EPOCH BASED DYNAMIC PROSODY MODIFICATION FOR NEUTRAL TO EXPRESSIVE SPEECH CONVERSION

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

GOVIND. D

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, INDIA

JULY 2013

Certificate

This is to certify that the thesis entitled "**EPOCH BASED DYNAMIC PROSODY MODIFICATION FOR NEUTRAL TO EXPRESSIVE SPEECH CONVERSION**", submitted by **Govind. D** (07610213), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

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

To

My beloved parents

N. T. Seethalakshmy and P. G. Divakaran

for their love and encouragements

My guide

Prof. S. R. Mahadeva Prasanna

for his guidance and inspiration

My wife

Gayathri. V

for her love and support

&

My Family

Gopal, Gautham, Smt. Bindu V, Sri. Ananda Padmanabhan, Smt. Remani Menon and Smt. Valsala Menon

for their love and continuous motivation

Acknowledgements

This thesis would not have been possible without the sincere efforts and constant motivation given by my guide, Prof. S. R. M. Prasanna. Therefore I dedicate this thesis to him. I was able to carry out a smooth research due to the immense facilities provided by him in Electro-Medical & Speech Technology Laboratory. I also would like to sincerely thank him for providing me with learning opportunities, financial and moral support throughout the time I had spent for PhD in IIT Guwahati. Among the many opportunities provided by him,I consider the fruitful interactions with Prof. Yegnanarayana, a renowned personality in speech processing research, as the golden opportunity which helped me to formulate the core part of my research work. Also I benefitted from the discussions with Prof. Sreenivasa Rao during his visits to IIT Guwahati.

I would like to express my sincere thanks to the doctoral committee chairman, Prof. S. Dandapat (sir), and other members, Dr. Rohit Sinha (sir) and Dr. P. K. Das (sir) for their suggestions in my research work.

I am also thankful to my teachers of EEE Department of IIT Guwahati, Prof. P. K. Bora (sir), Prof. Chithralekha mahantha (madam) for their support in my research work. I enjoyed working with Dr. Shakuntala Mahanta and Dr. Priyankoo Sarmah of Humanities & Social Science department in IIT Guwahati for speech prosody 2012.

I would like to acknowledge the sincere efforts by Dr. P. Krishnamurthy (PKM) for his suggestions during our discussions and providing a preliminary framework for my research. I also would like to thank my other senior members of EMST Lab Dr. Manikandan, Dr. Nirmala (madam), Dr. Jayanna (sir), Dr. Debadatta Pati (sir) for their overwhelming support during my PhD course. My special thanks to Dr. L. N. Sharma (sir) for maintaining an excellent computing facility and various resources useful for the research work.

I would like thank my friend Mr. Haris B. C. for his technical support during the final stages of my PhD. His efforts provided a smooth functioning of my work in the EMST Lab. Also I cannot forget the help and support I got from my colleague Mrs Sumitra Shukla (didi) for my research work.

I also would like to thank members of Center for Speech Technology Research (CSTR), University of Edinburgh, Prof. Simon King, Dr. Robert Clark, Dr. Korin Richmond and Dr. Junichi Yamagishi for the discussions and support during my stay in Edinburgh. I would like to acknowledge the help

Dr. Oliver Watts (Oliver) and Dr. Sebastian Anderson (Seb) for their help for collecting my data for speech synthesis. I would like to acknowledge the fruitful discussions that I had with Dr. J. P. Cabral during his visit to IIT Guwahati.

I also highly appreciate support of my friends Gayadhar Pradhan (Pradhanji), Deepak, Biswajit, Nagaraj, Ramesh and Syed, Rohan, Bandita, Aniruddh and Sandeep Reddy for my work.

I thank Dr. Felix Burkhardt, Technical University of Berlin to provide me with the EGG recordings of German emotional speech database.

I would like to convey my sincere thanks to Dr. Sathya Sai Prakash (sir) who was my teacher in Amrita Vishwa Vidyapeetham, for the encouragements and suggestions for joining PhD programme. I also would like to thank Dr. Suryakanth V Gangashetty (sir) for channeling me to pursue PhD studies in IIT Guwahati.

I cannot forget the basic signal processing lessons taught by Prof. K. P. Soman, Head CEN, Amrita Vishwa Vidyapeetham (University), which helped me lot later during my PhD studies. Also for the support and opening he has given me during the final stages of my PhD.

I also admit and respect the efforts by Adv. V. D. Satheesan (sir), M. L. A. of N. Parur in Eranakulam district of Kerala, during my hard times in final year B-Tech days.

Finally, for describing the amazing support given by my wife, Gayathri, during the final stages of my thesis is beyond my words.

Govind. D

Abstract

The objective of this thesis is to address the issues in the analysis, estimation and incorporation of prosodic parameters for neutral to expressive speech conversion. The prosodic parameters like instantaneous pitch, duration and strength of excitation are used as the expression dependent parameters. For the expressive speech analysis, refinements in the conventional methods are proposed to accurately estimate the prosodic parameters from different expressions. The variations in the prosodic parameters for different expressions are compared with respect to the neutral expression. The expressive speech is synthesized by modifying the prosodic parameters of the neutral speech according to the variations in the target expression. The variations in the prosodic parameters are incorporated by epoch based prosody modification. Epochs represent the instants of glottal closure in voiced speech and onset of burst or frication in unvoiced speech. The improved perceptual quality in the prosody modified speech is obtained by accurately estimating epochs location in epoch based prosody modification. A computationally efficient and perceptually improved epoch based prosody modification is initially proposed for incorporating static prosodic variations for different expressions. As the prosodic parameters of the expressions vary dynamically with respect to the corresponding neutral speech, an epoch based dynamic prosody modification method is then proposed for incorporating dynamic variations in the prosodic parameters. Finally, the significance of dynamic prosody modification is demonstrated and evaluated for neutral to expressive speech conversion for text dependent and speaker dependent, text dependent and speaker independent and text independent and speaker independent scenarios.

The major contributions of this thesis are as follows:

- Refined method for accurate estimation of prosodic parameters for expressive speech analysis
- Dynamic prosody modification method for incorporating dynamic variations in the prosodic parameters due to emotions.
- Demonstrated the effectiveness of the dynamic prosody modification for neutral to expressive speech conversion

The other contributions of this thesis are as follows:

- Identified degradation in the epoch extraction performance from expressive speech signals using conventional approaches .
- A computationally fast and perceptually improved epoch based static prosody modification.
- Significance of glottal activity detection for improving the naturalness of static and dynamic duration modification
- A general framework for dynamic prosody modification is proposed for the conventional methods

Keywords: Expressions, emotions, neutral to expressive speech conversion, prosody modification, epochs, dynamic prosody modification, glottal activity detection.

Contents

List of Figures

List of Tables

List of Acronyms

AMDF	Average Magnitude Difference Function
CART	Classification And Regression Tree
CMOS	Comparison Mean Opinion Scores
CREST	Core Research for Evolutional Science and Technology
DFT	Discrete Fourier Transform
DC	Direct Current
DTW	Dynamic Time Warping
DCT	Discrete Cosine Transform
DYPSA	DYnamic Programming based projected Phase Slope Algorithm
ESPS	Entropic Signal Processing System
EGG	Electroglottogram
EMA	Electro Magnetic Articulography
EPG	Electro Palatography
EGD-RM	Epoch based Group Delay Residual Modification
EZFF-RM	Epoch based Zero Frequency Filtering Residual Modification
EZFF-SM	Epoch based Zero Frequency Filtering Speech Modification
ESS	Expressive Speech Synthesis
ESP	Expressive Speech Processing
EMG	Electromyogram
FD-PSOLA	Frequency Domain Pitch Synchronous Overlap Add
FFNN	Feed Forward Neural Network
FFT	Fast Fourier Transformation
IFFT	Inverse Fast Fourier Transformation
FAR	False Alarm Rate

GA	Glottal Activity
GCI	Glottal Closure Instants
GOI	Glottal Opening Instants
GD	Group Delay
HE	Hilbert Envelope
HMM	Hidden Markov Model
IDCT	Inverse Discrete Cosine Transformation
IDFT	Inverse Discrete Fourier Transformation
IDR	IDentification Rate
IDA	IDentification Accuracy
JST	Japan Science and Technology
LP-PSOLA	Linear Prediction Pitch Synchronous Overlap Add
LP	Linear Prediction
LPC	Linear Prediction Coefficient
LF	Lijencrants and Fant
LSF	Line Spectral Frequency
LPCC	Linear Prediction Cepstral Coefficient
MLPG	Maximum Likelihood Parameter Generation Algorithm
MLSA	MeL Spectral Approximation
MRI	Magnetic Resonance Imaging
MOS	Mean Opinion Scores
MR	Miss Rate
NSS	Neutral Speech Synthesis
OQ	Open Quotient
OLA	OverLap Add
PSOLA	Pitch Synchronous Overlap Add
PSTS	Pitch Synchronous Time Scaling
RQ	Return Quotient
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
STFT	Short Time Fourier Transform

SIFT	Simple Inverse Filtering Technique
SQ	Speed Quotient
SOLA	Synchronous OverLap Add
TTS	Text To Speech
TD-PSOLA	Time Domain Pitch Synchronous Overlap Add
VT	Vocal Tract
ZFFS-SM	Zero Frequency Filtered Signal based Speech Modification
ZFF-RM	Zero Frequency Filtering Residual Modification
ZFF	Zero Frequency Filtering
ZFFS	Zero Frequency Filtered Signal
ZFR	Zero Frequency Resonator

List of Symbols

F_0	Fundamental frequency
F_{0Avg}	Mean value of F_0 contour
F_0max	Maximum value of the F_0 contour
F_0min	Minimum F_0 value of the F_0 contour
F_1	First formant frequency
F_2	Second formant frequency
F_3	Third formant frequency
F_4	Fourth formant frequency
F_5	Fifth formant frequency
α	Pitch modification factor
β	Duration modification factor
γ	Strength modification factor
$s(n)$	Speech signal
$x(n)$	Differenced speech signal
$e(n)$	Linear prediction residual
$E(\omega)$	Fourier transform of LP residual
$\tau(\omega)$	Group delay function
$\bar{\tau}(\omega)$	Phase slope function
N_a	Return phase
N_e	Peak flow duration
N_c	Closed phase
N_{op}	Opening phase
N_{cl}	Closing phase
A_k	Amplitude factor of k^{th} formant frequency

B_k	Bandwidth corresponds to k^{th} formant frequency
θ_k	Angle corresponds to k^{th} formant frequency in the unit circle
ζ	Epoch identification error
σ	Standard deviation
-$\phi'(\omega)$	Negative derivative of fourier transform phase

1

Introduction

Contents

1.1 Objective of the Thesis

In natural human-human communication, along with the message in speech, speaker conveys crucial non-linguistic information also to the listener. This information can be the mental state of the speaker, personality, emotional state of the speaker etc. and are conveyed through various expressions. The speaker conveys these expressions in terms of body reactions, gestures, facial expressions or through speech. The listener decodes these expressions, understands and accordingly adjusts his style of speaking in an effortless manner and makes the communication more natural and effective. The incorporation of these expressions related to the emotions, personality or style of the speaker into the speech is increasingly popular trend among the speech synthesis community to have an effective and natural human-computer interaction. Such a system which incorporates different expressions to the synthesized speech is termed as expressive speech synthesis (ESS) system. In ESS, the speech synthesized by the conventional text to speech synthesis system is converted by incorporating the expressive information related to the target expression. The neutral speech synthesized by the neutral text to speech synthesis (NSS) system followed by the neutral to expressive speech conversion form important modules of an ESS system. The three stages involved in neutral to expressive speech conversion are analysis, estimation and incorporation, of expression specific speech parameters into the neutral speech. In the analysis stage, different expressions are analyzed for identifying significant expression specific speech parameters from an expressive speech database. These expression specific speech parameters are accurately estimated from the expressive speech in the estimation stage. Finally, methods to incorporate these expressive parameters for the expressive speech conversion are devised in the incorporation stage. The present work reported in this thesis demonstrates the significance of incorporating dynamic variations in prosodic parameters for neutral to expressive speech conversion. To reduce the perceptual distortions in neutral to expressive speech conversion, the prosody modification is performed by anchoring around accurate epochs location estimated from the neutral speech. The work proposed in this thesis is therefore termed as **epoch based dynamic prosody modification for neutral to expressive speech conversion**.

1.2 Significance of Expressive Speech Synthesis

The expressions in speech carry extra-linguistic information about the context, add expressiveness and characterize the mental state of the speaker. The dictionary meaning of expression is conveying

a thought or an emotion. Expression is defined as the vocal indicators of various emotional state that reflect in the speech waveform [1]. Different emotions and speaking styles are also considered as expressions [2]. Hence in this thesis the term expression and emotion are interchangeably used. Expressive speech synthesis deals with incorporating these expressive information in speech. Listeners effortlessly detect the expressive content from the speech transmitted by the speaker, understand the behavioral characteristics of the speaker and adapt the speaking style according to the mental state of the speaker. Hence, expressive speech synthesis finds application in dialogue systems. Also incorporating emotions in the speech makes the synthesized speech more natural. The other applications of the expressive speech synthesis include,

- **Child Interfaces**: For story telling applications for children.
- **Animation Cartoons**: For dubbing the voice for different characters in an animation cartoon.
- **Call Centers**: Expressive speech analysis can be employed for determining the emotional state of the customers and give the reply by incorporating required expression to please the customer.
- **Public Address Systems**: Announcements in different styles. For example, for announcing the sad news in sad expression and happy news in happy expression.

1.3 Issues of Neutral to Expressive Speech Conversion

Text to speech synthesis is the process of converting the message in the textual form to the message in the spoken form. The quality of the synthesized speech is assessed in terms of intelligibility and naturalness. The intelligibility refers to how well the message in the text is conveyed in the synthesized speech. The naturalness refers to the similarity of the synthesized speech with the human speech. Due to various developments in the speech synthesis area, synthetic speech with sufficient intelligibility is already achieved such that the listeners can easily recognize the message in the generated speech [3]. Even though highly intelligible and naturally sounding speech can be synthesized by the state of the art unit selection or HMM based statistical parametric speech synthesis systems, the naturalness has not grown to the level of intelligibility that is achieved in the synthetic speech [4, 5]. Hence there is a great interest in the speech synthesis community to develop expressive speech systems that can incorporate various expressions related to human behavior. Such systems can be used effectively for human computer interaction [6].

The neutral to expressive speech conversion serves as the back end post processing unit of an ESS system. The front end processing module synthesizes the neutral speech from the input text. The synthesized neutral speech is then converted to expressive speech by the neutral to expressive speech conversion module according to the target expression. The speech parameters that vary according to expressions of the neutral speech are modified according to the target emotion. The neutral to emotion conversion is done in the following two stages [7, 8]:

- Analysis and estimation of expression dependent parameters
- Incorporation of expressive parameters for the neutral to emotion conversion

In the expressive speech analysis stage, the speech parameters of various expressions are analyzed for identifying the expression dependent speech parameters. The expression dependent parameters are estimated to analyze how these parameters are varying for different expressions with respect to neutral expressive speech. Variations due to expressive information are then incorporated in the neutral speech signal by an expressive parameter modification algorithm.

1.4 Neutral to Expressive Speech Conversion by Epoch based Prosody Modification

1.4.1 Expressive speech analysis

The expressive parameters are analyzed for each expression in the database for the same speaker and and same text. The parameters of the prosody of speech have been used as the important parameters related to expressions. In the past, researchers have shown that how prosodic parameters vary for different expressions in an average sense [9, 10]. Hence the prosodic parameters like average pitch, duration and intensity are analyzed across different expressions in the expressive speech analysis stage. The scaling value for each prosodic parameter is then derived by scaling the average value obtained for each prosodic parameter for each expression with the average prosodic parameter value obtained for the neutral speech. For neutral to expressive speech conversion, the prosodic parameters of the neutral speech are modified according to the corresponding factors derived for the target expression. The prosody modification algorithm is used to modify the pitch, duration and intensity parameters of the neutral speech.

1.4.2 Epoch based prosody modification

Prosody modification is the process of modifying the pitch, duration and intensity of a given speech without affecting the perceptual quality of the speech. There are both time domain and frequency domain methods for prosody modification of speech. Among these popular time domain approaches are time domain pitch synchronous overlap add (TD-PSOLA) and linear prediction pitch synchronous overlap add (LP-PSOLA). In PSOLA approach, the speech is divided into different analysis frames having length of 2 to 3 pitch periods. These analysis frames are centered around the pitch marks of the given speech. The synthesis pitch marks are generated according to the prosody modification factors. The analysis frames of the original speech are then copied to the nearest synthesis pitch marks in an overlap add manner with the preceding and succeeding analysis frames. The quality of the prosody modified speech depends on the accuracy with which the analysis pitch marks are estimated. Since the analysis pitch marks used in epoch based prosody modification are accurate, the epoch based prosody modification method can be used for improving the perceptual quality in prosody modification [11].

Epochs represent instants of glottal closure in case of voiced speech and onset of burst or frication in case of unvoiced speech [12]. Epochs provide accurate locations of pitch marks in voiced speech [13]. In epoch based prosody modification, the prosody modification of speech is performed using epochs as the analysis pitch marks. In the existing epoch based prosody modification, the epochs are estimated by the group delay analysis of the linear prediction (LP) residual [11]. Unlike in PSOLA method for prosody modification, epoch based prosody modification uses no voiced-unvoiced speech detection for prosody modification. Due to the improved perceptual quality in prosody modification, the work presented in this thesis uses the epoch based prosody modification for neutral to expressive speech conversion.

1.5 Scope of the Present Work

For addressing the issues presented for neutral to emotion conversion for expressive speech synthesis applications, the scope of the work presented in this thesis are the following:

(i) In general most of the expressive speech analysis methods use conventional methods such as autocorrelation, robust pitch estimation and etc. for estimating prosodic parameters for neutral to expressive speech conversion [8, 14, 15]. As the quality of the synthesized expression depends on the scaling factors derived in the expressive speech analysis stage, the prosodic parameters

have to be estimated more accurately in the expressive speech analysis stage for neutral to expressive speech conversion. Also, performance analysis of existing techniques for estimating prosodic parameters are not compared among different expressions.

(ii) As the epochs location in speech provide accurate pitch markers, the epoch based analysis is to be employed for deriving prosodic parameters. The accuracy of epoch estimation performance needs to be compared among neutral and different emotion speech signals.

(iii) In Most of the works, the static modification of the prosodic parameters of the neutral speech is performed according to fixed scaling factors for the neutral to expressive speech conversion. The dynamic variations in prosodic parameters due to different expressions are not considered in the existing methods for expressive speech conversion

(iv) For improved perceptual quality in the prosody modification, the epoch based prosody modification is employed for neutral to expressive speech conversion. Epoch based prosody modification introduces minimum distortion in the prosody modified speech at the cost of increased computational complexity. Hence a computationally fast prosody modification method is essential for real time neutral to expressive speech conversion. Also, the existing epoch based prosody modification is proposed for the static prosody modification of the prosodic parameters. Hence the method is not suitable for incorporating dynamic variations in the prosodic parameters due to expressions.

Motivated from these observations, the primary objective of the work presented in this thesis is to demonstrate the significance of incorporating dynamic prosodic variations for effective neutral to expressive speech conversion. The effectiveness of the neutral to expressive speech conversion is demonstrated through the following works:

- Refinement in the estimation of prosodic parameters for accurate expressive speech analysis
- Development of dynamic prosody modification tool for dynamic variations in prosodic parameters
- Demonstrating the effectiveness of dynamic prosody modification for neutral to expressive speech conversion

1.6 Organization of the Thesis

The rest of the thesis is organized as follows:

Chapter 2 provides a detailed review on the neutral to expressive speech conversion. After reviewing the ESS systems in Section 2.3, the chapter focuses on the issues in neutral to expressive speech conversion in Section 2.4. The rest of the chapter describes the various studies carried out in the literature for the analysis, estimation and incorporation of expressive parameters for neutral to expressive speech conversion. Summary of the review and scope for present work are given in Section 2.8. The organization of the present work is given in Section 2.9.

Chapter 3 describes expressive speech analysis based on the proposed method of accurate estimation of prosodic parameters from various expressions. The emotion speech analysis performed in different emotion speech databases are also presented in Chapter 3

Chapter 4 proposes epoch based dynamic prosody modification method for neutral to emotion conversion. Perceptually improved and computationally fast epoch based prosody modification is proposed in Section 4.3. Section 4.4 describes the proposed epoch based dynamic prosody modification for incorporating dynamic variations due to different expressions. The subjective studies conducted for evaluating the proposed static and dynamic prosody modification are given in Section 4.5.

Chapter 5 demonstrates the effectiveness of dynamic prosody modification in neutral to expressive speech conversion. The effectiveness of dynamic prosody modification in neutral to expressive speech conversion is demonstrated for text dependent and speaker dependent, text dependent and speaker independent and text independent and speaker independent scenarios. Section 5.6 describes the experimental results to demonstrate the effectiveness of dynamic prosody modification over static prosody modification for neutral to emotion conversion by comparative subjective evaluations.

Finally Chapter 6 summarizes the present work, lists the major contributions of the thesis and provides the scope for future work.

2

Expressive Speech Synthesis - A Review

Contents

2.1 Objective

The objective of this chapter is to provide a detailed review of expressive speech synthesis (ESS) by neutral to expressive speech conversion. Among various approaches for ESS, the present work focusses on the development of ESS systems by explicit control. In this approach, the ESS is achieved by modifying the parameters of the neutral speech which is synthesized from the text. This chapter also reviews the works addressing various issues related to the development of ESS systems by explicit control. The review provided in this chapter includes, review of various approaches for text to speech synthesis, various studies on the analysis and estimation of expressive parameters and various studies on methods to incorporate expressive parameters.

2.2 Introduction

Speech synthesis is the process of converting message written in text to equivalent message in spoken form. Expressive speech synthesis deals with synthesizing speech and adding various expressions related to different emotions and speaking styles to the synthesized speech [2] [7] [16] [17]. The dictionary meaning of expression is conveying a thought or an emotion. The expression is defined as the vocal indicator of various emotional states that reflect in the speech waveform [1]. The different emotions and speaking styles are also considered as expressions [2]. Based on this, in the present work, we have considered different emotions as the expressions and hence emotions and expressions are interchangeably used.

