008), Time-Frequency Analysis: Concepts and Methods, Wiley, New Jersey, USA.

[62] Oppenheim A.V., Schafer R.W. (1989), Discrete-Time Signal Processing, Prentice Hall, Englewood Cliffs, NJ.

[63] Potamianos A., Maragos P. (May 1994), A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation, Signal Process, 37(1), 95–120.

[64] Marple S.L. (September 1999), Computing the discrete-time 'analytic' signal via FFT, IEEE Trans Signal Process, 47(9), 2600–2603.

[65] Loutridis S.J. (July 2006), Instantaneous energy density as a feature for gear fault detection, Mechanical Sys and Signal Process, 20(5), 1239–1253.

[66] Jain P., Pachori R.B. (January 2012), Time-order representation based method for epoch detection from speech signals, J Intelligent Sys, 21(1), 79–95.

[67] Kominek J., Black A. (June 2004), The CMU-Arctic speech databases, In: Proceedings of 5^{th} ISCA Speech Synthesis Workshop, Pittsburgh, USA, pp. 223–224.

[68] CMU-Arctic (2013), www.festvox.org/cmu_arctic/.

[69] Rabiner L.R., Schafer R.W. (2009), Digital Processing of Speech Signals, Pearson Education, India.

[70] Fakotakis N., Tsopanoglou A., Kokkinakis G. (September 1991), Text-independent speaker recognition based on vowel spotting, In: Proceedings of Intl Conf on Digital Process of Signals in Comm, Loughborough, UK, pp. 272 – 277.

[71] NOISEX (2014), www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

[72] Ramirez J., Segura J.C., Benitez C., Torre A., Rubio A.J. (February 2004), A new Kullback-Leibler VAD for speech recognition in noise, IEEE Signal Process Letters, 11(2), 266–269.

[73] Sjolander K., Beskow J. (October 2000), Wavesurfer − An open source speech tool, In: Proceedings of Intl Conf on Spoken Lang Process, Beijing, China, pp. 464–467.

[74] Wavesurfer (2013), www.speech.kth.se/wavesurfer/.

[75] Rao A., Kumaresan R. (May 2000), On decomposing speech into modulated components, IEEE Trans Speech and Audio Process, 8(3), 240–254.

[76] Quatieri T.F. (2002), Discrete-Time Speech Signal Processing: Principles and Practice, Pearson Education, New-Dehli, India.

[77] Gilbert J., Gilbert L. (2005), Linear Algebra and Matrix Theory, Academic Press, New Dehli, India.

[78] Huang N.E., Shen Z., Long S.R., *et al* (November 1996), The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proceedings of Royal Society London A, 454(1971), 903–995.

[79] Pachori R.B., Sircar P. (January 2010), Analysis of multicomponent AM-FM signals using FB-DESA method, Digital Signal Process, 20(1), 42–62.

[80] Gopalan K., Anderson T.R., Cupples E. (May 1999), A comparison of speaker identification results using features based on cepstrum and Fourier-Bessel expansion, IEEE Trans Speech and Audio Process, 7(3), 289–294.

[81] Schroeder J. (April 1993), Signal processing via Fourier-Bessel series expansion, Digital Signal Process, 3(2), 112–124.

[82] Pachori R.B., Sircar P. (February 2008), EEG signal analysis using FB expansion and second-order linear TVAR process, Signal Process, 88(2), 415–420.

[83] Pachori R.B., Sircar P. (2010), Non-stationary Signal Analysis: Methods based on Fourier-Bessel Representation, LAP Lambert Academic Publishing, Saarbrucken, Germany.

[84] DiMonte C.L., Arun K.S. (April 1990), Tracking the frequencies of superimposed time-varying harmonics, In: Proceedings of Intl Conf Acoust, Speech and Signal Process, Albuquerque, USA, pp. 2539–2542.

[85] Zilca R.D., Kingsbury B., Navratil J., Ramaswamy G.N. (March 2006), Pseudo pitch synchronous analysis of speech with applications to speaker recognition, IEEE Trans Audio, Speech and Lang Process, 14(2), 467–478.

[86] Atal B.S., Hanauer S.L. (April 1971), Speech analysis and synthesis by linear prediction of the speech wave, J Acoust Soc Amer, 50(2), 637–655.

[87] Smits R., Yegnanarayana B. (September 1995), Determination of instants of significant excitation in speech using group delay function, IEEE Trans Speech Audio Process, 3(5), 325–333.

[88] Navarro-Mesa J.L., Lleida-Solano E., Moreno-Bilbao A. (August 2001), A new method for epoch detection based on the Cohen's class of time frequency representations, IEEE Signal Process Letters, 8(8), 225–227.

[89] Ang J., Dhillon R., Krupski A., Shriberg E., Stockle A. (September 2002), Prosody-based automatic detection of annoyance and frustration in human-computer dialog, In: Proceedings of Intl Conf Spoken Lang Process, Denver, USA, pp. 2037–2040.

[90] Nakatani T., Okuno H.G. (April 1999), Harmonic sound stream segregation using localization and its application to speech stream segregation, Speech Comm, 27(3-4), 209–222.