The objective of speech synthesis is to synthesize speech waveform from the text. The Schematic block diagram of a speech synthesis system is shown in Figure 2.1. The input text is first converted into abstract linguistic representation by the front end text processing stage. This linguistic representation is obtained by performing prosodic annotations on the syntactic, semantic and lexically analyzed text [18]. This linguistic representation drives the synthesis routines to get the speech waveform of the input text [4, 19, 20]. In the present work, such a system is termed as Neutral Speech Synthesis (NSS) system.

In expressive speech synthesis, along with text, the desired expression also forms an additional input to the text processing stage as shown in Figure 2.2. In expressive speech synthesis, along with the linguistic features of the input text, the expressive information is also incorporated, either before or after the synthesis of neutral speech. In the former case, the expressive information is coded along with

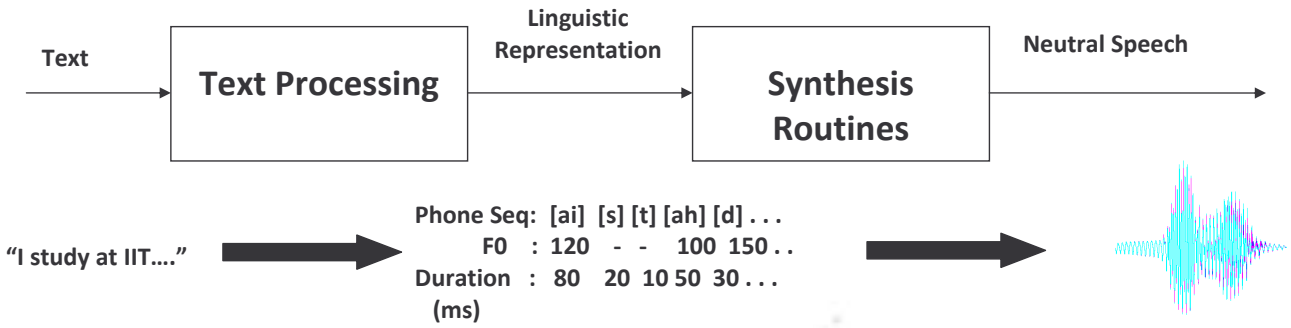

Figure 2.1: Schematic diagram of Neutral Speech Synthesis

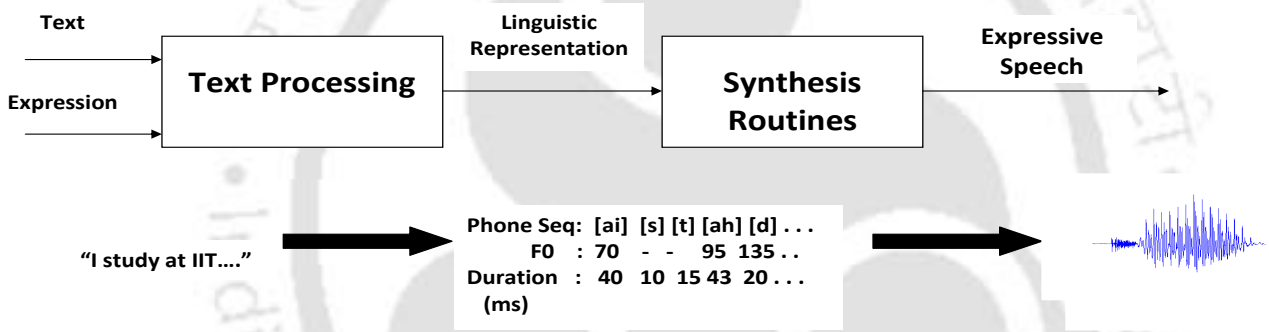

Figure 2.2: Schematic diagram of Expressive Speech Synthesis

the linguistic information and speech is synthesized from the text using the linguistic and expressive information. In the later case, the speech is synthesized initially without any expression, that is, neutral speech and then later the desired expression is added using a suitable voice transformation technique [7] [17].

Speech synthesized in different expressions can be used in story telling applications for children where for effectiveness and drawing attention, different expressions have to be generated in different contexts of the story [21]. ESS can be used as a part of dialogue system which makes the human computer interaction more natural and effective [22]. Expressive speech analysis can be utilized by the call center managers to identify the emotional state of the operators during conversing with the customers and valuate them based on their emotional maturity. ESS finds application in the financial

information system to make announcements in different speaking styles to the users [2].

The rest of the chapter is organized as follows: The review of existing approaches for the development of ESS systems are presented in Section 2.3. Among various approaches for the development of ESS systems, the present focus of the development of ESS by explicit control and the issues in that are described in Section 2.4. In ESS by explicit control approach, the perceptual quality of the synthesized expressive speech deeply depended on the quality of the synthesized neutral speech, Section 2.5 reviews different NSS approaches. The various works on analysis and estimation of expressive parameters are explained in Section 2.6. Section 2.7 reviews the methods to incorporate the expressive parameters. Finally, the scope for the present work is given in Section 2.9

2.3 Review of Existing Expressive Speech Synthesis Systems

This section reviews various existing approaches employed for expressive speech synthesis. According to Schroder, expressive speech synthesis approaches can be broadly classified into the following three categories [23].

- **Expressive speech synthesis by explicit control**
- **Expressive speech synthesis by playback approach**
- **Expressive speech synthesis by implicit control**

2.3.1 Expressive speech synthesis by explicit control

Here, the expressive speech is synthesized by modifying the neutrally synthesized speech based on the prosodic rules derived from the expressive speech database of the respective expressions. The expressive speech synthesis systems developed on formant synthesis [10, 24, 25] and diphone concatenation are examples of explicit control [15, 26]. Various methods developed for neutral to expressive conversion tasks [7, 8, 21] also falls under the category of explicit control.

Due to flexibility in controlling various source and system parameters, early developments of expressive speech synthesis systems were built on top of the formant speech synthesis systems [27]. The affect editor developed by Cahn was the first attempt to synthesize emotional speech using a formant synthesizer [10, 27]. The control parameters of formant synthesizer are manually tuned for each of the emotions to synthesize the expressive speech. Modification of each of the control parameters for each emotion is performed according to various acoustic profiles discussed in the literature [9] [28].

HAMLET, emotion speech synthesis system developed by Murray and Arnott, is a rule based system developed on commercial formant speech synthesis system called DECtalk [29]. In HAMLET the pitch and duration rules and voice quality rules are set in the formant synthesizer and quality of the synthesized emotions are improved heuristically by manual tuning. The development of these rules for emotions is as given in [24]. The objective of the perceptual experiments conducted by Burkhardt was to, find out the perceptually relevant acoustic features for each emotion by systematically varying these acoustic parameters during the synthesis of the neutral utterances and find the optimum values of each of the acoustic features for the emotional speech synthesis [25]. According to these perceptual experiments, suprasegmental features like mean pitch and pitch range, speech rate, and voice quality parameters like phonation and vowel precisions, are found to be significant for effectively synthesizing emotions using formant synthesizers. The studies conducted by Vroomen *et.al* on seven emotions (neutral, joy, boredom, anger, sadness, fear, indignation) showed that only intonation and duration are enough to express emotions in the synthesized speech using a diphone synthesizer [15]. Here emotional speech is synthesized by manipulating pitch and duration using Pitch Synchronous Overlap Add (PSOLA) of the neutrally synthesized speech. The significance of pitch and duration parameters in emotional speech synthesis is also shown in the studies by [26] in Spanish using diphone concatenation. This study also showed that the relative contribution of prosody and voice quality depends on the emotions to be synthesized [23].

Apart from expressive speech synthesis systems developed based on various speech synthesis approaches, there are some works described in the literature for neutral to target expressive speech conversion task using the explicit control approach. Tao *et al.* achieved expressive speech conversion by prosody (pitch and duration) modification of the neutral expressive speech [7]. This paper compares linear, Gaussian mixture model (GMM) and classification and regression tree (CART) methods for converting neutral speech to target expressive speech for Mandarin language. Apart from discrete emotions like angry, happy, sad and fear, the strong, medium and weak versions of each is also considered for synthesis. Direct scaling of sentence F_0 and syllable duration is done in linear modification model and other acoustic features of F_0 contour considered for modification are $F_{0topline}$,$F_{0baseline}$,F_{0avg} and intensity. Where, $F_{0topline}$ and $F_{0baseline}$ are the F_{0Max} and F_{0Min}, respectively [7]. In GMM based prosody modification, pitch target models are constructed from the tonal representation of the intonation pattern of each syllable for each expression. The pitch target model parameters generated

by GMM of the neutral syllable is mapped to that of the target expression to obtain intonation contour. In the case of CART, along with prosody information of target expression, linguistic information obtained from the text is also used to build trees. Listening test indicates that the speech synthesized using GMM (for small data set) and CART (large data set) sounds more expressive compared to linear prosody modification. Cabral *et al.* developed Emo Voice system to incorporate different emotions into the neutral expressive speech in German language [8]. In Emo Voice system the neutral speech is converted to expressive speech by modifying both prosody parameters (pitch, duration and intensity) and excitation source parameters (jitter, shimmer, and glottal wave parameters) by Pitch Synchronous Time Scaling (PSTS) method [30] [3]. The rules for the prosody and voice quality modification are derived based on the acoustic profiles presented in [31–33]. Theune *et al.* devised prosodic rules to generate expressions in the story telling style [21]. Story telling expressions are synthesized by modifying the pitch and intensity of various part of the story like suspense, climax etc..

2.3.2 Expressive speech synthesis by playback approach

In playback approach, the expressive speech is synthesized independently of expressive parameters using the respective expressive speech database. Here expressive speech synthesis is achieved either by merely playing back what is available in the database of the target expression or using the models which are trained using the target expression database. The unit selection based and HMM based expressive speech systems trained on the respective expressive database works on play back approach [2, 34–37].

For improved naturalness in the synthesized speech, the emotional speech synthesis systems developed based on unit selection concatenation were developed. A highly natural synthesized emotional speech is demonstrated by Lida *et al.* by storing large databases for each emotion [35]. For synthesizing the target emotion, the respective emotion database is loaded and selects units from the database to synthesize the speech in the target emotion. A good quality conversational speech is synthesized by Campbell using phrase unit selection based speech synthesizer from a very large database [38]. Hofer *et al.* used a blended database by mixing emotion databases of angry, happy and neutral speech for synthesizing speech in the target emotions [36]. For achieving this, target cost function is designed to give more penalty to select the units other than the intended emotion. The work done by Fernandez and Ramabhadran, also followed a similar approach by mixing the units of other emotions to synthesize the target emotional speech [37]. Pitrelli *et al.* proposed an unified approach

for expressive speech synthesis system by combining corpus driven and prosodic phonology approach [2]. The subjective studies described in the paper indicate that the use of corpus driven approach is effective for conveying good and bad news. Effective contrastive emphasis and Yes-No questions are achieved using prosody phonology approach. Similar to unit selection approach expressive speech synthesis systems are developed using statistical parametric (HMM) approach also. Yamagishi *et al.* trained HMM models for different speaking styles like reading, sad, joyful and rough and synthesized speech in the target styles using the respective trained HMM models [34]. Some of the synthesized expressive speech samples for happy and angry emotions are available for listening at the following link: http://www.iitg.ernet.in/stud/dgovind/emotionsynthesis.htm

2.3.3 Expressive speech synthesis by implicit approach

The implicit control based expressive speech system controls the expressivity by interpolation between two statistical models trained on different expressive databases. The expressive speech synthesized by the interpolation and adaptation of HMM models are examples of implicit control. HMM based speech synthesizers offer various adaptation techniques to adapt the average style model to a specific style. Miyanaga *et al.* proposed an HMM based style synthesis system using a style control vector estimated for each style [39]. During the synthesis the style control vector associated with the target style transforms the mean vectors of the neutral HMM models. The adaptation techniques provide flexibility to build the statistical models with a few minutes of data if an average model is available. As the speech synthesized using speaker adaptation are found to be more robust than speaker dependent case, these adaptation techniques can be used for synthesizing various styles also [40]. Apart from the adaptation techniques, HMM speech synthesis systems provide flexibility to synthesize various speaking styles or emotions by HMM interpolation or multiple regression of emotion vectors [41–43]. In spite of all these advantages for HMM based speech synthesis systems the notable disadvantage is the inherent over-smoothing of the spectral and excitation parameters by the HMM models [41]. This over-smoothing causes the reduced naturalness in the synthesized emotions. However, the perceptual studies presented by Barra-Chicote *et al.* shows that the emotional speech synthesized using HMM based speech synthesis system and unit selection based speech synthesis system provides almost similar emotion identification rates [41].

The present work focuses on the development of neutral to expressive speech conversion systems for ESS by explicit control of prosodic features. Here the issue will be framing of prosodic rules by

the analysis of each expression in the database and incorporating them into the neutral speech.

2.4 Issues in Expressive Speech Synthesis by Explicit Control

The ESS by explicit control is achieved by transforming the neutral speech by a signal processing approach according to the prosodic rules framed for the target expression. The various issues in the ESS by explicit control approach are the following:

- Synthesizing a good quality neutral speech
- Analysis and estimation of expressive parameters
- Incorporation of expressive parameters

The various issues and approaches for the development of neutral speech synthesizers are presented in Section 2.5. Based on this review, the speech synthesized either from a unit selection concatenative system or HMM based statistical parametric speech synthesis system is of good intelligibility and reasonably natural. Therefore any of the two systems can be used as the neutral speech synthesizer for the present work.

The analysis and estimation of expression specific parameters of various emotions are performed on an expressive database. Section 2.6 reviews various existing expressive databases used for expressive speech analysis. Expression specific parameters for each expression are analyzed with respect to the neutral expression. In this stage, the issues will be the accurate estimation of parameters across various expressions. Therefore the choice of signal processing tools that accurately estimate expressive parameters are important for analyzing the expressive parameters. Section 2.6 also reviews studies made on various expressive parameters in expressive speech analysis. Finally, the outcome of this study will be a set of rules on expressive parameters which can modify the parameters of neutral speech to synthesize the expressive speech.

The final stage in the ESS by explicit control is the incorporation of the rules for each expression on the parameters of neutral speech to obtain the speech in the target expression. This is typically achieved by a signal processing method. The issue in incorporating these expressive rules is to introduce minimum perceptual distortion without affecting the naturalness in the synthesized expressions. Section 2.7 reviews various methods for incorporating expressive parameters.

2.5 Review of Text to Speech Synthesis

The front end text to speech synthesis system serves as the NSS. The parameters of the neutral speech synthesized by the TTS system are modified according to target expression to generate the speech in the target expression. Every TTS has a front end text processing block, which converts the text to be synthesized to an abstract linguistic specifications. These abstract linguistic specifications could be a sequence of phonemes or any sub-word unit and also it could be annotated with the prosodic information [4, 18, 20]. The text processing stage generally includes the text normalization, phrasification and lexical analysis modules. The role of the text processing module is to provide a unique contextual description about the sound units across the entire utterance. This abstract linguistic representation drives the waveform generation module to synthesize the speech in accordance with the text given. For the waveform generation from the abstract linguistic representation, there are four approaches namely,

- Articulatory speech synthesis
- Formant speech synthesis
- Concatenative speech synthesis
- Statistical parametric speech synthesis

2.5.1 Articulatory speech synthesis

The objective of articulatory speech synthesis is to model various articulatory processes involved during speech production and use that knowledge to synthesize good quality speech sounds. Various stages in the development of articulatory speech synthesizers are the following:

- Articulatory data acquisition
- Geometric modeling of vocal tract
- Acoustic modeling for the synthesis

In articulatory data acquisition, the positions of various articulators are studied during the production of various sound units. In articulatory data acquisition, snapshots of the speech production organs are taken during the production of various sound units. The articulatory data is acquired using

various sensors like fiberscope [44], x-ray [45–47] and Magnetic Resonance Imaging (MRI) [48]. The popular commercially available devices for measuring articulation are developed using electro magnetic articulography (EMA) and electro palatography (EPG). After acquiring the articulatory data, geometric models are built for the vocal tract which is in turn used for the acoustic synthesis of the sound units. There are a number of 2D and 3D models proposed in the literature for the accurate geometric modeling of vocal tract system using the available articulatory data [49–51]. After generating a complete geometrical model from the articulatory data for various sound units, these parameters in terms of the area functions have to be mapped into acoustic parameters for the speech synthesis which is the final stage of the articulatory speech synthesis. The source filter theory of speech production proposed by Fant is the basis for the speech synthesis from the acoustic parameters [52]. The vocal tract (VT) tube acoustics is obtained by solving Websters horn equations for the sound pressure. The area functions of the geometrical models are mapped to simple 2D circular cross-sectional areas for applying Websters Horn equations. Once the acoustic parameters of VT tube is obtained, the electrical analogue circuit can be designed for the synthesis [53]. The acoustic parameters can also be simulated by estimating the digital filter coefficients [54, 55]. The source information (pitch and intensity) required for the acoustic synthesis can be directly computed for each sound unit from the recorded data. Palo provides a detailed review of articulatory speech synthesis in his MSc Thesis [49].

Even though articulatory synthesis is based on physical theory, construction of geometrical models and their mapping makes it computationally complex. Also improper co-articulation modeling causes degradations in the synthesized speech. Computational complexity and reduced naturalness makes the articulatory speech synthesis approach less popular compared to other existing approaches for speech synthesis.

2.5.2 Formant Speech Synthesis

Formant speech synthesizers are example of the speech synthesizers using a speech production model. The formant speech synthesizer is developed based on source filter theory of speech production [52]. From a historic perspective Dudley's channel vocoder developed in 1939 is a primitive form of formant synthesizer [56]. In [56], the distribution of the formant energies and voicing are adjusted by an expert human to synthesize speech like waveform. Formant speech synthesis involves the simulation of formant frequencies, formant amplitudes and glottal source characteristics for each sound unit. The

vocal tract is simulated using a set of resonators connected in cascade or parallel. The popular technique for the formant synthesis is developed by Klatt in 1980 [19]. The parameters corresponds to formants and voicing source are tuned manually for synthesizing a good quality speech. After the development of Klatt formant synthesizer, Fant and Liljencrants came up with an improved parametric glottal model to provide a better shape for the glottal waves used in the Klatt synthesizer [57]. As the formant synthesizers provide flexibility to vary the voice qualities of the synthesized speech by varying the control parameters of the source and the system, formant synthesizers are used in emotive speech synthesis applications [10]. Recent development in the formant speech synthesis is the data driven formant synthesis [58]. Here formant parameters stored in the units library are selected and set as the control parameters for the formant synthesizer.

Even though formant synthesizers provide flexibility for varying voice qualities in the synthesized speech, increases complexity due to large number of control parameters. This necessity of setting the control parameters for speech synthesizer increases the time required to build speech synthesizer with good intelligibility and improved naturalness. Even though formant synthesized speech is observed intelligible, but sounds unnatural which is its main drawback.

2.5.3 Concatenative Speech Synthesis

The basic idea in concatenative speech synthesis is synthesis by joining the segments of the natural speech waveform that are stored in the database [4,59]. These segments can be words, subword units like phonemes, diphones and syllables. The widely used concatenative speech synthesis works on the principle of *unit selection*. The popular unit selection speech synthesis systems are *clunits* and *multisyn* [4,59–61]. These systems differ to each other in terms of the type of unit, database and unit selection criteria used for synthesizing the speech. The unisyn concatenative speech synthesis system uses diphones as the basic units for concatenation. A diphone is defined from the stable middle region of one phone to the stable middle region of another phone. Unisyn attempts for building a diphone synthesizer by storing fixed size diphone units obtained from natural recordings stored in the repository [60,62]. However, due to the availability of only single example of diphone units representing all the phonetic contexts, causes unnaturalness in the synthesized speech. To overcome this problem, *clunits* speech synthesis system is proposed in [59,61]. In clunits, the basic units of concatenation are mono phones. Here a large phonetically labeled database of 4-5 hours of continuous speech is used as unit inventory. The similarly sounding phonemes of different phonetic contexts in the entire database

are clustered for the same phoneme class. During the synthesis, according to the phonetic context, the appropriate cluster of the same phone class are picked from clustered database and optimum units are selected for the reduced temporal and spectral discontinuities (join cost) at the concatenation points by the efficient Viterbi search algorithm. In multisyn based speech synthesis systems, use diphones as the basic unit of concatenation. Here the diphones required for the concatenation are selected from a large diphone labeled database of 4-5 hours of continuous speech recordings. The two cost functions are used for the optimum unit selection by the viterbi search are join cost and the target cost functions. The join cost computes the penalty cost in terms of the spectral and temporal distortion while joining two units and target cost gives the penalty cost of the unit with respect to the target diphone context. Figure 2.3 shows the example of unit selection for the text "two". For each target unit, the example candidate units are listed from the database. The variable size of the candidate units shown in Figure 2.3 indicate that the size of the listed candidate units from the database need not be fixed size units. Also variable number of candidate units per target unit indicate that, the number of example units available in the database are different for each target unit having specific linguistic context. The path in bold dotted line indicates the optimum path obtained by the Viterbi search algorithm which has total minimum sum of target and join costs. As unit selection speech synthesis approach requires a large repository of labeled database, it demands higher memory requirements. The unit selection systems also suffer from the spectral and temporal mismatches at the concatenation points to some extent. Signal processing techniques, like Time Domain Pitch Synchronous Overlap Add (TD-PSOLA), are used to smooth the discontinuities at the concatenation points [63, 64]. Even though there are advances in the articulatory and formant synthesis approaches, unit selection based speech synthesis approach remain as the mostly used speech synthesis approach.

2.5.4 Statistical Parametric Speech Synthesis

Statistical parametric speech synthesizers follow a model based approach for speech synthesis. In contrast to concatenative systems, instead of storing the units here the models corresponding to each unit will be stored in the repository. In the model based approach, the speech is parameterized and uses statistical methods to build models for those speech parameters, hence the name *statistical parametric speech synthesis* [20]. In statistical parametric speech synthesis, the statistical parametric

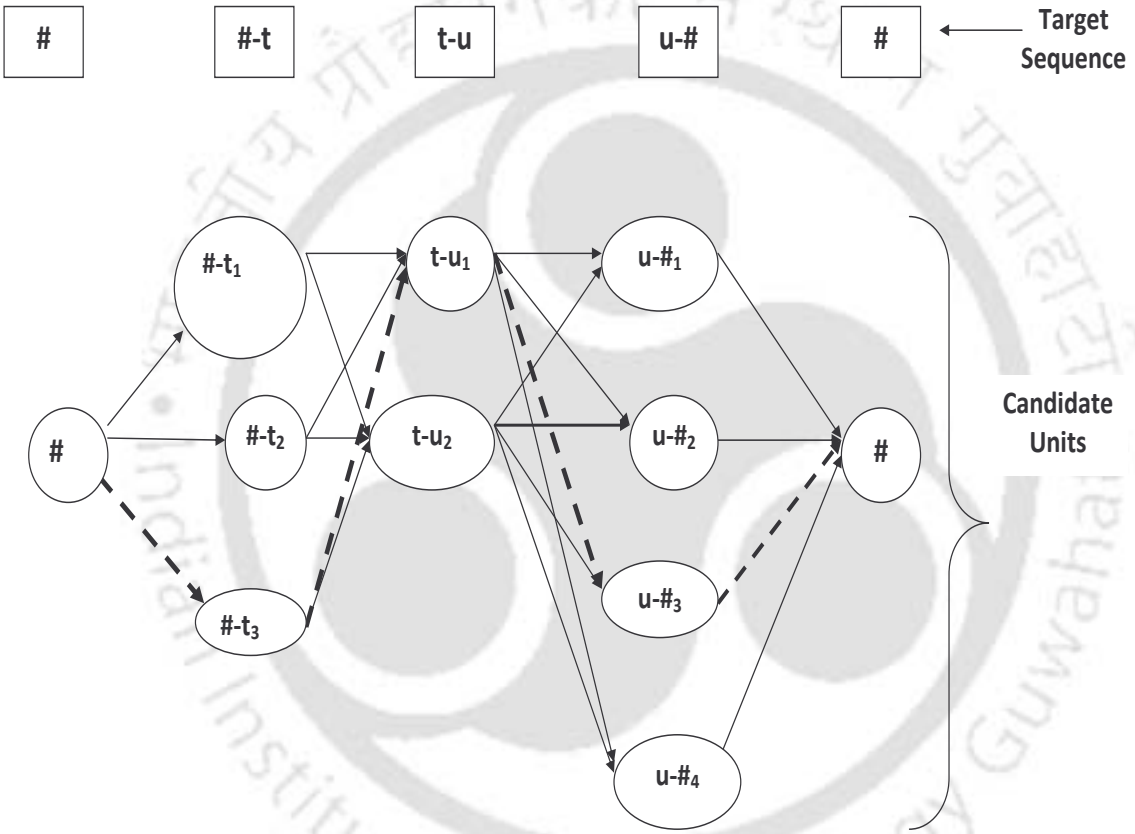

Figure 2.3: Unit selection in concatenative speech synthesis system: The bold-dotted lines indicate the optimum path of the diphone units to be concatenated for the text "two"

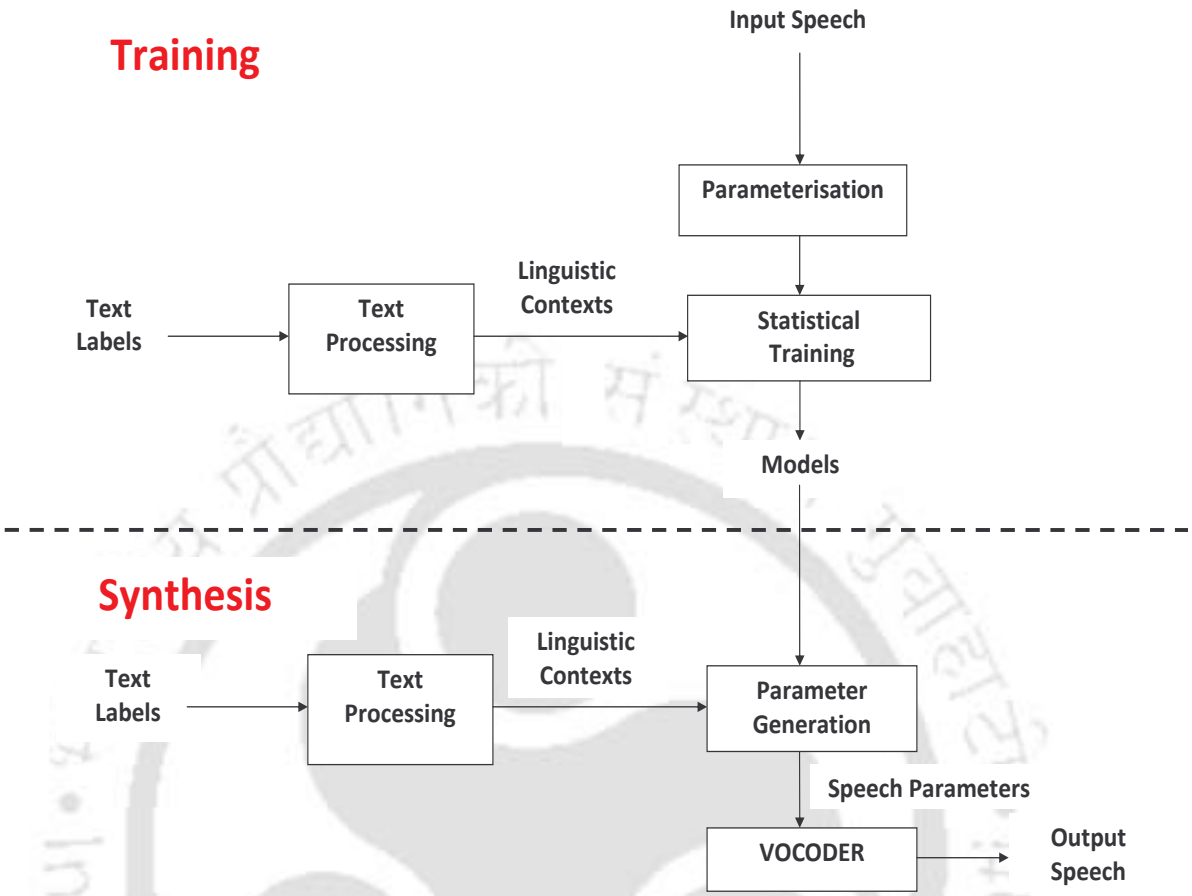

Figure 2.4: Statistical parametric speech synthesis: The block diagram showing training and synthesis phases in building a statistical parametric speech synthesizer [20]

models are built using HMM models[1]. Hence statistical parametric speech synthesis is also known as HMM-based speech synthesis. The schematic block diagram of statistical parametric speech synthesis is given in Figure 2.4. The HMM based speech synthesis is proposed in [65]. The increased popularity of the HMM in speech recognition and availability of efficient learning algorithms (Forward-Backward algorithm, Baum-Welsch re-estimation), computationally efficient search algorithms (Viterbi search) and parameter tying methods by decision tree clustering, are the motivation behind the development of HMM based speech synthesis systems [20, 66].

In HMM based speech synthesis, the speech in the database is parameterized into system and excitation source components. The context dependent HMM models are built by training the HMMs

[1]In contrast with the HMM based speech recognition, HMM based speech synthesis uses Hidden Semi Markov Models (HSMM) for representing the speech parameters for each sound unit [20]. The terminology of HMM models used in this chapter refers to HSMM models that is used for the speech synthesis.

simultaneously with source and system components for all the training data set. The HMM parameters during training are estimated using maximum likelihood criterion.

The speech parameters used for HMM training include, value of F_0 and 5 parameters for spectral envelope of the aperiodic excitation as the excitation parameters and 40 to 60 parameters are used for the spectral envelope (mel cepstral coefficients) [67,68]. For the natural synthesis of speech dynamic features (delta and delta-delta coefficients) of both F_0 and spectrum are also used for modeling. These parameters are extracted typically at 5 ms frame rate. Like in speech recognition HMM models are trained with labeled speech data. Unlike the speech recognition case, here full context labels are used for the training. Use of these full contexts for the HMM modeling increases the complexity as compared to speech recognition case where simple context models like triphone HMM models are used. For the model complexity control, model parameter tying techniques are adopted where model parameters shared among models having similar contexts. Decision tree based clustering techniques are commonly used for parameter tying for HMM based speech synthesis. These parameter tying is responsible for retrieving models corresponding to the unseen contexts (for which there were no examples in the training data) during the synthesis. To synthesize a sentence, the text processing block generates the context dependent phoneme sequence. The corresponding stored HMM models are then retrieved (decision tree based clustering is used to find the model parameters for the phonemes with the unseen context factors) and concatenated to form the sentence HMM. From the sentence HMMs the speech parameters are generated. The number of frames of speech parameters to be generated is determined by the explicit duration model. The speech parameter trajectories are generated based on maximum likelihood parameter generation algorithm (MLPG) using the model parameters for dynamic features [69]. The generated spectral and F_0 parameters are used for vocoding to synthesize the speech. The STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) vocoder is generally used for synthesizing the speech in HMM based speech synthesis systems [70]. Some HMM based systems also use MLSA (MeL Spectral Approximation) algorithm for synthesizing the speech [71].

Statistical parametric speech synthesis offers more flexibility to adjust the speech characteristics as compared to unit selection synthesis systems. Due to these parameter flexibility, speaker adaptive speech synthesis systems are developed using various adaptation techniques (speaking style adaption, speaker adaption) using limited training data. Since statistical parametric speech synthesizers use

vocoders for synthesizing the speech, synthesized speech sounds a little unnatural as compared to the speech synthesized using concatenative speech synthesizers. However, different techniques are being developed for improving the quality of the synthesized speech in statistical parametric synthesizer by incorporating glottal source parameters [72], articulatory parameters [73], etc. .

Despite the development in articulatory and formant speech synthesis in the recent years, currently the unit selection based concatenative and HMM based speech synthesis systems are the mostly used approaches in the area of speech synthesis. Hence to develop a good quality neutral speech synthesizer in the context of expressive speech synthesis, we can use either unit selection based speech synthesis system or HMM based speech synthesis system.

2.6 Analysis and Estimation of Expressive Parameters

2.6.1 Expressive Speech database

As analysis and estimation of expressive parameters are performed on an expressive speech database to frame the explicit rules for the ESS, the development of expressive speech database is a crucial step for the present work. Hence the present section starts with the review of expressive speech databases used for the ESS. Very few works related to ESS used commonly available database for ESS. Most of the works are based on the data collected locally and are publically unavailable. These expressive databases differ by the language, type of expressions considered, type of text materials used, number of speakers and so on [23].

In the literature two types of expressive data are collected. One is the expressive data simulated by actors [2, 28, 32, 74, 75] and the second is the spontaneous expressive data collected from a real life scenario [9, 76–78]. Most of the ESS systems described in Section 2.3 used expressions simulated by actors. Angry, happy, sad, fear and disgust are the commonly used emotions for the analysis in the case of simulated emotions. Williams *et al.* compared the spontaneous fear and sorrow emotions obtained from the radio announcer recording of the Hindenburg airship disaster, with the same sentences simulated by professional artists [9]. This work concluded that, emotion specific parameters estimated from simulated emotion speech data is comparable with that of the real life emotion speech data. Johnston collected multimodal spontaneous data from the subjects by making them participate in a competitive computer game [76]. The various instants of the game are manipulated in order to obtain various emotional responses from the subjects. Speech, electroglottogram (EGG) and

electromyogram (EMG) are collected for tense, neutral, irritated, happy, depressed, bored and anxious expressions. Despite the practical difficulties in inducing the emotions in speaker, the combined analysis of acoustic features (from speech) and physiological features (EGG and EMG) gave a clearer indications of emotional states of the speakers. JST/CREST database collected by Campbell consists of natural telephonic conversation of various social interactions [77]. The databases of spontaneous expressions are used for synthesizing expressive speech by unit selection approach.

If the goal of ESS system is to deploy in cartoon animations, call center applications or any other commercial applications the simulated expressions can be used for analysis. Since actors are well trained to produce emotions effectively, the use of these simulated emotions by them are recommended. The expressive data collected from non-professionals can also be used. If the application of ESS system is to deploy a dialogue system where the machine should interact more naturally with the user, the analysis of spontaneous expressions is better. Theune *et al.* described an interesting application of ESS aimed at children story telling [21]. The database collected for this application is recordings of story narrated by professional artists. Johnson *et al.* discussed the ESS for military applications to simulate the shouted commands, shouted conversation, normal spoken commands and normal spoken conversation for animated characters [79]. The data set used for training is the recordings of simulated shouting and normal commands.

2.6.1.1 Berlin Emotional Speech Database

Burkhardt *et al.* described the development of acted emotion speech corpus in German language [74]. The database is created with 10 professional actors (5 Males and 5 Females) of 10 emotionally unbiased sentences in six target emotions (Angry, Happy, Fear, Boredom, Sad and Disgust). Each file is recorded at 48 kHz sampling rate is downsampled to 16 kHz sampling rate with 16 bits per sample resolution. The listening test conducted on the recorded emotions gave more than 80% emotion recognition rate by the listeners. Each recorded speech file is annotated at the word and syllable levels.

2.6.1.2 LDC Emotional Prosody Speech Transcripts Database

The data collected in this database are grouped into distance/dominance category and emotional state category [80]. In the distance category the speakers have to give data by imagining whether speaker is speaking in a close room environment with a single listener or speaking with one or more

people or speaking to someone who is standing far way of a room. In the emotional category, speakers have to give data in 14 different emotions with each of the emotions are well defined with a context. The data is collected from 8 professional actors (three males and five females). The actors were asked to speak semantically neutral English phrases with dates and numbers in a given category (emotional or distance category). 14 emotional states are hot anger, cold anger, panic, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust and contempt. Each speaker is given a script card in which the emotion category and phrases to be spoken are written and they are allowed to utter the phrase until the speaker is satisfied about the emotional category conveyed. The data is recorded in 22.5 kHz with 16 bits per sample bit resolution.

2.6.2 Studies on the analysis of expressive parameters

The studies on the expressive speech parameters are classified based on following speech parameters,

- Prosodic parameters
- Excitation source parameters
- Vocal tract parameters

2.6.2.1 Studies on prosodic parameters

The typical prosodic features used for expressive analysis are parameters of F_0 contour, duration (sentence duration, syllable duration etc.) and intensity [9,14,28,81]. F_0 is the average rate at which vocal folds vibrate for voiced sounds. The F_0 contour refers to the variation of F_0 with respect to time. The characteristics of this F_0 can be considered as the prosodic parameters. The duration parameters can be the total duration of the utterance or duration of the sound units like phones, syllables or words etc. The intensity parameter of the prosody is measure of loudness in the utterance. Fairbanks *et al.* studied the F_0 characteristics of five expressions (anger, fear, indifference, grief and contempt) simulated by actors [28]. They found that the expressions can be classified on the basis of F_{0range} (absolute difference between F_{0min} value and the F_{0max} value), F_{0avg}, overall F_0 inflections (variations in F_0 values) and F_0 slope of F_0 contour. Based on the analysis of F_0 characteristics, the indifference expression showed lowest F_{0avg} and narrowest F_{0range}. Fear expression exhibited widest F_{0range} and highest F_{0avg}. The angry expression shows highest F_0 inflection.

Studies on duration characteristics by Fairbanks *et al.* on the same simulated expressions showed that the duration features like speech rate (number of words per minute), variation in number of pauses, length of the pauses and ratio of pause duration to total phonation time affects expressions [75]. For instance, anger, indifference and fear expressions showed higher speaking rate and the expressions grief and contempt showed lower speaking rates [75]. Williams explored the effect of prosodic parameters on actor simulated emotional expressions such as neutral, sorrow, angry and fear. The prosodic parameters considered for the study are F_{0med} (median of F_0 values), F_{0range} and speech rate. According to this study, the angry expressions found to have increased F_{0med} and F_{0range} and sorrow expressions showed reduced F_{0med} and reduced F_{0range} [9]. The duration of the utterance spoken in fear expressions found to be longer than that of the anger expression.

The intensity parameter computed as the average spectral energy, found to be higher for anger for some syllables as compared to neutral speech [9]. The proposed prosodic parameters for the simulated expression case is compared with various emotion contexts of the broadcast recordings during the Hindenburgh airship crash disaster. The study concluded that the characteristics of F_0 contour alone gives the indication of the emotional state of the speaker [9]. Vroomen *et al.* showed that emotions can be accurately expressed by manipulating F_0 contour and sentence duration in a rule based manner. The seven expressions (neutral, joy, anger, boredom, indignation, sadness and fear) of two sentences in Dutch recorded by an actor are considered as expressive data for their work. Based on perceptual and acoustic analysis, the F_0 contour of each expression is modeled using Dutch intonation rules [15]. During the synthesis, the F_0 contour for each expression is generated and used to manipulate the F_0 contour of the neutral speech. The duration is modified by the linear compression according to the optimum modification factors obtained from the analysis stage. Murray *et al.* considered prosodic parameters like F_{0avg} , F_{0range} , F_0 changes (F_0 inflections, F_0 inclination, F_0 declination), intensity and speech rate [24] for the analysis of various expressions. Six expressions (anger, sadness, happiness, fear and disgust) along with the neutral expression are considered for the analysis. According to the analysis, the fear expressions and disgust expressions showed the highest and lowest speech rates, respectively. The F_{0avg} was highest for anger and lowest for disgust emotion. F_{0range} was narrower for the sad expressions. Anger and happiness got the highest intensity whereas disgust showed low intensity. Among the F_0 changes, anger showed abrupt F_0 changes in stressed vowels and happiness showed smooth and upward F_0 inflections. Sadness and disgust showed

downward deflections whereas fear expression showed normal F_0 changes. The rule based expressive speech synthesis system developed by Murray *et al.* in [29], modifies the prosodic parameters at the phoneme level according to the prosodic analysis described in [24]. The rules are set for each prosodic parameter in the HAMLET neutral formant speech synthesizer for synthesizing the speech in the target expression [29]. Hashizawa *et al.* considered F_{0max}, speech rate and F_0 of the pitch accented syllables for the analysis. The analysis showed that, the F_{0max} is the highest for anger emotion, F_0 and pitch accents are enhanced for joy and F_{0max} is minimum for sad emotion [81]. Tao *et al.* used F_{0avg}, $F_{0topline}$, $F_{0baseline}$, syllable duration and intensity as the prosodic parameters [7]. The $F_{0topline}$ is the mean of the line connecting the peaks in F_0 contour and $F_{0baseline}$ is the mean of the line connecting the valleys in F_0 contour. According to Tao *et al.*, the F_{0avg} and $F_{0topline}$ provide more classification abilities for five emotions (neutral, anger, happiness, fear and sadness). Murtaza *et al.* showed the significance of F_{0range} than F_{0avg} in classifying four emotions (neutral, anger, happy and sad) of two sentences from two speakers.

2.6.2.2 Studies on excitation parameters

Excitation parameters refer to the parameters representing the characteristics of the excitation source. The excitation source parameters are analyzed at subsegmental and segmental levels. The parameters computed within 2 to 3 pitch periods (10-20 ms) of speech are termed as the segmental parameters. The parameters such as jitter and shimmer are examples of excitation parameters estimated at the segmental level. The parameters estimated within a pitch period of speech are subsegmental parameters. The glottal flow parameters like open quotient (OQ), Return quotient (RQ) and speech quotient (SQ) are examples of excitation parameters at the subsegmental level.

Jitter is the average change of F_0 from one cycle to another, where as, shimmer is the change in the excitation strength from one cycle to another. Whiteside has shown the significance of jitter and shimmer in discriminating various emotions [32]. Seven expressions (neutral, cold anger, hot anger, happiness, sadness, interest and elation) of 5 sentences from two speakers are used for this study. The prosodic parameters like mean intensity, standard deviation of intensity and F_{0avg} are also used along with mean of shimmer and jitter for the analysis of seven expressions. According to the analysis done in this work, hot anger possessed highest mean jitter and mean shimmer and sadness showed minimum mean jitter and mean shimmer. Using these five parameters, the expression discrimination accuracy was found to be 88.9% and 85.7% for the two speakers, respectively [32]. Jhonston *et*

al. performed the expressive analysis on both spontaneously recorded expressions and simulated expressions. The first part of the paper used multimodal (speech, EGG, EMG) data of spontaneous emotions collected from subjects by exposing them to different instants of a manipulated computer game [76]. The participants were asked to pronounce the sentences to be recorded and asked them to choose one expression from the list of expressions (irritated, disappointed, surprised, relieved, helpless and alarmed). Based on the choice of expressions made by the participants at different situations of the game, the recorded expressions are categorized into low coping, high coping, obstructive and constructive responses. The parameters used for the analysis are glottal slope, F_{0range}, heart period, respiratory period and respiratory depth. According to the analysis, the glottal slope obtained from the EGG and the heart rate tend to be higher for obstructive situations and the low coping situations are characterized by the longer respiratory cycle. The second part of the study consisted of glottal analysis on EGG data of seven expressions (tense, neutral, happy, irritated, depressed, boredom and anxious) of 5 digit strings, short phrases and sustained vowel /a/. The expressive data of these seven expressions were collected from eight speakers. The speakers were asked to imagine the emotions for recording the expressive data. The excitation parameters used for the analysis are mean jitter, closing quotient (glottal closing time of the glottis as a percentage of pitch period, T_0). According to the analysis, mean jitter was highest for happy and anxious expressions and lowest for boredom and depressed expressions. The depressed and boredom expression showed the higher values of closing quotient and anxious expression showed lowest values for closing quotient. Cabral *et al.* used jitter, shimmer and glottal flow parameters like OQ, RQ and SQ as the excitation parameters for synthesizing emotion [17]. In this work, the excitation parameters are extracted from seven emotions (angry, happy, fear, boredom, neutral, sad and disgust) of German emotional speech database. According to the excitation parameter analysis presented in this work, happy and fear expressions tend to show decrease in OQ . The breathy quality of the anger expressions are confirmed by the decrease in SQ and RQ compared to other expressions. Along with prosodic parameters Tao *et al.* used jitter to analyze five emotional expressions like neutral, anger, happiness, fear and sadness for the task of neutral to expressive speech conversion [7]. According to this study, the happiness expression tend to have highest jitter and sadness showed lowest jitter.