[91] Rabiner L., Cheng M., Rosenberg A.E., McGonegal C. (October 1976), A comparative performance study of several pitch detection algorithms, IEEE Trans Acoust, Speech and Signal Process, 24(5), 399–418.

[92] Hess W. (1983), Pitch Determination of Speech Signals: Algorithms and Devices, Springer-Verlag, Berlin, Germany.

[93] Veprek P., Scordilis M.S. (July 2002), Analysis, enhancement and evaluation of five pitch determination techniques, Speech Comm, 37(3-4), 249–270.

[94] Ross M., Shaffer H., Cohen A., Freudberg R., Manley H. (October 1974), Average magnitude difference function pitch extractor, IEEE Trans Acoust, Speech and Signal Process, 22(5), 353–362.

[95] Noll A.M. (February 1967), Cepstrum pitch determination, J Acoust Soc Amer, 41(2), 293–309.

[96] Markel J. (December 1972), The SIFT algorithm for fundamental frequency estimation, IEEE Trans Audio and Electroacoustics, 20(5), 367–377.

[97] Gopalan K. (August 2000), Pitch estimation using a modulation model of speech, In: Proceedings of IEEE Intl Conf Signal Process, volume 2, pp. 786–791.

[98] Hermes D.J. (January 1988), Measurement of pitch by subharmonic summation, J Acoust Soc Amer, 83(1), 257–264.

[99] Shimamura T., Kobayashi H. (October 2001), Weighted autocorrelation for pitch extraction of noisy speech, IEEE Trans Speech and Audio Process, 9(7), 727–730.

[100] Nakatani T., Irino T. (December 2004), Robust and accurate fundamental frequency estimation based on dominant harmonic components, J Acoust Soc Amer, 116(6), 3690–3700.

[101] Shahnaz C., Zhu W.P., Ahmad M.O. (January 2012), Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme, IEEE Trans Audio, Speech and Lang Process, 20(1), 322–335.

[102] Krubsack D., Niederjohn R.J. (February 1991), An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech, IEEE Trans Signal Process, 39(2), 319–329.

[103] Ghosh P.K., Ortega A., Narayanan S. (August, 2007), Pitch period estimation using multipulse model and wavelet transformation, In: Proceedings of Interspeech, Antwerp, Belgium, pp. 2761–2764.

[104] Plante F., Meyer G.F., Ainsworth W.A. (September 1995), A pitch extraction reference database, In: Proceedings of European Conf Speech Comm, Madrid, Spain, pp. 837–840.

[105] Bagshaw P.C., Hiller S.M., Jack M.A. (September 1993), Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching, In: Proceedings of European Conf Speech Comm, volume 2, Berlin, Germany.

[106] Bagshaw P.C. (2013), Evaluating pitch determination algorithms, http://www.cstr.ed.ac.uk/research/projects/fda/.

[107] de Cheveigne A., Kawahara H. (April 2002), YIN, a fundamental frequency estimator for speech and music, J Acoust Soc Amer, 111(4), 1917–1930.

[108] Boersma P., Weenink D. (2012), Praat: Doing Phonetics by Computer (Version: 5.3.21) [Computer Program], http://www.fon.hum.uva.nl/praat/.

[109] Freund R.J., Wilson W.J., Mohr D.J. (2010), Stastical Methods, Academic Press, Burlington, USA.

[110] Childers D.G., Varga R., Perry N.W. (December 1970), Composite signal decomposition, IEEE Trans Audio and Electroacoustics, 18(4), 471–477.

[111] Pantazis Y., Rosec O., Stylianou Y. (February 2011), Adaptive AM-FM signal decomposition with application to speech analysis, IEEE Trans Audio, Speech and Lang Process, 19(2), 290–300.

[112] Pachori R.B., Bajaj V. (December 2011), Analysis of normal and epileptic seizure EEG signals using empirical mode decomposition, Computer Methods and Programs in Biomedicine, 104(3), 373–381.

[113] Stankovic L., Thayaparan T., Dakovic M. (November 2006), Signal decomposition by using the S-method with application to the analysis of HF radar signals in sea-clutter, IEEE Trans Signal Process, 54(11), 4332 – 4342.

[114] Song J., Lin S., Zhao C., Liu H. (May 2011), Decomposition of seismic signal based on Hilbert-Huang transform, In: Proceeding of Intl Conf Business Management and Electronic Inf, volume 1, pp. 813–816.

[115] Gaouda A.M., Salama M.M.A., Sultan M.R., Chikhani A.Y. (October 1999), Power quality detection and classification using wavelet-multiresolution signal decomposition, IEEE Trans Power Delivery, 14(4), 1469–1476.

[116] Bajaj V., Pachori R.B. (November 2012), Classification of seizure and non-seizure EEG signals using empirical mode decomposition, IEEE Trans Inf Tech in Biomedicine, 16(6), 1135–1142.

[117] Tan A.W.C., Rao M.V.C., Sagar B.S.D. (March 2007), A signal subspace approach for speech modelling and classification, Signal Process, 87(3), 500 – 508.