2.6.2.3 Studies on Vocal tract parameters

Formant frequencies (F_1, F_2, F_3, F_4 and F_5) and bandwidth associated with each formant form important characteristics of the vocal tract system. Mean F_1, mean F_2 and F_1 bandwidth are the vocal tract cues reviewed by Scherer [1]. The acoustic characteristics of around 14 expressions are reviewed in [1]. Compared to other emotions, lower mean F_1 was observed for happy and elation expressions and higher mean F_1 was observed for other expressions. Whereas lower mean F_2 was observed for all the expressions other than happy and elation. The expressions, hot anger, cold anger, disgust and fear, tend to show narrower F_1 bandwidth [1]. The parameters considered for these articulatory stimuli are F_1 mean, F_2 mean and corresponding formant bandwidths. Ishii *et al.* used a subset of spontaneous expressions collected in JST (Japan Science & Technology) CREST (Core Research for Evolutional Science and Technology) ESP (Expressive Speech Processing) project. The spontaneous expressive data is collected by recording subject's daily spoken conversations using mini recording devices and wearing head mounted close speaking microphones. After recording the data, speakers were asked to label the expressions based on their mood at various times in the conversations such as neutral, worried, content, happy, bright, sad, angry, tension, energy ("Energy" is categorized based on the global intensity of the speech) etc. By analyzing the average F_3 and average F_4 parameters of words in bright, energy and tension expressions, Ishii *et al.* found that the average F_4 is higher for bright expressions than the expressions labeled with energy. There was no correlation observed for average F_3 values indicating inconsistency of F_3 parameter for the same expressions. Erickson *et al.* studied the effect of formant frequencies on spontaneous sad emotions [82]. The spontaneous sad emotional data is collected in two sessions through the telephonic conversions with the subject. The spontaneous sad emotions were evoked by asking about the sad demise of the subject's mother. Lowering of F_2, F_3 and F_4 was observed for the sad emotions when compared with the non emotional data.

Table 2.1 presents the summary of the review of the studies made on expressive parameters. The columns given in the table represent the contributors, choice and type of expressive data used in their work, expressive parameters considered and the important findings of their work.

Table 2.1: Summary of various studies about expressive parameters

Author	Expressions	Type of expressive data used	Expressive parameters explored	Findings
C. E. Williams and K. Stevens (1972) [9]	Anger, sorrow, fear and neutral	Simulated by actors	F_0 mean, F_0 range, speech rate and Energy	1. F_0 contour as the indicator of different emotional states 2. F_0 parameters of simulated and real emotions are similar
Scherer (1986) [1]	Happy, cold anger, hot anger, anxiety, disgust and sad	Simulated by actors	F_0 parameters, F_1 mean, F_2 Mean and Formant Band width	F_0 parameters along with VT parameters represents the acoustic properties of emotion
I. R. Murray and J. L. Arnott (1993) [24]	Angry, happy, sad, fear and disgust	Simulated by actors	Speech rate, F_0 mean, F_0 range and Intensity	Emphasizes the role of prosodic parameters in synthesizing emotions
Whiteside (1998) [32]	Cold anger, hot anger, happy, sad, interest and elation	5 Short sentences simulated by two speakers	Mean of overall jitter, Mean of overall shimmer	Significance of jitter and shimmer in discriminating the emotions
T Johnston and K. R Scherer (1999) [76]	Tense, neutral, irritated, happy, depressed, bored, anxious	data (EGG, EMG and speech) collected during computer game events	Jitter, Glottal closing Time	EGG signal gives emotion dependent characteristics
Ishii and Campbell (2003) [78]	Neutral, worried, bored, polite, depressed, angry	Natural telephonic conversation recorded	F_0 parameters, F_3 mean , F_4 mean	F_4 influences different voice qualities
Y. Hashizawa *et al.* (2004) [81]	Angry, happy and sad	Isolated Words by professional announcers	Speech rate, F_0 max and Pitch Accent	1. F_0 max is higher for anger 2. For happy both accents and F_0 will be enhanced 3. F_0 and accents were suppressed for Sad.
J P Cabral [8] *et al.* (2005)	Angry, happy, sad, fear, surprise, boredom, disgust	Simulated by actors	Jitter, shimmer, glottal wave parameters (OQ, SQ, RQ)	Better recognition rates obtained for happy, angry and fear
M. Bulut and S. Narayanan (2008) [14]	Angry, happy, sad and neutral	Simulated by professional and non-professional actors	F_0 mean, F_0 range, F_0 stylization characteristics	Changes in F_0 range significantly changes perceived emotions

2.6.3 Estimation of Expressive Parameters

This section reviews various studies made on estimation of prosodic, excitation and VT parameters from speech.

2.6.3.1 Estimation of prosodic parameters

As most of the works related to expressive speech synthesis use prosodic parameters as the expression dependent parameters, it is essential to accurately estimate these prosodic parameters for expressive speech analysis. The features of F_0 contour, speech rate and intensity are the prosodic parameters reviewed in this section.

F_0 or pitch is the fundamental frequency of vibration of the vocal folds during the production of voiced sounds. Since vocal folds vibrate only during the production of voiced sounds, F_0 is defined only for voiced sounds. F_0 is undefined for the unvoiced sounds such as fricatives. Representation of F_0 values versus the time instants at which they are calculated is termed as F_0 contour or pitch contour. In order to derive F_0 contour, F_0 values have to be accurately estimated from speech. F_0 estimation techniques described in the literature are broadly classified into block processing based approach and event based approach [83] [13]. Block processing approach computes the average F_0 from block of speech segment where as event based approach accurately determines the instantaneous F_0 by processing entire speech utterance. Most of the earlier works employ block processing approach for estimating F_0 for expressive speech analysis [1, 9, 14, 24, 32, 76, 78, 81].Auto correlation [83, 84], cepstral analysis [84], simplified inverse filtering (SIFT) [85] and average magnitude difference function (AMDF) [86] are the popular methods for estimating F_0 by block processing. A robust method, exploiting the properties of Hilbert envelope (HE) of LP residual for reliably estimating average F_0 in adverse conditions is proposed by Prasanna *et al.* [87].

In order to accurately estimate all the instantaneous F_0 values for the entire speech utterance, the event based approach is used. The instantaneous pitch period is defined as the interval between glottal closing instant of one cycle to the next. As the discontinuities related to pitch occur at the instants of glottal closure where the maximum excitation of the vocal tract occurs, accurate determination of these instants of significant excitation or epochs are essential for computing the instantaneous F_0. The epochs or instants of significant excitation can be defined as instants of glottal closure incase of voiced speech or onset of burst or frication incase of unvoiced speech [12, 13]. The interaction of

vocal tract in the speech produced makes the estimation of epochs location a challenging task. There are several methods proposed in the literature to estimate the epochs location accurately. The epoch estimation using group delay (GD) functions [88], dynamic programming based projected phase-slope algorithm (DYPSA) [89], HE based method and Zero Frequency Filtering (ZFF) [12] of speech based methods are popular existing methods for the epoch estimation. The instantaneous pitch period is computed as the interval between successive epochs location [11]. The instantaneous pitch period is also termed as the epoch interval [11]. The instantaneous F_0 is computed by scaling the reciprocal of epoch interval with F_s. The representation of instantaneous F_0 values at the corresponding epochs location gives the instantaneous F_0 contour of the utterance. The significant F_0 parameters derived from the F_0 contour are F_{0avg}, F_{0max}, F_{0min}, F_{0range}. These F_0 parameters can be computed using the following equations,

$$F_{0avg} = \frac{1}{N} \cdot \sum_{i=1}^{N} F_{0i} \tag{2.1}$$

$$F_{0max} = \max \{F_{0i}, i = 1, 2, ..., N\} \tag{2.2}$$

$$F_{0min} = \min \{F_{0i}, i = 1, 2, ..., N\} \tag{2.3}$$

$$F_{0range} = F_{0max} - F_{0min} \tag{2.4}$$

Where N is the total number of pitch cycles in the speech utterance. Various duration parameters that are used for expressive analysis are the speech rates at sentence, syllable and phoneme levels and number of pauses. Unlike the F_0 parameters, the estimation of duration parameters are mostly measured directly from the database. Fairbanks *et al.* computed the speech rate by counting the number of words uttered per second for analyzing emotional expressions [75]. Burkhardt *et al.* used syllable duration as the prosodic parameter for the analysis of the expressions given in German emotional speech database [74]. The syllable boundaries are labeled manually by listening, analyzing spectrograms and simultaneous EGG recordings. Cabral *et al.* used sentence duration of each utterance as the duration parameter for analyzing various expressions. Murray *et al.* analyzed duration at the phoneme level [24]. The duration of phonemes are estimated directly from the phone boundary labeling of the utterances. The automatic phone boundary marking can also be done using HMM based force alignment techniques [4]. Other prosodic parameter used for emotion analysis is the intensity. The intensity is measured by computing the energy of the utterance.

2.6.3.2 Estimation of excitation parameters

According to the review of excitation parameters given in Section 2.6.2.2, various excitation parameters used for expressive speech analysis are the segmental parameters like jitter and shimmer, and subsegmental parameters related to glottal flow and strength of excitation. This section reviews the method employed to estimate these parameters.

Shimmer is a measure of strength of excitation [90] of the glottal wave which is defined as the change of strength of excitation pulses from one cycle to another. One of the method to characterize the glottal activity is from the LP residual obtained by the LP analysis of speech [91]. During the glottal activity, the LP residual has high energy region and during non glottal activity region, LP residual shows noisy characteristics. Analysis of the excitation source based on LP residual depends on the accuracy of LP analysis. Murty *et al.* described a method to compute the glottal activity and strength of excitation in speech based on the ZFF of speech [12]. Since the rate of vibration of the vocal folds is proportional to the glottal air flow, the excitation strength can be found by measuring the sharpness with which glottal closure occurs. This can be computed by measuring the slope of the ZFF signal around the epochs location. Now shimmer can be measured as the change of excitation pulse between successive epochs location. Farrus *et al.* computed shimmer as the variation of the peak to peak amplitude values in consecutive pitch period and then proposed shimmer measurements in the various levels for speaker verification [92]. The average shimmer measurement described in [92] is given by the Equation 2.5.

$$S_{avg} = \frac{1}{N-1}\sum_{i=1}^{N}|A_i - A_{i+1}| \tag{2.5}$$

where A_i is peak to peak amplitude in the i^{th} pitch period and N is the total number of pitch periods.

Jitter is estimated by measuring the average change in pitch period from one pitch cycle to another. Jitter is derived from the instantaneous F_0 contour. Method to estimate the jitter in different levels for speaker verification task is described in [92].

$$J_{avg} = \frac{1}{N-1}\sum_{i=1}^{N}|T_{0i} - T_{0i+1}| \tag{2.6}$$

where T_{0i} is i^{th} pitch period and N is the total number of pitch periods

These shimmer and jitter measurements can be used to find the variation of target expressions with respect to the neutral speech.

There are several methods discussed in the literature to estimate glottal waveform parameters directly from speech. Fant *et al.* developed LF (Liljencrants and Fant) model to uniquely represent glottal flow derivative for a given pitch period [57]. The LF model is a four parameter model developed based on the glottal closure instants and glottal closure discontinuity points. The four parameters are frequency, amplitude, growth constant of sinusoid and recovery time constant. Cabral *et al.* estimated the expression dependent glottal flow parameters from the LP residual. These parameters are estimated by integrating the LP residual. These glottal flow parameters are measured by estimating the following time instants first:

- **Glottal closure instants**, n_o: By estimating the instants of glottal closure from the LP residual
- **Closed phase instants**, n_{cl}: It is the instant at which closed phase of the glottis starts. n_{cl} is calculated by finding the instant of the first peak after the zero crossing.
- **Glottal Opening instants**, n_{op}: The time instant of the opening phase is calculated by setting a positive (th_{pos}) and negative threshold (th_{neg}) on the short time signal. where th_{pos} is 75% of the maximum value of the signal energy and $th_{neg} = -th_{pos}$. The first point of the positive growing part that intercept with the negative threshold and the last point of growing part of the opening phase that intercepts the positive threshold are calculated [3]. The average signal amplitude value, k_{mean} between the two estimated points is calculated. The last positive crossing point of the signal over the k_{mean} axis is estimated to be n_{op}. The calculation of n_{op} is shown in Figure 2.5
- **Maximum of glottal flow**,n_p: The zero crossing of DC value between n_{op} and end of the short time signal.

The estimated time instants are used to compute the duration of the glottal cycle phases, Return phase (N_a), Peak flow duration (N_e), Closed phase (N_c), Opening phase (N_{op}) and Closing phase, (N_{cl}).

$$N_a = n_{cl} \tag{2.7}$$

$$N_e = N - n_{op} \tag{2.8}$$

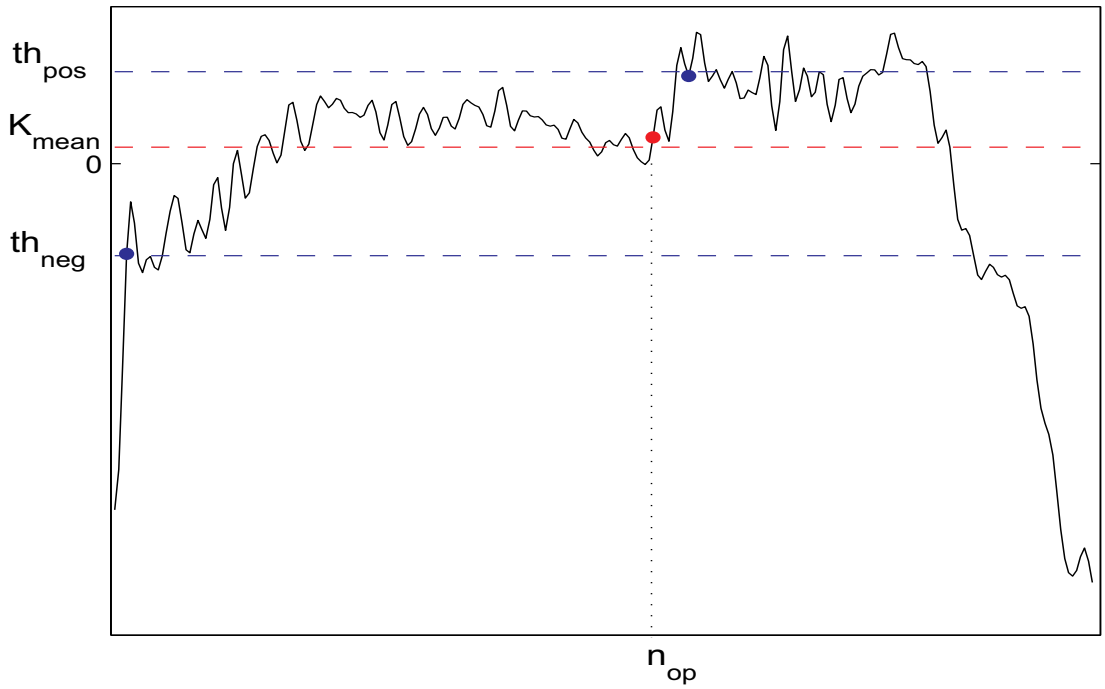

Figure 2.5: Locating the instant of glottal opening in a short time segment of LP residual (Figure used with permission of J. P. Cabral).

$$N_c = N - N_a - N_e \tag{2.9}$$

$$N_{op} = n_p - n_{op} \tag{2.10}$$

$$N_{cl} = N - n_p \tag{2.11}$$

where N is the total duration of the glottal cycle. The duration of these phases in a glottal cycle is pictorially represented in Figure 2.6. The glottal flow parameters like OQ, RQ and SQ are calculated as given by the following equations,

$$OQ = \frac{N_a + N_e}{N} \tag{2.12}$$

$$RQ = \frac{N_a}{N} \tag{2.13}$$

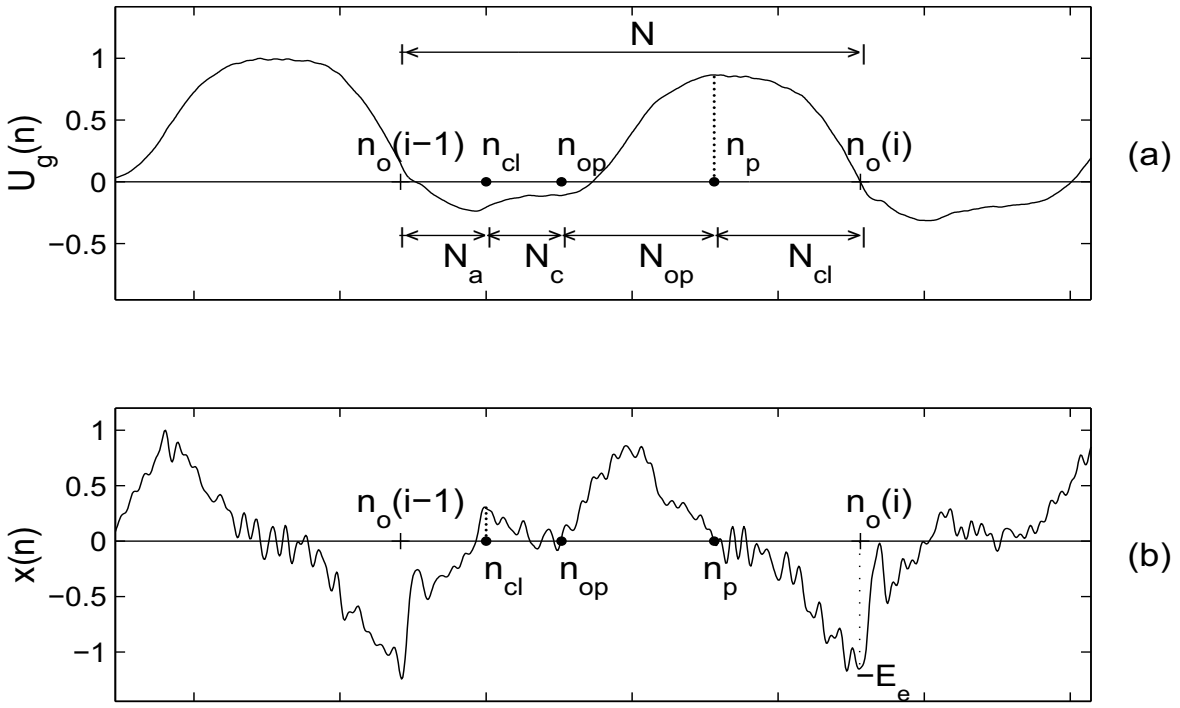

Figure 2.6: Representation of glottal phases in a (a) glottal cycle and (b) in its derivative for a short time segment of LP residual (Figure used with permission of J. P. Cabral)

$$SQ = \frac{N_{op}}{N_{cl}} \tag{2.14}$$

The other important excitation source parameter is the strength of excitation. The strength of excitation is a subsegmental feature which is the strength with which the vocal folds are vibrating during the production of voiced speech [90]. One of the methods to compute the excitation strength is from the LP residual obtained by the LP analysis of speech [93]. In the glottal activity region, the LP residual has high energy and during non-glottal activity region, LP residual shows low energy noisy characteristics [93]. The strength of excitation is computed by computing the energy of the residual samples in the region around the glottal closure instants. Murty *et al.* described a method to compute the glottal activity and strength of excitation in speech using the ZFF based epoch extraction [90]. Since rate of vibration of the vocal folds is proportional to the glottal air flow, the excitation strength can be found by measuring the sharpness with which glottal closure occurs [90]. This can be computed by measuring the slope of the ZFFS around the epochs location.

2.6.3.3 Estimation of vocal tract parameters

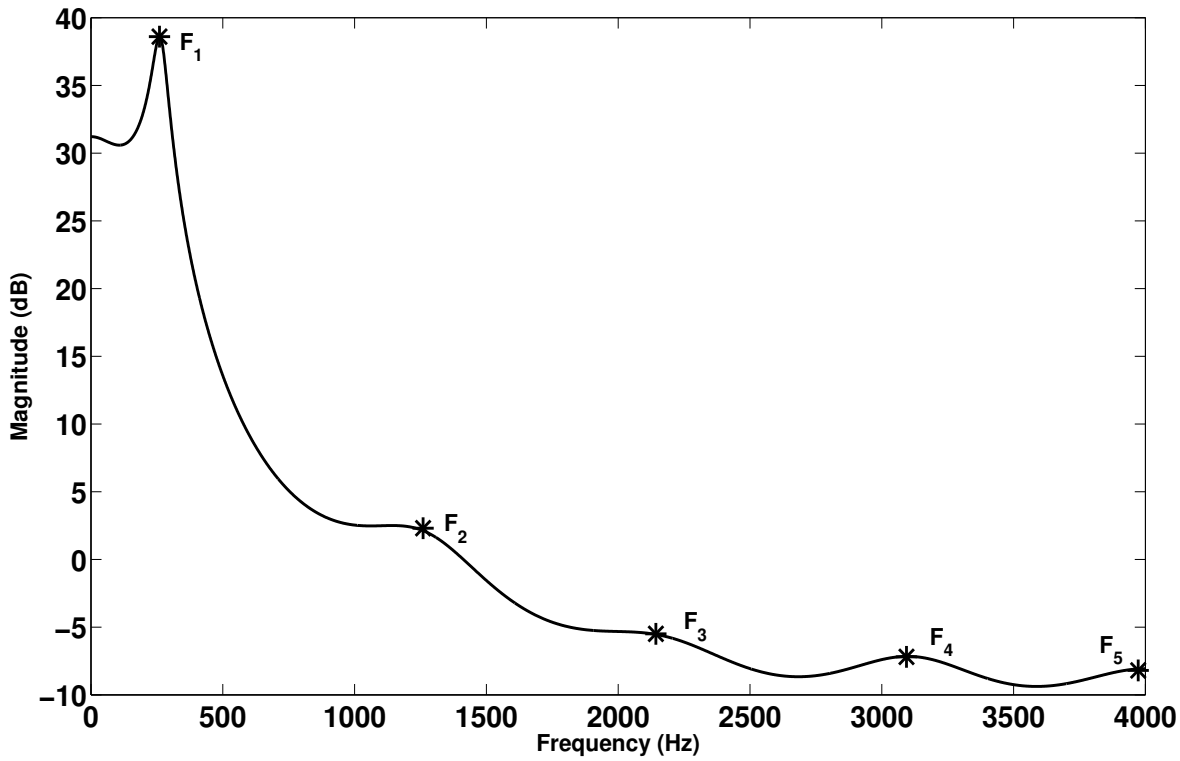

Figure 2.7: The formant estimation from LP spectrum: This figure shows the linear prediction spectrum and Formant locations (indicated by '*') obtained from the peaks in the LP spectrum

There are many methods discussed in the literature to estimate the formants from speech. In [94] used a method to extract formants by picking peaks from smoothed log-spectra obtained by cepstral analysis. Figure 2.7 indicates how the formants are located by picking the peaks in the LP spectrum. The spurious peaks in the log magnitude spectrum causes wrong identification of the formants and is the main disadvantage of the approach. Formant extraction by the linear prediction analysis solved the issue of spurious peaks appearing in the LP spectrum [95]. But here also peak picking wrongly estimates the formants in the case of merged peaks in the LP spectrum. Yegnanarayana showed that the differentiated linear prediction phase spectrum can clearly resolve the merged peaks because of the additive nature of the phase of cascaded digital resonators [96]. The Formants can be extracted by picking peaks of the differentiated LP phase spectrum . To further resolve the spectral peaks in the smoothed log spectra, properties of the group delay function of minimum phase signals are utilized in [97]. The significance of measuring the free resonances of the vocal tract to analyze various regions

like consonant vowel transitions is described in [98]. The speech produced during the closed phase of the vocal folds are mainly due to vocal tract. At this time the vocal tract tube is closed at one end and hence the resonances produced are free from the vocal fold vibrations and glottal air flow that occur during the opening phase of the glottis. The formant parameters are extracted from the analysis segments taken around the glottal closure instants. The pole zero model of these analysis frames are determined. This can be represented mathematically as follows [98]:

$$r(n) = \sum_{k=1}^{p} A_k \rho_k^n e^{i\theta_k n} = \sum_{l=1}^{p/2} \rho_l^n (A_l e^{i\theta_l n} + \overline{A}_l e^{-i\theta_l n}) \tag{2.15}$$

where n is the discrete time index and $p/2$ is the number of formants. The variable θ_k is the k^{th} formant frequency such that -π <$\theta_k \leq$+ π. As $r(n)$ is real, it can be represented as complex conjugate pair. The factor A_k is the formant amplitude and ρ_k is the formant damping factor, where 0<$\rho_k \leq 1$. It can be seen that the formant frequency F_k and bandwidth B_k can be computed from the Equations (2.16) and (2.17) as given in [98].

$$F_k = \frac{F_s}{2\pi}\theta_k \tag{2.16}$$

$$B_k = -\frac{F_s}{\pi}\ln(\rho_k) \tag{2.17}$$

2.7 Incorporation of Expressive Parameters

The expressive parameters have to be incorporated into the neutral speech according to the scale factors set in the expressive analysis stage for effective synthesis of the speech in the target expression. The incorporation of expressive parameters are performed at the prosodic, excitation and vocal tract levels. This section reviews methods used to incorporate the expressive parameters at each level.

2.7.1 Methods to incorporate prosodic parameters

The expression specific F_0, duration and intensity parameters can be incorporated by prosody modification algorithms. Manipulation of F_0, duration and intensity of the given speech without affecting the perceptual quality is termed as prosody modification [11,99]. There are several methods discussed in the literature for prosody modification [11] [100] [99]. The approaches like Over Lap Add (OLA), Synchronous Over Lap Add (SOLA) and Pitch Synchronous Over Lap Add (PSOLA) operate directly on the speech waveform to modify the prosodic parameters [99]. The OLA and SOLA

methods are mainly used for time scale modification of the given speech signal [100]. The duration modification here is achieved by overlap adding the analysis frames having time scaled length chosen from the cross correlation with the actual frames of the given speech signal. Development of the PSOLA allowed both time scale and pitch scale modification by using pitch marks as the anchor points [11] [99]. In PSOLA method, in general, the pitch modification is achieved by placing the analysis windows around the modified pitch marks and adding the overlap regions. In the timescale modification, first resampling the actual pitch mark locations according to the desired timescale and then the analysis frames around the actual pitch mark is copied and overlap added to the new pitch locations closest to the original location. The resulting signal obtained will be duration modified according to the desired modification factor.

Depending on the domain in which PSOLA is applied there are Time Domain PSOLA (TD-PSOLA), Frequency Domain (FD-PSOLA) and Linear Prediction PSOLA (LP-PSOLA) [99]. In FD-PSOLA prior to overlap add, the spacing between the pitch and harmonics of excitation signal obtained by the source-filter decomposition are modified according to the desired pitch modification factors by resampling in the frequency domain [99]. Unlike TD-PSOLA, LP-PSOLA operates on the LP residual of the speech signal to be prosody modified. As the LP residual samples are less correlated than speech samples, the overlap-adding of residual analysis frames give less distortion. Kawahara developed a method to manipulate the speech parameters like F_0, speech rate and vocal tract length using speech representation and transformation using adaptive interpolation of weighted spectrum (STRAIGHT) [101]. Here a pitch adaptive speech analysis is carried out for speech parameter manipulations. The instantaneous F_0 estimation method developed as part the work uses Gabor filters for finer frequency resolutions. The manipulated speech is obtained by reconstructing the smooth time frequency representation using bilinear transformations. Smoothing is done to remove the pitch periodicity effects in the time-frequency surface representation of the original speech. Muralishankar *et al.* proposed F_0 modification method using discrete time cosine transformation (DCT) of pitch synchronous residual frames [102]. Here the DCT coefficients are estimated from the pitch synchronous residual frames obtained after the LP analysis. According to the pitch modification factors the DCT coefficients are either truncated (increase in F_0) or padded with zeros (decrease in F_0). For instance, if N_1 is the number of DCT coefficients in residual frame, N_2 point IDCT is taken where N_2 is N_1 divided by the F_0 modification factor. N_2-N_1 trailing end DCT coefficients are removed for increasing

F_0 and N_1-N_2 zeros are padded to decrease F_0. Each frame of speech is then synthesized by the LP filtering with the LP coefficients computed.

A method of prosody modification by accurately determining the epochs location is proposed in [11] [88] [103]. This type of prosody modification is generally known as epoch based prosody modification. The steps involved in the epoch based prosody modification are the following:

- Finding the accurate epochs location
- Modifying the epochs location according to the desired prosodic parameters
- Reconstruct the prosody modified speech

2.7.1.1 Estimating epochs location

As described earlier epochs in speech can be defined as the instants of glottal closure in voiced speech and onset of burst or frication in unvoiced case [12,88]. Due to the effect of vocal tract, the accurate estimation of epochs location from speech is a challenging task. There are many methods proposed for the estimation of the epochs location from speech [12,88,89]. Smit *et al.* proposed group delay (GD) based approach to estimate the epochs location from LP residual of speech [88]. In GD method, GD function is computed from the LP residual by considering in blocks of about 1-2 pitch periods length with shift of every sample. If $e(n)$ is the LP residual, the fourier transform of $e(n)$ and its time weighted function are given by the Equations (2.18) and (2.19).

$$E(\omega) = FT[e(n)] = E_R + jE_I \tag{2.18}$$

$$F(\omega) = FT[ne(n)] = F_R + jF_I \tag{2.19}$$

The computation of GD function $\tau(\omega)$ is given in Equation 2.20.

$$\tau(\omega) = -\phi'(\omega) = \frac{E_R F_R + E_I F_I}{E_R^2 + E_I^2} \tag{2.20}$$

After removing isolated peaks from $\tau(\omega)$ using a 5 point median filter, average value of the GD function is computed for each block of residual with a shift of one sample. The average GD function obtained for every sample shift is known as the phase slope function. The epochs are estimated as the zero crossings of the phase slope function. Later, the robustness of GD based method against various degradation are studied by Satyanarayana *et al.* in [93]. As the GD function is computed from LP residual frames for every sample shift, the GD based epochs estimation is a computationally

complex task [88]. To reduce the time complexity in estimating the epochs from LP residual, a two stage processing approach is proposed by Rao *et al.* [104]. In the first stage, the approximate epochs location are estimated from the HE of LP residual and in the next phase, the GD function calculated around the approximate epochs location obtained from the HE of LP residual. Dynamic programming based projected phase slope algorithm (DYPSA) is proposed by Naylor *et al.* found to provide better epochs estimation accuracy than GD method [89].

Recently, a simple, fast and accurate method for estimating epochs from speech is proposed by Murty *et al.* in [12]. In ZFF method, the speech is passed through the cascade of two zero frequency resonators (ZFR). The ZFR output $y(n)$ is given by the Equation (2.21)

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{2.21}$$

where $a_1 = 4, a_2 = -6, a_3 = 4, a_4 = -1$ and $x(n)$ is difference speech obtained by the successive difference of samples in the speech signal which is given by $x(n) = s(n) - s(n-1)$ The variations in the ZFR output due to epochs are obtained by subtracting local mean from the ZFR output. This local mean subtracted ZFR output is termed as the zero frequency filtered signal (ZFFS). The local mean substraction from ZFR output can be expressed as,

$$\hat{y}(n) = y(n) - \frac{1}{2N+1} \sum_{n=-N}^{N} y(n) \tag{2.22}$$

Here $2N+1$ corresponds to the size of window used for computing the local mean, which is typically the average pitch period computed over a long segment of speech. Thus, the epochs location will be the positive zero crossings of the ZFFS. The accuracy of the epochs estimated using ZFF method is better compared to DYPSA and GD method [12]. In ZFF method, as the epochs are estimated directly from speech without computing LP residual, the method is found to be computationally fast as compared to DYPSA and GD methods [12].

2.7.1.2 Modifying epochs location for prosody modification

Rao *et al.* described a method to modify the epochs location according to the desired prosody modification factors [11]. In this method the modified epochs location are obtained by deriving the epoch intervals. The epoch intervals are derived as the difference between successive epochs location. The epoch interval plot is then generated by interpolating the epoch intervals of successive epochs

location. In the case of F_0 modification, this epoch interval plot obtained for the entire utterance is scaled according to desired pitch modification factor. The epoch interval plot is resampled according to the desired duration modification factor in case of duration modification. Modified epochs location are obtained from the resampled and/or scaled interpolated epoch interval plot. For instance, if A is the starting sample index of the modified epoch interval plot, the modified epoch location B is computed by adding, the modified epoch interval number of samples at A^{th} location in the modified epoch interval plot, to A^{th} time index. Similarly, modified epoch location sample index C is obtained by adding modified epoch interval number of samples at B^{th} location in the modified epoch interval to the location sample index B.

2.7.1.3 Reconstructing the prosody modified speech

The prosody modified speech is synthesized by generating the modified residual waveform. For generating the modified LP residual, the modified epochs location that are nearest to the original epochs location are found. Perceptually relevant (20% epoch interval region starting from the epoch) residual samples starting from original epochs location are copied to the new modified epochs location. The perceptually relevant residual samples in the epoch interval refers to human listening in TTS and speech enhancement tasks. The remaining residual samples in the original epoch interval other than the samples in the perceptually relevant region, are resampled to fill up the modified epoch interval. Resampling is used mainly to avoid the spectral discontinuities introduced due to truncation of epoch intervals (in case of raising F_0) and replication of samples (in case of lowering F_0). This way the prosody modified LP residual is reconstructed. In pitch modification, as the duration of the utterance remains same, the LPCs computed from the original speech signal are excited by the modified LP residual to synthesize the pitch modified speech. In duration modification, since overall duration of the utterances are changing, LPCs of original speech are updated for very frame shift according to duration modification factors. These new LPCs are excited by the modified LP residual to synthesize the duration modified speech. The perceptual quality of the synthesized files for various pitch and duration modification factors are evaluated using perceptual tests. 25 research scholars of the lab participated in the perception experiments. Here, the subjects were asked to evaluate the prosody modified speech files based on the distortion present in the speech file. The filenames of each method are coded to avoid the biasing towards a particular method. The subjects were asked to judge their opinion scores on a five point scale where score 1 corresponds to poor quality with objectionable

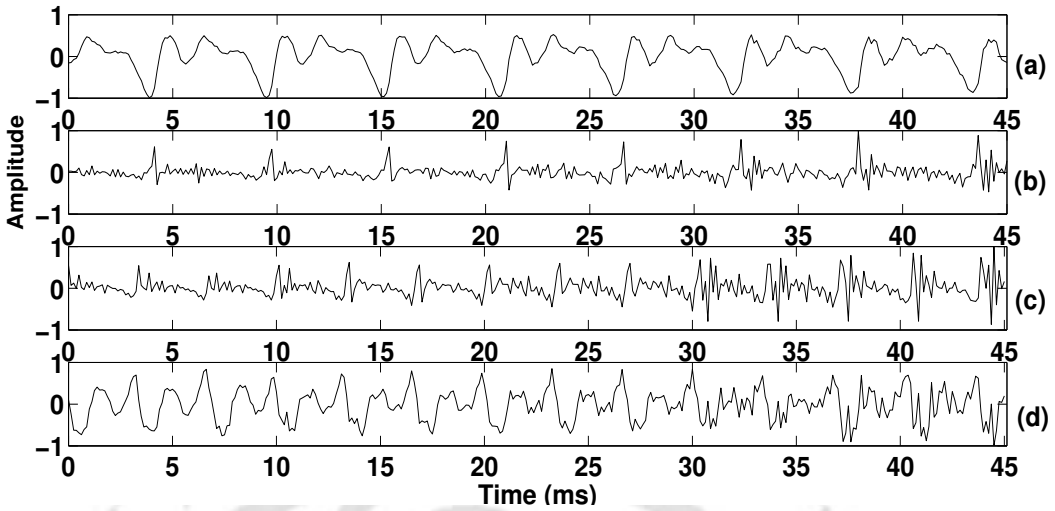

Figure 2.8: Pitch Modification: (a) Long Segment of a voiced speech, (b) its LP residual, (c) modified LP residual by increasing the pitch by 1.5 times and (d) reconstructed pitch modified speech.

distortion and score 5 is excellent quality with no perceptual distortion. For moderate pitch and duration modification factors, the epoch based residual modification and LP-PSOLA based approaches provide almost equal mean opinion scores (MOS). As compared to LP-PSOLA approach, epoch based prosody modification approach had higher MOS scores for the case of extreme modification factors (modification factors greater than 2 and less than 0.5).

Figure 2.8 plots the pitch modification example by increasing the pitch of the original speech segment by factor of 1.5. As we can observe that there are around 13 pitch cycles in Figure 2.8(c) which is 1.5 times than the number of pitch cycles (8 pitch cycles) in the LP residual segment of the original signal as shown in Figure 2.8(b).

Figure 2.9 plots the duration modification example by increasing the duration of the original speech segment by factor of 2. As we can observe that the overall duration of the signal is doubled compared to the original duration of the signal. Also it has to be observed that the pitch intervals in the duration modified speech remain unaltered as in original speech. Figure 2.9(c) has duration that is 2 times that of the LP residual segment of the original signal as shown in Figure 2.9(b).

2.7.2 Methods to incorporate excitation parameters

The rules related to the excitation parameters like jitter, shimmer and glottal flow parameters have to be incorporated to effectively convey the expressive information in the synthesized speech. Cabral described one method to incorporate the jitter into the neutral speech by adding a random

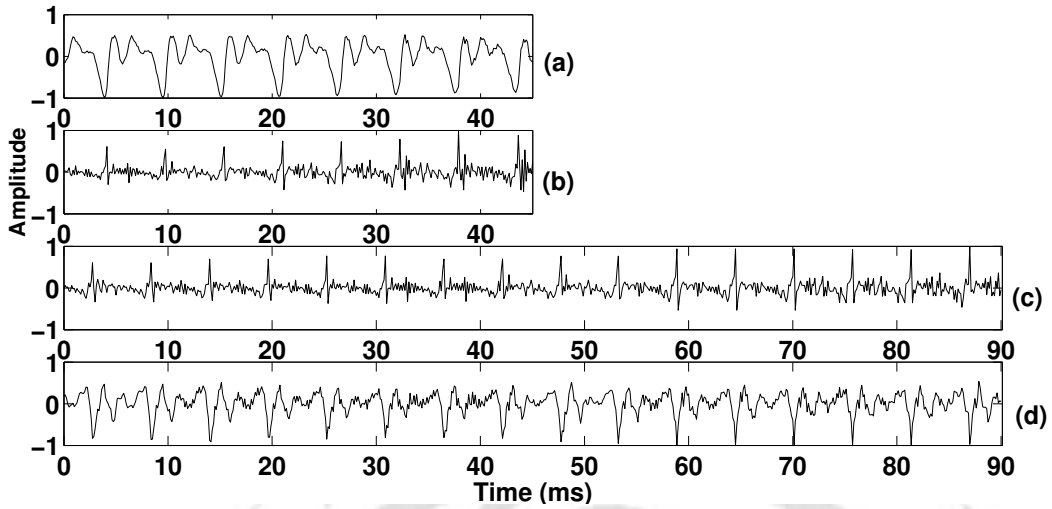

Figure 2.9: Duration Modification: (a) Longer Segment of a voiced speech, (b) its LP residual, (c) modified LP residual by increasing the duration by 2 times and (d) reconstructed duration modified speech.

value to the pitch period [17]. Here the time index of the synthesis pitch marks are randomly varied (according to jitter modification factor) for incorporating the voice quality that is related to jitter. Also the shimmer is incorporated by scaling the energy envelope of the short time signal by a random number (whose variation is according to shimmer modification factor). The glottal flow parameters like OQ, SQ and RQ can be modified by scaling the time indices used to estimate them [17].

Ruinskiy et al. brought another method to incorporate shimmer and jitter for simulating hoarse voice quality in a given speech. Here the jitter is introduced by relative stretching and shortening of pitch cycles [105]. The jitter modification factor for the modification is retrieved from a jitter bank which stored the trends in the jitter values for every 2 to 4 pitch cycles. The residual samples in the given pitch cycle is resampled to modify jitter. Similarly, the shimmer modification factors are also stored across consecutive pitch cycles. The shimmer is introduced by multiplying each pitch cycle by a window function with varying peak amplitudes according to the shimmer factor [105].

2.7.3 Methods to incorporate vocal tract parameters

Even though there are little works discussed in the literature towards incorporating VT parameters for ESS, there are some works in the literature towards voice conversion [106] [107]. Rao *et al.* achieved vocal tract modification by the linear transformation of formant locations and bandwidth in the z-plane [106]. The formant frequencies are modified by shifting the polar angle that the corresponding conjugate pair of poles make with respect to positive real axis of the z-plane. Formant frequency

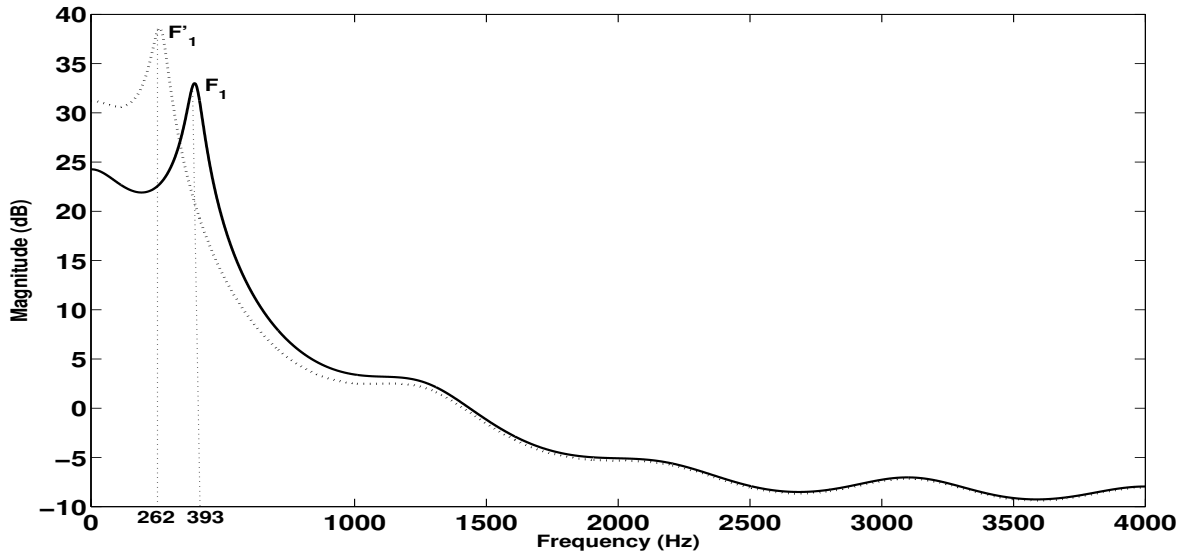

Figure 2.10: The formant frequency modification: This plot demonstrate the shifting of the first formant (F_1) by a factor of 1.5 times the actual formants locations (dotted plot)

shifting of the LP spectrum is demonstrated in Figure 2.10. In the Figure 2.10, the first formant of the modified LP spectrum, F_1', located at 393 Hz (shown by the thicker plot) is shifted 1.5 times the the formant value of the original LP spectrum, F_1 which is at 262 Hz (shown by the dotted plot).

The formant bandwidth is modified by scaling the magnitude of the conjugate pair of poles for the corresponding formant. The bandwidth scaling corresponding to F_2 in the LP spectrum is demonstrated in Figure 2.11. In Figure 2.11, the bandwidth of F_2, B_2' scaled 0.25 times (thick plot) the original F_2 bandwidth, B_2 (dotted plot) .

In a recent work on voice conversion, Rao *et al.* achieved the VT modification by deriving mapping functions using feed-forward neural networks (FFNN) [107]. Here, the line spectral frequencies (LSF) derived from LPCs are used to represent vocal tract characteristics. The mapping functions that represent the relation between the VT characteristics of source and target speakers are derived by feeding the time aligned LSFs of both source and target speakers to the FFNN using a database containing 500 Hindi utterances [107]. The dynamic time warping (DTW) is used to time align the LSFs of source and target speakers. Joseph *et al.* used multi-layer FFNN to map the vocal tract parameters of throat microphone to close speaking microphone to enhance throat microphone speech [108]. Here, the objective was to bring a nonlinear relationship between source and system features of throat microphone data and speech from close speaking microphone for the enhancement of throat

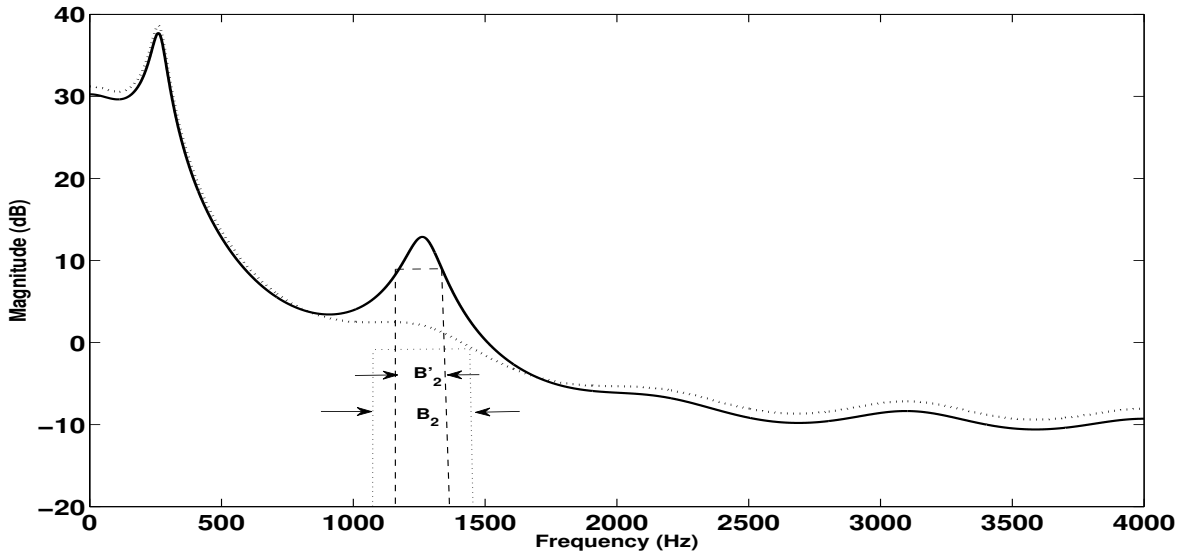

Figure 2.11: The formant frequency modification: This plot demonstrate the bandwidth scaling corresponds to the second formant (F_2) by a factor of 0.25 times the actual formant bandwidth (dotted plot)

microphone data. The LP cepstral coefficients (LPCC) derived from LPCs are used to characterize the vocal tract parameters.