[118] Loughlin P.J., Davidson K.L. (June 2001), Modified Cohen-Lee time-frequency distributions and instantaneous bandwidth of multicomponent signals, IEEE Trans Signal Process, 49(6), 1153–1165.

[119] Cohen L. (1995), Time-Frequency Analysis, Prentice Hall, New-York.

[120] Mallat S. (1998), A Wavelet Tour of Signal Processing, Academic Press, San Diego.

[121] Nielsen M., Kamavuako E.N., Andersen M.M., Lucas M.F., Farina D. (July 2006), Optimal wavelets for biomedical signal compression, Medical and Biological Engineering and Computing, 44(7), 561–568.

[122] Singh B.N., Tiwari A.K. (May 2006), Optimal selection of wavelet basis function applied to ECG signal denoising, Digital Signal Process, 16(3), 275–287.

[123] Choi H., Williams W.J. (June 1989), Improved time-frequency representation of multicomponent signals using exponential kernels, IEEE Trans Acoust, Speech and Signal Process, 37(6), 862–871.

[124] Cohen I., Raz S., Malah D. (January 1999), Adaptive suppression of Wigner interference-terms using shift-invariant wavelet packet decompositions, Signal Process, 73(3), 203–223.

[125] Pachori R.B., Sircar P. (November 2008), Time-frequency analysis using time-order representation and Wigner distribution, Hyderabad, India, pp. 1–6.

[126] Gomez S., Naranjo V., Miralles R. (July 2011), Removing interference components in time-frequency representations using morphological operators, J Visual Comm and Image Representation, 22(5), 401–410.

[127] Deering R., Kaiser J.F. (March 2005), The use of a masking signal to improve empirical mode decomposition, In: Proceedings of IEEE Intl Conf Acoustics, Speech, and Signal Process, volume 4, pp. 485–488.

[128] Gupta R., Kumar A., Bahl R. (March 2012), Estimation of instantaneous frequencies using iterative empirical mode decomposition, Signal, Image and Video Process, 6(1), 1–14.

[129] Rilling G., Flandrin P. (January 2005), One or two frequencies? The empirical mode decomposition answers, IEEE Trans Signal Process, 56(1), 85–95.

[130] Wu Z., Huang N.E. (January 2009), Ensemble empirical mode decomposition: a noise-assisted data analysis method, Advances in Adaptive Data Analysis, 1(1), 1–41.

[131] Friedlander B., Francos J.M. (April 1995), Estimation of amplitude and phase parameters of multicomponent signals, IEEE Trans Signal Process, 43(4), 917–926.

[132] Gazor S., Far R.R. (March 2006), Adaptive maximum windowed likelihood multicomponent AM-FM signal decomposition, IEEE Trans Audio, Speech and Lang Process, 14(2), 479–491.

[133] http://perso.ens-lyon.fr/patrick.flandrin/emd.html.

[134] Wood L.C., Pearce D.J.B. (April 1989), Excitation synchronous formant analysis, IEE Proceedings I Comm, Speech and Vision, 136(2), 110–118.

[135] Kaiser J.F. (April 1990), On a simple algorithm to calculate the 'energy' of a signal, In: Proceedings of IEEE Intl Conf on Acoust, Speech and Signal Process, Albuquerque, NM, pp. 381–384.

[136] Maragos P., Kaiser J.F., Quatieri T.F. (October 1993), Energy separations in signal modulations with application to speech analysis, IEEE Trans Signal Process, 41(10), 3024–3051.

PUBLICATIONS

Journal Publications

1. P. Jain and R.B. Pachori, "Time-order representation based method for epoch detection", *Journal of Intelligent Systems*, vol. 21, no. 1, pp. 79-95, February 2012.

2. P. Jain and R.B. Pachori, "Marginal energy density over the low frequency range as a feature for voiced/non-voiced detection in noisy speech signals", *Journal of the Franklin Institute*, vol. 350, no. 4, pp. 698-716, May 2013.

3. P. Jain and R.B. Pachori, "Event-based method for instantaneous fundamental frequency estimation from voiced speech based on eigenvalue decomposition of the Hankel Matrix", *IEEE Trans Audio, Speech and Lang Process*, vol. 22, no. 10, pp. 1467-1482, October 2014.

4. P. Jain and R.B. Pachori, "A novel iterative Approach for decomposition of multi-component non-stationary signals based on eigenvalue decomposition of the Hankel matrix", Second Revision Submitted in Feb 2015, *Journal of the Franklin Institute*.

Conference Proceedings

1. P. Jain, R.B. Pachori, "A new approach for glottal closure instants detection from speech signals", *Proceedings of Indian International conference on Artificial Intelligence*, pp. 1216-1231, Bangalore, India, December 2011.

2. P. Jain and R.B. Pachori, "GCI Identification from voiced speech using the eigen value decomposition of Hankel matrix", *Proceedings of IEEE Intl Symposium on Image and Signal Process and Analysis*, pp. 371-376, Trieste, Italy, September 2013.