2.8 Summary of the Works Related to Neutral to Expressive Speech Conversion for ESS

As the unit selection based neutral speech synthesis provides more naturalness compared to statistical parametric speech synthesis systems, unit selection based speech synthesis systems can be used as NSS system for the ESS by explicit control. From the studies on expressive parameters, most of the works use prosodic parameters as common expressive parameters. Also, many of the works use the expression dependent excitation parameters and vocal tract parameters as the supplementary features with prosodic parameters. Hence it is necessary to accurately estimate and analyze these prosodic parameters for various expressions. It has been observed that almost all the studies use conventional methods for estimating prosodic parameters for expressive speech analysis. As expressions are characterized by the presence of prosodic variations that are much more than that of the neutral expressions, the accuracy of the estimated parameters using conventional methods have to be verified for different expressions. The recently developed ZFF method which provides best accurate estimates of various prosodic parameters as compared to other existing methods. This can be used for accu-

rate speech analysis of various expressions. Most of the expressive speech systems use conventional PSOLA based methods to incorporate the expression specific prosodic variations for ESS. Since the epoch based prosody modification provides improved perceptual quality for moderate prosody modification factors than PSOLA based methods, epoch based prosody modification can be applied for ESS. The GD method of epochs estimation increases the computational complexity of the existing epoch based prosody modification. Hence more accurate and computationally faster ZFF epochs can be used instead of epochs estimated using GD method for epoch based prosody modification. Based on the review of incorporation of prosody parameters for ESS, the ESS achieved in most of the studies are by modifying the prosody parameters for fixed scale factors. However, fixed scaling of the prosodic parameters will not capture the dynamics of the prosody due to various expressions. Hence the prosody modification methods that incorporate the time varying dynamics of the prosodic parameters have to be used for ESS.

To evaluate the epoch estimation performance across various expressions, five expressions (Neutral, Angry, Happy, Boredom and Fear) of German emotional speech database having simultaneous EGG recordings can be used. The epochs estimated from EGG recordings of these expressions, can be used as the reference epochs location for measuring epoch estimation performance. If there are degradations in the epochs estimation performance for expressive speech, a refined method has to be proposed to improve the epochs estimation performance in expressive speech. The expressive parameters like F_{0Avg}, sentence duration and strength of excitation have to be estimated and analyzed across various expressions. A dynamic prosody modification method has to be devised to incorporate dynamic variations of these parameters in the neutral speech to synthesize the speech in the target expression. The synthesized expressions have to be subjectively evaluated for the degree of expressiveness present in the speech.

2.9 Organization of the present work

Chapter 3 presents the refined ZFF method for epochs estimation from expressive speech and, analysis and estimation of expressive parameters. Chapter 3 starts by using conventional ZFF method for estimating epochs from various expressions. In order to reduce the degradation in the epoch estimation performance due to rapid pitch variations in expressive speech, a refined ZFF method is proposed to accurately estimate the epochs location from various expressions. The gross level

variations of F_{0Avg}, sentence duration and strength of excitation parameters are analyzed across various expressions. The gross level scaling factors for neutral to target expressive speech conversion are derived by comparing the parameters of the target expressions with that of neutral expression.

Chapter 4 describes the epoch based prosody modification algorithm for the incorporation of prosodic parameters. Chapter 4 starts by demonstrating the improved perceptual quality when ZFF epochs are used in the existing epochs based prosody modification using GD based epochs. The significant improvement in the computational efficiency of the epoch based prosody modification is demonstrated by performing prosody modification directly on the speech waveform and using epochs estimated using ZFF. In order to incorporate dynamic variations of the prosodic parameters, in the second part of chapter 4 a dynamic prosody modification method using ZFFS is proposed. As the zero crossings of the ZFFS provides accurate epochs location, these zero crossings from the ZFFS are used as the reference locations for deriving the modified epochs location according to desired dynamic prosody modification factors. The perceptual quality of the prosody modified speech is improved by confining the prosody modification only in the glottal activity regions derived from the ZFFS. The general procedure devised to derive the modified epochs location, is used in existing epoch based prosody modification method and TD PSOLA for dynamic prosody modification. The perceptual quality of the dynamically prosody modified speech in each case is compared with the proposed dynamic prosody modification method using ZFFS.

Chapter 5 demonstrates the effectiveness of the epoch based dynamic prosody modification in neutral to expressive speech conversion for text dependent and speaker dependent, text dependent and speaker independent and text independent and speaker independent scenarios. The effectiveness of the of the neutral to expressive speech conversion using dynamic prosody modification and static prosody modification are compared by subjective evaluations.

Chapter 6 summarizes the works done towards the neutral to expressive speech conversion using epoch based dynamic prosody modification. The major contributions and the scope of future work also listed in Chapter 6.

3

Analysis and Estimation of Expressive Parameters

Contents

3.1 Objective

Based on the review, the works done for expressive speech analysis show that the F_0 contour, duration and intensity are important expressive speech parameters [9, 14, 15, 28, 81]. The objective of this chapter is to analyze the effect of various expressions on prosodic parameters like F_0 contour, duration and intensity. For accurate F_0 analysis, instantaneous F_0 contour is computed from each emotion. The strength of excitation, accurately computed for each epoch interval is used as a measure of intensity. Effect of overall duration characteristics are obtained by measuring overall duration of each emotional utterance. In order to accurately compute the instantaneous F_0 and strength of excitation for emotional speech analysis, epochs location need to be reliably estimated from various emotions. As the conventional zero frequency filtering (ZFF) method provides accurate epochs location from neutral speech, ZFF method is used for estimating epochs from various emotions. Due to the rapid variations in the prosodic parameters in emotional speech, spurious zero crossings are introduced in the zero frequency filtered signal (ZFFS) which in turn result in the unreliable epochs estimation. In the present work, a modified ZFF method is proposed by smoothing the conventional ZFFS to remove the spurious zero crossings. The expressive parameters like instantaneous F_0 contour and strength of excitation are then derived from the modified ZFFS for the expressive speech analysis. The estimated parameters from the emotional speech are compared with the ground truth electroglottogram (EGG) recordings. Five emotions of German emotional speech (neutral, angry, boredom, fear and happy) are initially considered for the study presented in this chapter. Also four emotions (neutral, angry, boredom and happy) of recorded Hindi emotion speech database with EGG recordings are also considered for the expressive speech analysis.

3.2 Introduction

Expression or emotion is the hallmark of human speech. Both emotion and naturalness provide completeness and expressiveness to speech. The emotion part of speech plays an important role in conveying the real intent and meaning of the speaker, not readily available as part of the message. In human-human communication, such intent and meaning are extensively used as guiding factors for deciding the future course of action. For instance, in customer care division, the angry speech by a customer mostly indicate the dissatisfaction and hence immediate need in the change of action. Like this the use of emotional information in human-human communication is extensive and effortless.

However, the human-machine communication falls behind in this aspect significantly. There are several attempts to incorporate the same [9,14,15,28,29,81,109]. According to Vroomen *et al.*, F_0 contour and sentence duration (prosody) parameters are sufficient to effectively synthesize the emotional speech from neutral speech [15]. Cabral *et al.* achieved neutral to emotion conversion based on the analysis of the prosodic parameters of various emotions made in [31, 32]. Also the significance of prosodic parameters in emotional analysis are described in [9,14,15,28,81]. Hence it is necessary to accurately analyze the effect of prosody over various emotions. The availability of EGG signals under different emotions [74], make the estimation of prosodic parameters more accurate. However, EGG signals for various emotions may not be available in practice for accurately estimating expressive parameters. Hence we have to rely on the best available tools for accurately estimating the expressive parameters from various emotions. Since the ZFF method is proved to be the most accurate method for estimating the prosodic parameters for neutral speech signals, it can be used for the accurate estimation of these parameters from various emotions.

The near-periodic nature of vocal folds vibration and the strength with which the closure of vocal folds occur is better estimated computationally using the derivative of the glottal wave [12]. The associated periodicity is characterized grossly in terms of average and standard deviation values of F_0 [87] and finely in terms of instantaneous F_0 [13]. The strength can be characterized in terms of the amplitude of the associated epoch at the closure [90]. Apart from this, the duration of the utterances also varies from one emotion to other. All these parameters may be distinct for each expression and hence useful for analyzing expressive information of speech. These parameters can be reliably and accurately estimated using the recently developed methods for finding epochs, instantaneous F_0 and strength of excitation, all based on the concept of zero frequency filtering (ZFF) [12]. In expressive speech signals, due to rapid variations in pitch as compared to the neutral speech, the epochs estimation performance of the conventional ZFF method has to be verified with respect to that of the neutral speech. In case of degradation in the epochs estimation performance in emotional speech, method has to be developed to improve the epoch estimation which in turn improves the accuracy of prosodic parameters when derived using the knowledge of the epochs location. As described later the present chapter addresses the issue of degradation in epochs estimation performance and methods to improve the same for emotional speech. The estimation and analysis of expressive parameters using the modified ZFF method is described for various emotions. The novelty of the work includes the

following:

- Instantaneous F_0, strength of excitation and duration are found to be expression dependent parameters
- The degradation in the performance is observed for the parameters estimated from emotional speech using conventional methods
- Modified ZFF method is proposed for the accurate analysis and estimation of expressive parameters

The rest of the chapter is organized as follows: Section 3.3 provides the analysis of expressions on the glottal wave and its derivative. Section 3.4 describes the modified ZFF method for the estimation of expressive parameters from various emotions. The various parameters estimated from the derivative of glottal wave are studied in Section 3.5. A comparative study with the parameters estimated from the speech waveform is made in Section 3.6. Finally Section 3.7 summarizes the chapter.

3.3 Analysis of Expressions on the Glottal wave and its derivative

The significance of prosodic parameters in generating expressions can be studied by analyzing the EGG signals of various emotions. EGG represents the actual glottal waveform that is produced by closing and opening of the vocal folds during speech production. The near impulse-like excitation produced due to the closing of vocal folds forms an important characteristic of excitation source. The location of these excitation instances is represented by impulse-like discontinuity in the derivative of the glottal waveform. Figure 3.1 shows the speech waveform, LP spectrum, glottal waveform and glottal wave derivative of five different expressions, namely, neutral, angry, happy, boredom and fear, for the same speaker and sound. The nature of the speech waveform is different across different emotions. The variations in the vocal tract characteristics is represented by the LP spectrum and that of excitation characteristics by the glottal wave and its derivative. Since the speaker and sound are same, the variations across different plots are primarily due to emotions. The first level visual inspection of Figure 3.1 indicates that the emotions significantly affect the glottal wave and its derivative as well as the vocal tract spectrum. From the derivative of glottal wave it can be seen that the amplitude of the excitation pulses (termed as strength of excitation), their periodicity (pitch contour) are also affected by the expressions. The glottal waveforms and their derivative indicate that the boredom and

neutral expressions have relatively higher excitation strength and fear emotion has the least strength of excitation compared to all other emotions. Also glottal wave derivative shows that the angry and fear have the highest F_0 and lowest F_0 is by the boredom and neutral.

An interesting observation is the strength of excitation of angry emotion. By perception of speech we feel its strength of excitation should be largest, but comes out to be smallest. This is true because, for producing speech in angry emotion, the pitch increases and hence pitch periodicity decreases. To cater to the smaller values of periodicity, the vocal folds may not be closing with high suction and hence low strength for impulse-like excitation. This interpretation may be further appreciated by analyzing the boredom case. In case of boredom, by perception of speech we feel its strength of excitation should be smallest, but comes out to be largest. This is due to the large pitch periodicity associated with boredom and hence more time for closing and opening. This leads to closing of vocal folds with high suction and hence high strength.

Expressive analysis of relatively smaller segments of the same syllable of various emotions showed that instantaneous F_0 and strength of excitation parameters significantly affect various emotions. In order to study the duration characteristics and nature of instantaneous F_0 contour and strength of excitation contour, longer speech segments of various emotions have to considered. The instantaneous F_0 contour and strength of excitation can be computed by estimating epochs location from emotional speech. Next section describes the estimation of epochs location from various emotions using ZFF method.

3.4 Zero Frequency Filtering Method for Epoch Estimation From Emotional Speech

3.4.1 Conventional ZFF Method for Epoch Estimation

The algorithmic steps to estimate the epochs in speech by ZFF, which hereafter will be termed as *conventional ZFF*, are as follows [12]:

- Difference input speech signal $z(n)$

$$x(n) = z(n) - z(n-1) \tag{3.1}$$

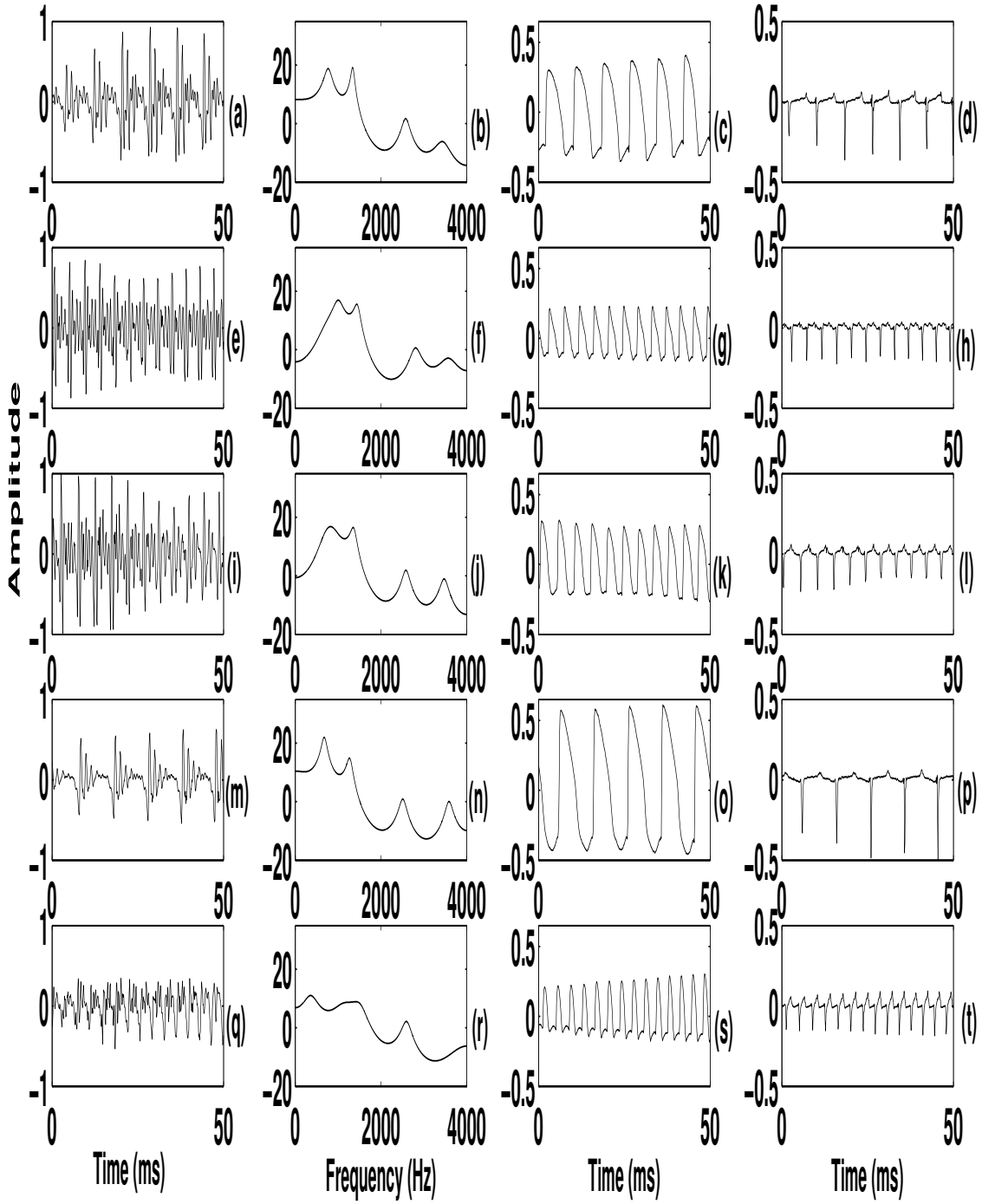

Figure 3.1: *Speech waveforms, LP spectrum, glottal waveform and glottal wave derivative of the expressions for Neutral ((a)-(d)), Angry ((e)-(h)), Happy ((i)-(l)), Boredom ((m)-(p)) and Fear((q)-(t)), respectively.*

- Compute the output of cascade of two ideal digital resonators at 0 Hz

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{3.2}$$

where $a_1 = 4$, a_2 = -6, $a_3 = 4$, a_4 = -1

- Remove the trend i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n) \tag{3.3}$$

where $\bar{y}(n) = \frac{1}{(2N+1)} \sum_{n=-N}^{N} y(n)$ and $2N+1$ corresponds to the average pitch period computed over a longer segment of speech

- The trend removed signal $\hat{y}(n)$ is termed as ZFFS.

- The positive zero crossings of the filtered signal will give the location of the epochs.

These epochs are periodically located in case of voiced speech representing the glottal closure instants and are randomly located in case of unvoiced speech representing the onset of burst or frication [12]. Figure 3.2 shows the epochs estimated from a voiced segment and an unvoiced segment of speech. It can be noted that ZFFS gives periodic zero crossings in case of voiced segment and random zero crossings in the unvoiced segment.

The epoch estimation performance is evaluated for five different expressions (Neutral, Angry, Happy, Boredom and Fear) of German emotional speech corpus having simultaneous EGG recordings [74]. Approximately 100 speech files of 10 speakers and 10 texts per emotion were used for the performance evaluation. For evaluating the estimated epochs from the speech, following measures are used [89].

- Larynx cycle: The range of sample (1/2) $(l_{r-1} + l_r) < n < (1/2)(l_{r+1} + l_r)$ where l_r, l_{r-1} and l_{r+1} are the current, preceding and succeeding reference epoch locations respectively

- Identification Rate (IDR): The percentage of larynx cycles for which exactly one epoch is detected.

- Miss Rate (MR): The percentage of larynx cycles for which no epoch is detected.

- False Alarm Rate (FAR): The percentage of larynx cycles for which more than one epoch is detected.

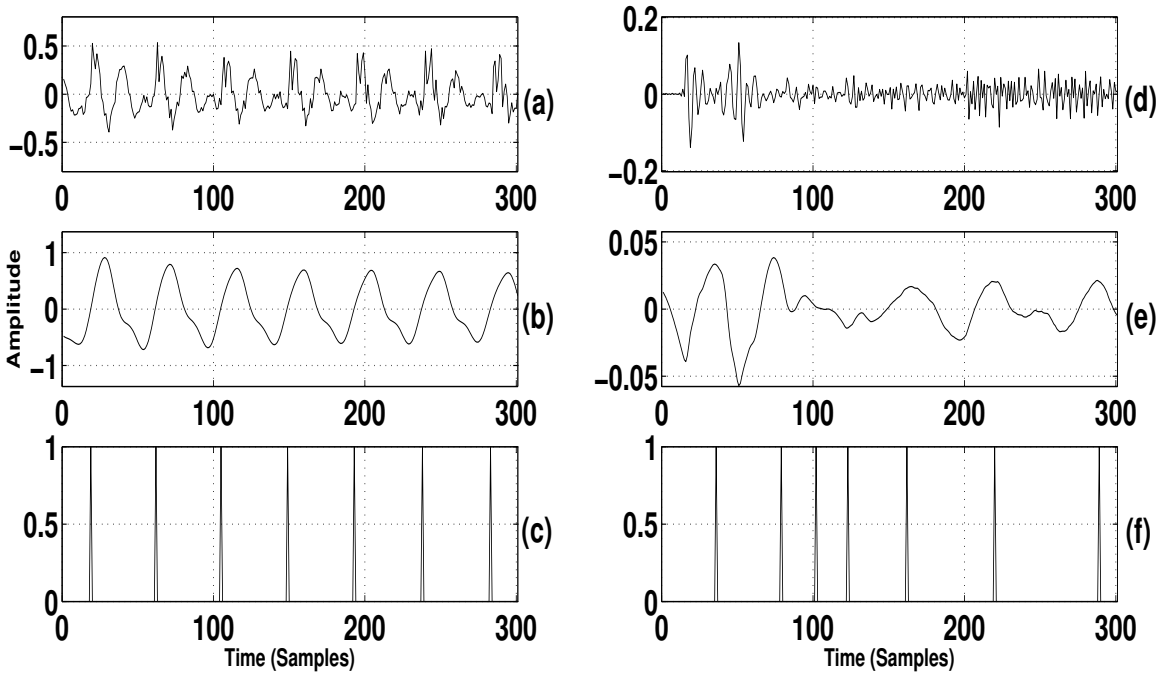

Figure 3.2: Epochs from voiced and unvoiced segments of speech. (a) A voiced speech segment, its (b) ZFFS and (c) epochs. (d) An unvoiced speech segment, its (e) ZFFS and (f) epochs.

- Identification Error (ζ): The timing error between the reference and detected instants of significant excitation in larynx cycles for which exactly one epoch was detected.
- Identification Accuracy (σ) (IDA): The standard deviation of the identification error ζ. Small values of σ indicate high accuracy of identification.

The performance of conventional ZFF are tabulated in the Table 3.1. As expected the conventional ZFF method gives better epoch estimation for neutral emotion speech and the epoch estimation performance degrades in the case of other emotions, except boredom. The reason for this is due to the rapid pitch variations in the emotional speech compared to the neutral speech. In the conventional ZFF method of epoch estimation, due to fixed window size used for local mean subtraction, some of the epochs are either missed or spuriously hypothesized. For this reason an appropriate window size should be selected for the optimum epoch estimation from speech signals with larger pitch variations. As the pitch variation of boredom emotion is similar to that of the neutral, the epoch detection performance of both the emotions is nearly same. This epoch estimation performance from various emotions are compared with the DYPSA algorithm, another popular method for estimating epochs [89]. The

Table 3.1: Epoch estimation performance of conventional ZFF and DYPSA algorithms for different emotional speech signals taken from German database.

Emotion	IDR (%)	MR (%)	FAR (%)	IDA (ms)
ZFF				
Neutral	99.12	00.08	00.79	0.3194
Angry	87.93	00.41	11.66	0.4115
Happy	90.66	00.33	09.02	0.3858
Boredom	98.75	00.04	01.20	0.3495
Fear	94.90	00.13	04.97	0.2774
DYPSA				
Neutral	96.25	0.84	2.92	0.3727
Angry	88.43	5.11	6.46	0.3824
Happy	87.87	4.68	7.45	0.3828
Boredom	96.66	0.63	2.71	0.4057
Fear	88.62	4.38	7.00	0.4297

Table 3.2: Epoch estimation performance of conventional ZFF method on Hindi emotional speech database

Emotion	**IDR** (%)	**MR** (%)	**FAR** (%)	**IDA (ms)**
Neutral	99.82	0.032	0.14	0.300
Angry	96.71	0.62	2.68	0.3460
Happy	92.17	0.21	7.62	0.3420
Boredom	99.78	0.02	0.20	0.2984

DYPSA algorithm used in this work is implemented using the programs given in the Voicebox speech processing toolbox [110]. The epochs estimated using DYPSA also show the same trend, confirming that the degradation is due to the large pitch variations in the emotional speech.

A similar trend in the epoch estimation performance can be observed in the Hindi emotional speech database collected for four speakers (2 males and 2 females) in 4 different emotions (Neutral, Angry, Happy and Boredom) having simultaneous EGG recordings. Ten randomly selected sentences from the Hindi broadcast news database, are used for the emotional speech recording. As the speech is recorded in three sessions, there are 120 files (3x4x10) available for each emotion. Table 3.2 shows epoch estimation performance obtained for Hindi emotional speech database. The degradation in the epoch estimation performance can be observed here also for angry and happy emotions. Even though the level of degradation performance is different, it follows a similar trend as in German emotional speech corpus.

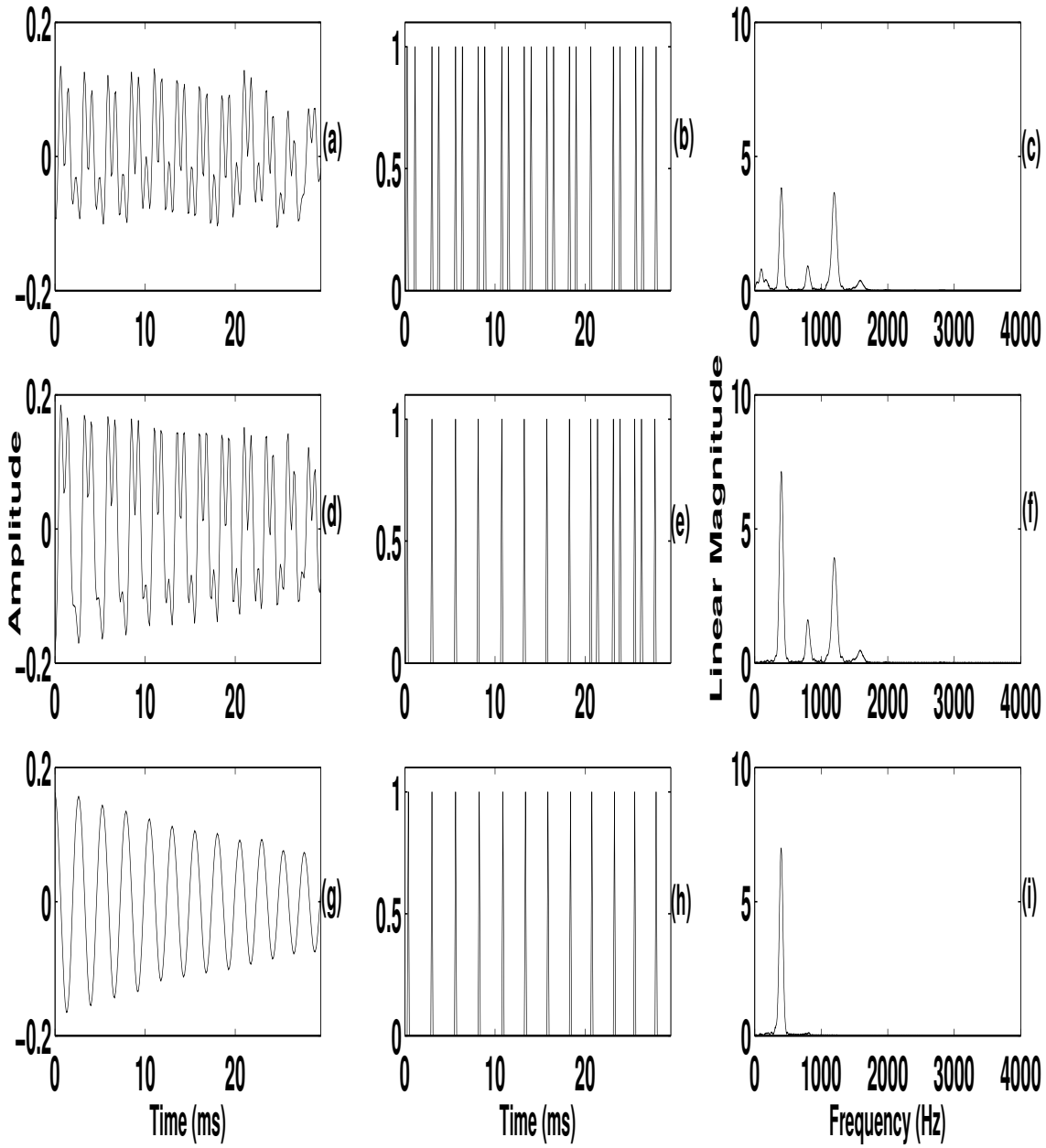

Figure 3.3: Comparison of conventional ZFF and the modified ZFF approaches. (a) The ZFFS obtained from a voiced segment of angry speech showing the spurious zero crossings, its (b) epochs and (c) STFT magnitude spectrum. (d) The modified ZFFS obtained by updating the window length, its (e) epochs and (f) STFT magnitude spectrum. (g) The ZFFS obtained by low pass filtering the modified ZFFS segments, (h) epochs estimated and its (i) STFT magnitude spectrum showing no frequency components beyond F_0.

3.4.2 Modified ZFF Method for Epoch Estimation in Emotional Speech

For reliably estimating the epochs from emotional speech, a refinement to the conventional ZFF algorithm is developed here. Instead of using fixed average pitch period as the window length for the trend removal of the zero frequency resonator output, the window length is updated for every short time segment of length 20-30 ms. A robust method to find the F_0 values from the ZFFS for acoustically degraded speech is described in [111]. In this method F_0 is computed as the frequency value corresponding to the highest magnitude in the short time fourier transform (STFT) of the ZFFS segment obtained from the conventional ZFF method and the window length is computed as the reciprocal of the F_0 value. This window length is used for the trend removal of the zero frequency resonator output to get the ZFFS for that particular speech segment. In the present work we use this approach for the emotional speech case which can be treated as a degraded speech due to the change in the psychological state of the speaker. The performance of this method for the estimation of epochs in emotional speech is given in Table 3.3. The performance improves significantly compared to the fixed window case demonstrating the significance of variable window length for trend removal in case of emotional speech.

Table 3.3: Epoch estimation performance of modified ZFF method by updating the window length for 25 ms segment of speech from different emotions.

Emotion	**IDR** (%)	**MR** (%)	**FAR** (%)	**IDA (ms)**
Neutral	99.56	0.04	0.29	0.2493
Angry	94.47	0.40	5.12	0.3746
Happy	94.36	0.48	5.16	0.3622
Boredom	99.57	0.03	0.40	0.2682
Fear	96.95	0.26	3.51	0.2792

Figure 3.3 compares the epoch estimation using conventional ZFF method and the proposed modified ZFF method. Figure 3.3(a) shows the ZFFS of a segment of angry speech. The corresponding zero crossings termed as epochs are plotted in Figure 3.3(b). The fixed window length results in spurious zero crossings. Figure 3.3(d) shows the ZFFS for the same segment obtained from the method given in [111]. Even though, the number of spurious zero crossings are reduced, it still leaves some spurious zero crossings unaltered. To analyze this, the corresponding magnitude of STFT of ZFFS for the both the methods are plotted in Figures 3.3(c) and (f). As it can be observed from both the figures, the magnitude of the harmonics beyond the fundamental frequence are comparatively stronger. To

alleviate this problem the trend removed ZFFS segment is passed through a low pass filter having cut off frequency equal to $1.05 \times F_0$. This is to suppress the strength of the pitch harmonics that come beyond F_0 as shown in Figure 3.3(i). This resulted in further reduction of spurious zero crossings as given in Figure 3.3(h). Similarly, all the ZFFS segments obtained are concatenated together to obtain the modified ZFFS. Modified epochs are estimated by finding the positive zero crossing in the modified ZFFS. The performance after the low pass filtering of ZFFS is given in Table 3.5. Table 3.6 also shows the improved epoch estimation performance for the Hindi emotional speech database. As given, it further improves with low pass filtering. Thus the proposed *modified ZFF method* employs both variable window length and low pass filtering for the epoch estimation.

The steps in the modified ZFF method can be summarized in Table 3.4

Table 3.4: The steps in the modified ZFF method for epoch extraction from emotional speech

- Compute the ZFFS using conventional ZFF method.
- Compute F_0 as the highest magnitude frequency value in the STFT of each 25 ms non-overlapping ZFFS frames.
- Derive window length as the reciprocal of F_0 for each frame.
- Remove the trend in the corresponding segment of the resonator output signal $y(n)$ (Equation (3.2)) using a moving average filter of length equal to the window length of that segment, as given in Equation(3.3).
- Low pass filter each of the trend removed ZFFS segment with a cut off frequency equal to $1.05 \times F_0$.
- Concatenate all the modified ZFFS segments to obtain the modified ZFFS signal.
- Hypothesize the negative to positive zero crossings of the modified ZFFS as the estimated epoch locations.

It is to be noted that, even though there is a significant improvement in the epoch estimation due to the modified ZFF method for the emotional speech case, still the performance in case of angry, happy and fear are not comparable with that of the neutral or boredom. To study the reason for this, the glottal waves from these five emotions are analyzed. Figure 3.1 shows the speech waveform, glottal wave and difference of glottal wave of five emotions under the condition of same speaker, text and syllable. The prominent features in the difference glottal wave are the impulse-like discontinuities.

Table 3.5: Epoch estimation performance of refined ZFF method on various emotions

Emotion	IDR (%)	MR (%)	FAR (%)	IDA (ms)
Neutral	99.61	0.09	0.29	0.2422
Angry	96.20	0.37	3.43	0.3569
Happy	95.27	0.43	4.30	0.3544
Boredom	99.55	0.06	0.39	0.2688
Fear	96.95	0.25	2.80	0.2721

Table 3.6: Epoch estimation performance of refined ZFF method on various emotions from Hindi emotional speech database

Emotion	IDR (%)	MR (%)	FAR (%)	IDA (ms)
Neutral	99.56	0.64	0.37	0.2549
Angry	98.33	0.68	0.99	0.2668
Happy	99.27	0.08	0.65	0.2313
Boredom	99.54	0.05	0.41	0.2799

The amplitude of these impulse like discontinuities give an indication about the intensity with which the closing of vocal folds occur. Hence the difference glottal wave may be treated as the representative of *strength of excitation*. The strength of excitation of the emotions like angry, happy and fear are low as compared to neutral and boredom emotions. This is due to the rapid pitch variations and/or the difference in the nature of vocal folds activity for these emotions. The rapid pitch variations cause the vocal folds to vibrate with the lower suction pressure and hence the reduced strength of excitation. Depending on the psychological state of particular emotion, the tension associated with the vocal folds and associated muscle structure may be different. Because of these factors, the impulse strength may not be as prominent in the case of neutral and boredom emotions. These can be observed by comparing the difference glottal waves of different emotions. This in turn may result in the reduction of energy around the zero frequency region, leading to the spurious epochs detection. This may be the reason for the reduced epoch estimation performance in case of angry, happy and fear emotions. Further exploration and understanding is required in this direction.

3.5 Estimation of Expressive Parameters from Modified ZFFS

Modified ZFF method is used to compute the instantaneous F_0 contour and strength of excitation from the emotional speech utterances for the expressive speech analysis. The sentence level duration of emotional utterances are also used for expressive speech analysis.

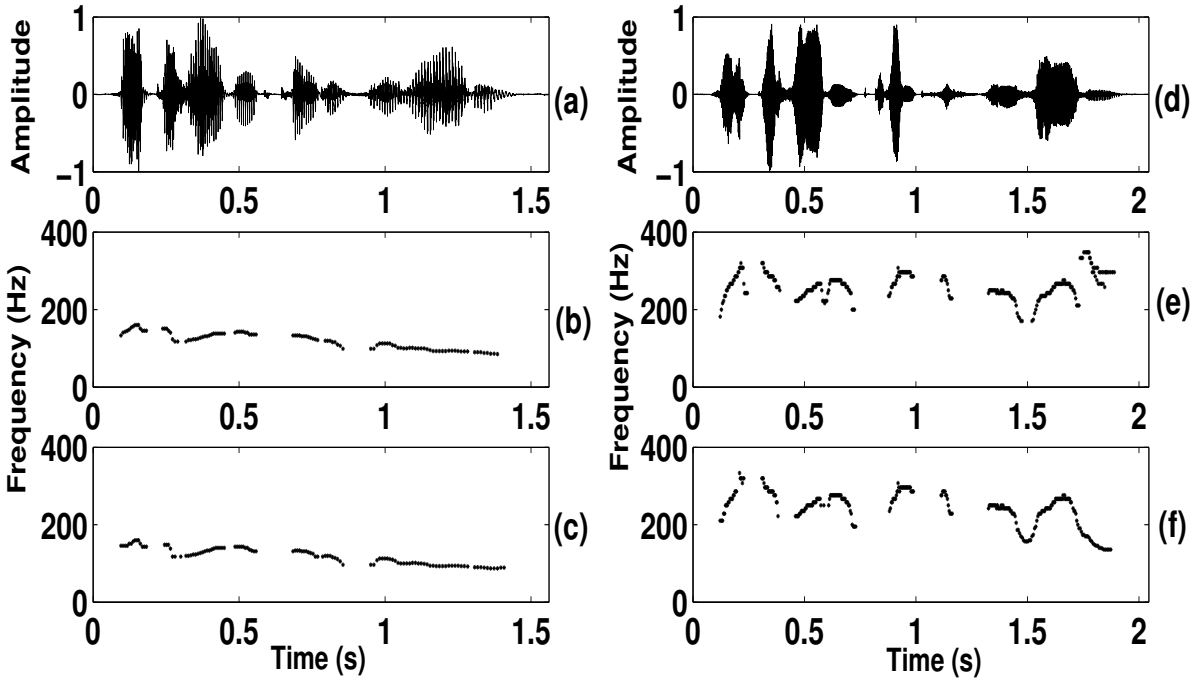

Figure 3.4: Comparing the F0 contour obtained using conventional and refined ZFF method. The F0 contour obtained from, a neutral ((a)-(c)) and angry ((d)-(f)) speech signals using conventional and refined ZFF methods.

3.5.1 F_0 Parameters [13]

The instantaneous pitch period or the epoch interval can be found by the successive difference between the estimated epoch locations. Taking the reciprocal of each epoch interval, multiplied by the sampling frequency gives the fundamental frequency (F_0) [13]. Figure 3.4 shows the F_0 contours derived from the estimated epochs using conventional and modified ZFF methods for neutral and angry emotional speech signals. It is to be observed from the Figures 3.4(b) and (c) that the F_0 contours obtained using conventional and modified ZFF methods remain nearly same for the neutral emotion. For angry emotion, the Figures 3.4(e) and (f) indicate that the F_0 values obtained using the modified ZFF method are more continuous than that obtained using conventional ZFF method. Hence the merit of the modified ZFF method.

3.5.2 Strength of Excitation [90]

The strength of excitation is referred to as the amplitude of the glottal impulses of the excitation waveform. Hence the strength of excitation is otherwise known as the rate of glottal closure [90]. The strength of excitation is proportional to the peak intensity of the glottal wave derivative at the glottal

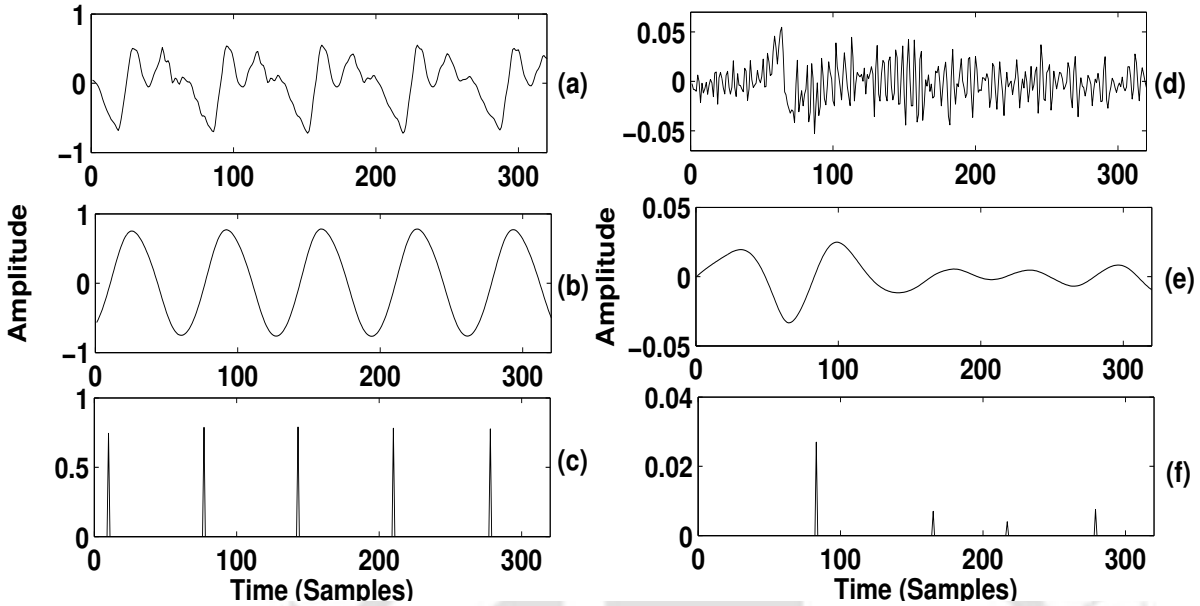

Figure 3.5: Strength of excitation in GA and non-GA Regions. (a) a voiced segment of speech waveform , (b) corresponding ZFFS segment and (c) Strength of excitation. (d) An unvoiced waveform segment, (e) corresponding ZFFS segment and (f) strength of excitation.

closure instants. The strength of excitation can be computed as the slope of the ZFFS computed around each epoch location. The regions of the glottal activity (GA) can be determined from the strength of excitation. Figure 3.5 plots the strength of excitation of the GA and non-GA regions. The lower strength values in the non-GA regions (Figure 3.5(f)) have to be observed as compared to strength of excitation in the GA regions. Also the ZFFS segment (Figure 3.5(e)) in the non-GA region is more random as compared to the GA region ZFFS (Figure 3.5(b)).

As the strength of excitation is significant in the GA regions than the non-GA region, the GA detection can be done by putting a threshold in the strength of excitation values. Typical threshold for GA detection is experimentally found to be 10% of the maximum strength of excitation value of the entire speech utterance. The characteristics of the strength of excitation values and instantaneous F_0 values in the GA regions are considered for the expressive speech analysis.

3.5.3 Duration parameters

For the present work, the sentence duration of the emotional speech utterances are considered as the duration parameters for the expressive speech analysis. The sentence duration of various emotions are measured directly by computing the length of emotional speech utterances in the database. The

Table 3.7: Expressive parameters of different emotions from EGG and speech of German emotional speech corpus.

Emotion	F_{0Avg} (Hz)		Std. Dev. F_0 (Hz)		Epoch Strength (Norm. Amplitude)		Dur. (s)
	EGG	Speech	EGG	Speech	EGG	Speech	
Neutral	174.53	180.86	30.73	35.10	0.60	0.51	2.26
Angry	279.67	301.59	57.09	60.44	0.50	0.41	2.59
Happy	259.76	287.17	47.70	55.58	0.47	0.38	2.43
Boredom	168.08	175.60	36.34	39.36	0.59	0.52	2.70
Fear	243.19	249.10	45.66	45.56	0.54	0.42	2.240

sentence duration of various utterances are measured in seconds (s).

3.6 Expressive Parameters from EGG and Speech

Figure 3.6 plots the EGG signals, strength of excitation and F_0 contours estimated from the EGG signals using ZFF method for five different emotions, for a given speaker and sentence. The envelopes of EGG signals are different indicating different way of modulation of air flow for each emotion. The variation in the nature of strength of excitation contours across different emotions indicate the difference in the rate of closure of vocal folds. It is interesting to observe that the maximum values of strength of excitation is even more compared to that of EGG. Hence the amplitudes in the two cases may represent different aspect of source information. The F_0 contours are different indicating the influence of emotion on the pitch periodicity. Figure 3.7 represents the same expressive parameters derived from the speech signals using ZFF method. The difference in the envelopes of speech signals and strength of excitation contours are attributed to the effect of emotion. The difference between the envelopes of strengths of excitation derived using EGG and speech indicate that the observation using speech case may not directly correlate well. However, the general trend remains same in both the cases. The similarity in the F_0 contours can be observed for both the cases. Apart from F_0 and strength of excitation contours, the overall duration of the utterances are also vary for different expressions. From Figures 3.7 and 3.6, the variations in the length of emotional speech utterances indicate the effect on duration in various emotions.

3.6.1 Comparison of Expressive Parameters

As per Figures 3.6 and 3.7, angry and fear emotions have highest range of pitch values, whereas boredom has the lowest pitch contour values. In the case of strength of excitation, boredom has the

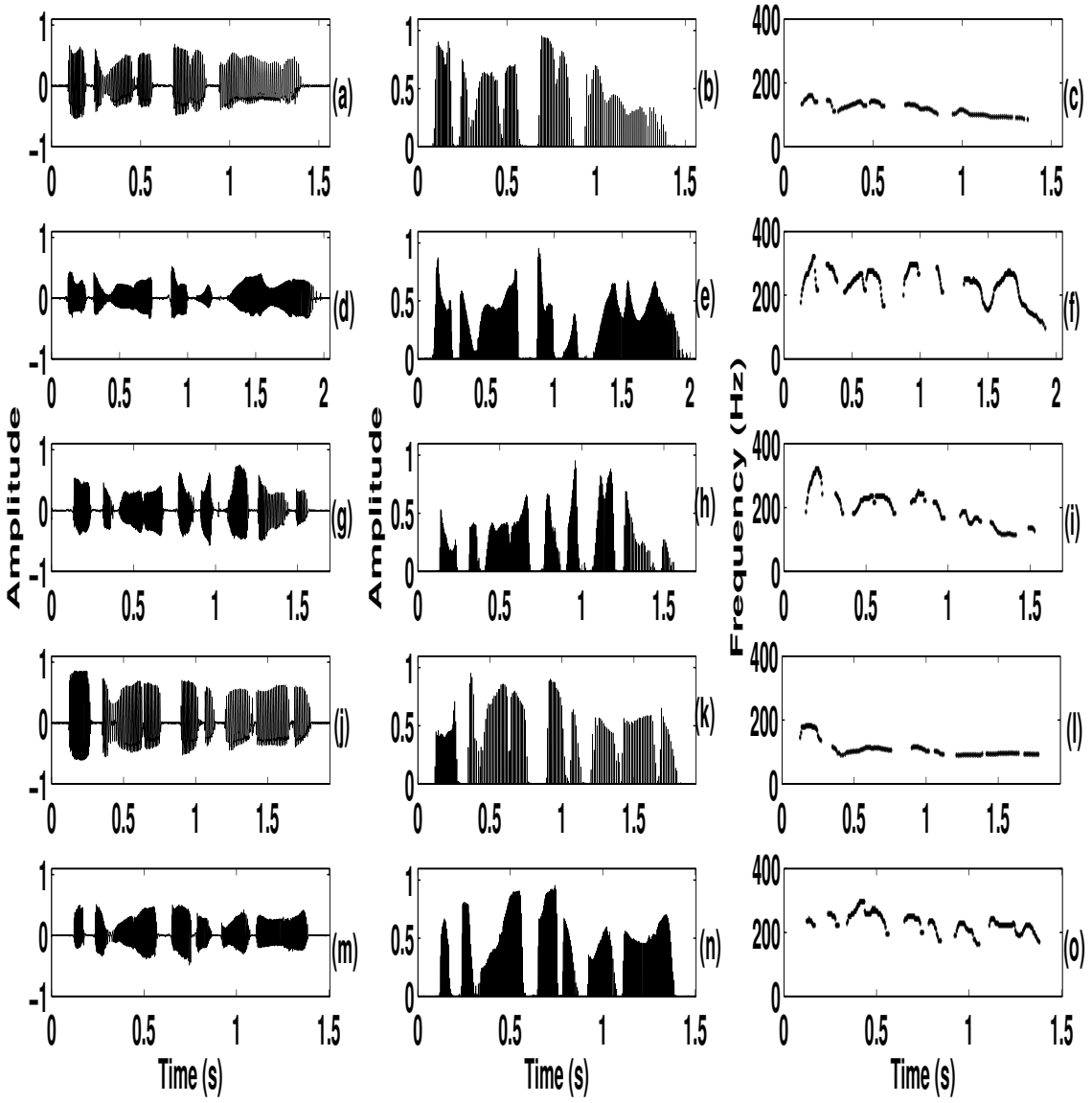

Figure 3.6: *EGG, estimated strength of excitation and instantaneous F_0 contours from EGG of Neutral ((a)-(c)), Angry ((d)-(f)), Happy ((g)-(i)), Boredom ((j)-(l)) and Fear((m)-(o)) emotions, respectively.*

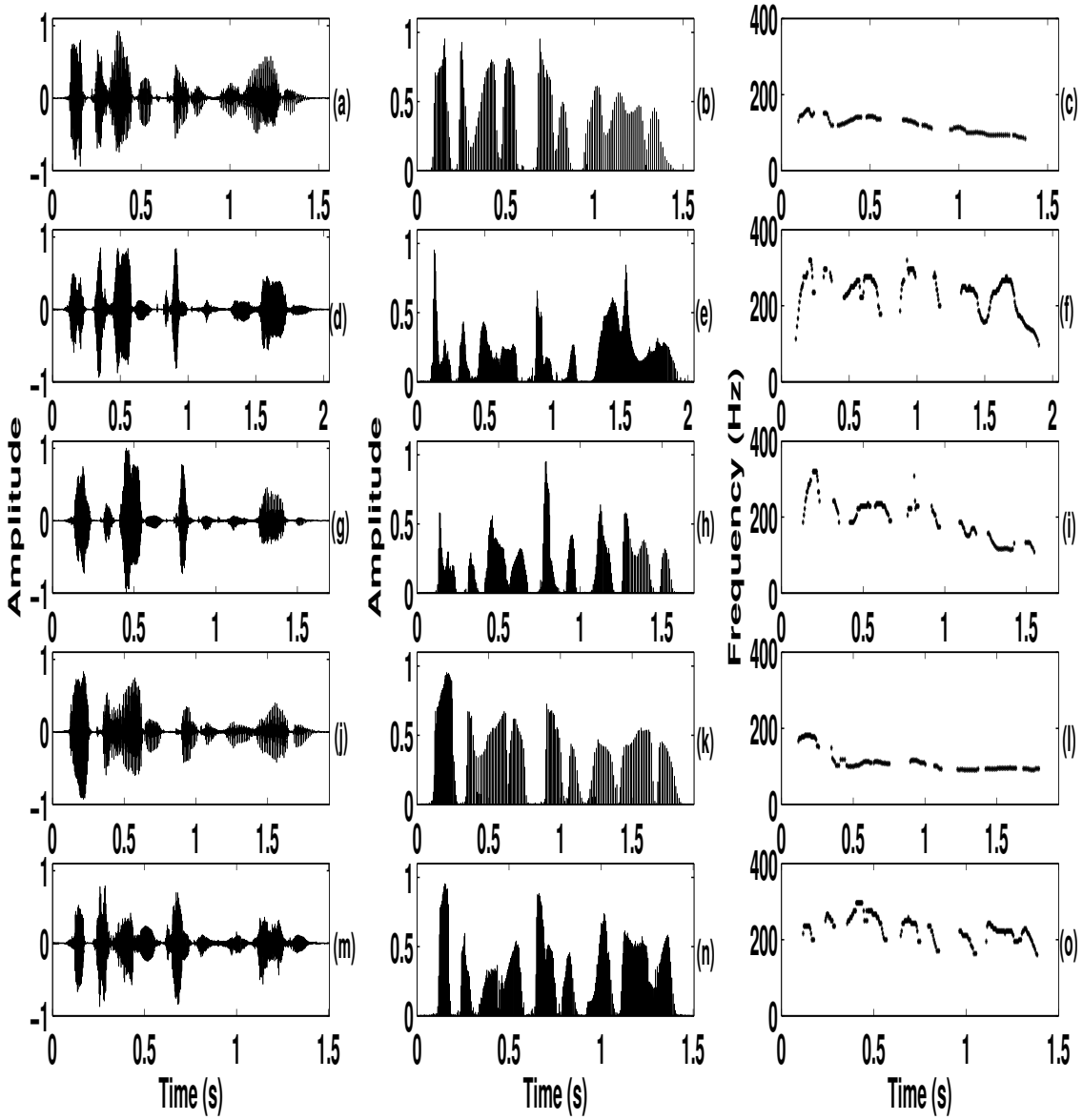

Figure 3.7: *Speech waveforms, estimated strength of excitation and instantaneous F_0 contours of Neutral ((a)-(c)), Angry ((d)-(f)), Happy ((g)-(i)), Boredom ((j)-(l)) and Fear((m)-(o)).*

Table 3.8: Expressive parameters of different emotions from EGG and speech of German emotional speech corpus for Male speaker case. The prosodic parameters are computed for two texts and three male speakers

Emotion	F_{0Avg} (Hz)		Std. Dev. F_0 (Hz)		Epoch Strength (Norm. Amplitude)		Dur. (s)
	EGG	Speech	EGG	Speech	EGG	Speech	
Neutral	107.61	108.86	18.81	18.71	0.52	0.50	2.17
Angry	189.93	196.03	48.11	46.96	0.38	0.38	2.72
Happy	180.81	199.02	51.14	47.03	0.35	0.42	2.33
Boredom	110.70	111.44	26.15	26.56	0.51	0.57	2.83
Fear	161.26	162.15	26.55	28.83	0.44	0.43	2.26

highest, and angry and fear have the lowest values. Tables 3.7 provide F_{0Avg}, standard deviation of F_0, strength of excitation and the overall duration of the utterances computed across nine sentences and eight speakers (5 males and 3 females) using EGG and speech signals of German emotional speech database. In case of duration, boredom showed the increased sentence duration compared to other emotions in the database. Even though the respective values are different in the two cases of EGG and speech, the trend is same across all the parameters. This infers that if trend is the required information, then we can derive the same either using speech or EGG. Table 3.10 presents the same expressive parameters estimated from Hindi emotion database with four emotions across 10 sentences and 4 speakers (2 males and 2 females). From Tables 3.7 and 3.10, it has to be observed that the expressive parameters estimated from corresponding emotions of German and Hindi languages follow almost similar trend. However, a varied trend in the overall duration characteristics is observed for angry emotions of German and Hindi languages. Table 3.8 and Table 3.9 show the variations of the prosodic parameters across various expressions in German emotional speech database for male and female speakers respectively. Three male and female speakers for two different texts are used for the analysis of prosodic parameters. The observed trend for whole German emotional database remained same when the prosodic parameters are separately computed for the male and female speakers.

3.7 Summary & Conclusions

In this chapter, variations in the expressive parameters like instantaneous F_0, strength of excitation and duration are observed for various emotions by analyzing the EGG recordings. To estimate these expression specific parameters ZFF method is used. Even though conventional ZFF method reliably estimates accurate epochs location from neutral speech signals, it shows degradation in the

Table 3.9: Expressive parameters of different emotions from EGG and speech of German emotional speech corpus for Female speaker case. The prosodic parameters are computed for two texts and three female speakers

Emotion	F_{0Avg} (Hz)		Std. Dev. F_0 (Hz)		Epoch Strength (Norm. Amplitude)		Dur. (s)
	EGG	Speech	EGG	Speech	EGG	Speech	
Neutral	191.38	193.54	36.94	36.85	0.57	0.50	2.06
Angry	296.50	301.59	63.98	58.51	0.32	0.39	2.22
Happy	270.79	277.65	58.28	57.13	0.45	0.36	2.20
Boredom	174.67	177.93	37.21	38.27	0.58	0.55	2.34
Fear	241.70	242.35	40.79	36.52	0.41	0.44	2.06

Table 3.10: Expressive parameters of different emotions from EGG and speech of Hindi emotional speech corpus.

Emotion	F_{0Avg} (Hz)		Std. Dev. F_0 (Hz)		Epoch Strength (Norm. Amplitude		Dur. (s)
	EGG	Speech	EGG	Speech	EGG	Speech	
Neutral	190.9	196.7	67.7	72.8	0.58	0.53	3.44
Angry	239.4	246.7	68.0	90.3	0.41	0.38	2.81
Happy	222.6	236.1	63.6	79.3	0.50	0.54	4.11
Boredom	151.4	156.9	60.1	76.7	0.62	0.52	4.73

estimated epochs in case of emotional speech. In this chapter, a modified ZFF method is proposed by smoothing the ZFFS obtained by the conventional ZFF method for removing the spurious zero crossings introduced due to rapid pitch variations in the emotional speech. Performance evaluation of refined ZFF method indicates the robustness of the epoch extraction for various emotions in two languages. The expression specific features are estimated from the refined ZFFS. The next chapter presents the proposed epoch based dynamic prosody modification tool using ZFFS to incorporate the variations in the expressive parameters for neutral to expressive speech conversion.

4

Epoch Based Dynamic Prosody Modification

Contents

4.1 Objective

Modifying the prosody parameters like pitch, duration and strength of excitation by desired factor is termed as prosody modification. The objective of this chapter is to develop an epoch based dynamic prosody modification method which can be used for effectively incorporating the dynamic variations in the prosodic parameters with reduced perceptual distortion. A computationally fast static prosody modification method with reduced perceptual distortions using ZFF method is developed in the first part of the chapter. The perceptual quality of the existing epoch based prosody modification is improved by using the accurate epochs location estimated using ZFF method instead of the GD based epochs locations as pitch marks in the existing epoch based prosody modification. The computational complexity is further reduced by performing the prosody modification directly on the speech waveform samples. The second part of the chapter develops a dynamic prosody modification method based on zero frequency filtered signal (ZFFS), a byproduct of zero frequency filtering (ZFF). The existing epoch based prosody modification techniques use epochs as pitch markers and the required prosody modification is achieved by the interpolation of epoch intervals plot. Alternatively, this work proposes a method for prosody modification by the resampling of ZFFS. Also the existing epoch based prosody modification method is further refined for modifying the prosodic parameters at every epoch level. Thus providing more flexibility for prosody modification. The general framework for deriving the modified epoch locations can also be used for obtaining the dynamic prosody modification from existing PSOLA and epoch based prosody modification methods. The quality of the prosody modified speech is evaluated using waveforms, spectrograms and subjective studies. The dynamic prosody modified speech files synthesized using the proposed, epoch based and TD-PSOLA methods are available at http://www.iitg.ac.in/eee/emstlab/demos/demo5.php.

4.2 Introduction

Prosody modification is the process of changing the pitch, duration and strength of excitation of a recorded or synthesized speech according to the given modification factor [112]. The resulting prosody modified speech should be natural, free from spectral and temporal distortions, and preserve the speaker dependent features of the original speech signal [112] [113]. The pitch and duration modifications are extensively used as part of unit selection based text to speech synthesis (TTS) engine to remove the discontinuities due to the concatenation of various sound units selected from

different contexts of a large database [64]. The emotion conversion can be achieved by modifying the prosody parameters of neutral speech according to the target expression [7,17,21]. The prosodic information in emotional speech changes continuously. The well known existing prosody modification methods work well for static modification factors. On the other hand what is needed in tasks like emotional speech conversion is a *dynamic prosody modification* method. Such a method offers flexibility for prosody modification and hence the motivation for this work.

There are several methods discussed in the literature for prosody modification [11,63,99]. Among these PSOLA and epochs based prosody modification methods are popular methods for prosody modification. The perceptual quality of the prosody modified speech using PSOLA methods depends on the accuracy of the pitch markers estimation. As estimating epochs from speech provide more accurate pitch marker locations, prosody modification can also be performed using epochs as the anchor points [11,30,114]. The epochs in speech refer to the instants of glottal closure in case of voiced speech and random instants like onset of frication or burst in case of unvoiced speech [11,12,88,104]. The prosody modification approach developed in [11] uses epochs estimated using group delay (GD) analysis as the pitch markers.

The prosody modification techniques discussed above are proposed mainly for prosody modification by time invariant modification factor where the pitch, duration and energy of the original waveform are scaled by static values, uniformly across the whole utterance [113,115–117]. However, in real scenario, for instance, emotional speech, the prosody parameters vary continuously and non-uniformly. Thus incorporating the prosody parameters by static prosody modification may not be effective in the applications like neutral to target emotional speech conversion. The dynamic prosody modification technique capable of non-uniformly modifying the prosody parameters is required. The works reported in [115,116] propose an improvement over PSOLA for achieving time varying prosody modification. Here, the prosody parameters of the given syllable are modified dynamically according to the prosodic trajectory predicted by the prosodic processing unit of the text analysis stage of a concatenative TTS system. The shape invariant prosody modification methods based on sinusoidal analysis and synthesis are proposed in [113,117]. Instead of using arbitrary phase values for sinusoidal synthesis, the shape invariance is achieved by keeping the phase relations of excitation and system components consistent with the prosody modification factors during the sinusoidal synthesis for every pitch interval [113]. Along this direction, the present work proposes an epoch based dynamic prosody modification method

using zero frequency filtered signal (ZFFS). Also the proposed method provides a general frame work where any existing technique can be used to modify according to desired dynamic prosody modification factors.

The epochs and ZFFS are the byproducts of the ZFF approach. The epoch based approaches in [11] compute the epoch intervals plot and interpolate it according to the prosody modification factor. The modified epochs location are then derived and used for generating the prosody modified speech. Alternatively, this work proposes that the ZFFS itself can be directly resampled according to the modification factors and used for deriving the modified epochs location. Hence the basis for the proposed dynamic prosody modification method. Further, as will be described later, in the proposed method, the prosody modification factors can be changed dynamically and hence offers more flexibility. The novelty of this work include the following:

- Use of ZFF epochs to improve the perceptual quality and reduce the computational complexity in the existing epoch based prosody modification.
- A method for deriving modified epoch locations using ZFFS.
- A method for dynamically changing the prosody modification factor using ZFFS.

Apart from the above mentioned novel aspects, the existing solution in the literature for identifying glottal activity (GA) regions using ZFFS [118] is integrated into the epoch based prosody modification approach for performing prosody modification only in the voiced regions.

The rest of the chapter is organized as follows: A computationally fast static prosody modification method is proposed in Section 4.3. Section 4.4 describes the development of the proposed epoch based dynamic prosody modification method using ZFFS. The experimental results and discussion are given in Section 4.5. Finally Section 4.6 summarizes the works done in this chapter towards epoch based prosody modification.

4.3 Computationally Fast Static Epoch Based Prosody Modification

There are three main steps involved in the proposed fast prosody manipulation.

(i) Deriving the epochs from the speech signal by the ZFF method.

(ii) Deriving a modified (new) epoch sequence according to the desired prosody (pitch and duration).

(iii) Deriving a modified speech signal from the modified epoch sequence.

4.3.1 Deriving the epochs using ZFF method

The epochs estimated using the GD based approach used as the pitch marks in the existing epoch based prosody modification is replaced with the ZFF epochs in the proposed fast prosody modification. The reasons for using ZFF epochs for prosody modification are the following:

- Epochs estimated using ZFF method are more accurate than the GD based epochs
- ZFF method is computationally less complex than the GD method

Table 4.1: Performance of ZFF and GD methods for determining instants of significant excitation.

Method	IDR (%)	MR (%)	FAR (%)	IDA (msec)
GD Method	94.48	4.07	1.45	0.45
ZFF Method	99.67	0.02	0.31	0.26

The CMU-Arctic database having simultaneous recordings of speech and EGG signals [119] was used to evaluate the GD and ZFF methods for determining the epochs location. This database consists of 1132 phonetically balanced English sentences, spoken by two male and one female talkers. The duration of the speech utterance is about 3 sec. For each speaker 100 sentences are randomly selected forming a set of 300 sentences. The reference epochs location are extracted from the voiced segments of the EGG signals by finding the peaks in the differenced EGG signal. The performance of epochs estimation was evaluated only in the voiced segments which contains a total of 42065 reference epochs. Table 4.1 compares the performance of epochs estimated using ZFF and GD methods. The improved epoch identification rate and identification accuracy and reduced epoch miss rate and false alarm rate of the ZFF based method as compared to the GD based method can be observed from the Table 4.1.

4.3.2 Deriving the modified epochs location for prosody modification

The modified epochs location for the epochs based prosody modification is obtained by deriving the epoch interval plot from the estimated epochs locations.The term epoch interval refers to the interval between successive epochs [11]. This epoch interval corresponds to the instantaneous pitch period. For duration modification, the epoch interval plot obtained is interpolated and resampled according to the static duration modification factor. For pitch modification, the interpolated epoch interval plot is scaled according to the static pitch modification factors [11]. Starting from a point,

new locations are derived from the modified epoch interval plot (interpolated and resampled original epoch interval). These new locations will be the modified epoch locations for the desired pitch and duration modification.

4.3.3 Waveform Generation

After obtaining the modified epochs, the next step is to derive the speech signal. For this, the original epochs closest to the modified epochs are determined. The speech samples around the original epoch are placed starting from the corresponding new epoch. Since the value of the desired epoch interval is different from the value of the corresponding original epoch interval, it is necessary to either delete some speech samples or append some new speech samples to fill the new epoch interval. Deletion of required number of speech samples is made in the tail portion of the selected speech samples. Insertion of required number of speech samples is achieved by suitably resampling about 10% of the tail portion of the selected speech samples and appending them to the end. Changes in the spectral features are visible for large modification factors of pitch period and duration as can be seen in the narrowband spectrograms for pitch period modification by 2.0 given in Fig. 4.1. The degradation seems to be more for residual modification compared to waveform modification.

4.3.4 Computational efficiency of the proposed fast prosody modification method

The first observation is that the ZFF method works directly on the speech signal, and hence does not need LP analysis as in the GD method. The ZFF does not need GD analysis which is a computation intensive process. Finally, the ZFF method does not employ block processing for every sample shift to determine the instants. A speech signal (about 3 sec duration) for the text *"Don't ask me to carry an oily rag like that"* taken from the TIMIT database is used for determining the instants of significant excitation using both the GD and ZFF methods. Both the Matlab programs were run on the same computer. The time taken for determining the instants was about 4.83 secs in the case of the GD method, and it was only about 15.6 msec in the case of the ZFF method, demonstrating the computational efficiency of the ZFF method for finding the instants of significant excitation.

For analyzing the computational efficiency, the same speech utterance from TIMIT database is used for prosody modification (pitch period by 0.66, and duration by 2) using (i) epochs from the GD method and residual modification (EGD-RM), (ii) epochs from the ZFF method and residual modification (EZFF-RM), and (iii) epochs from the ZFF method and speech waveform modification

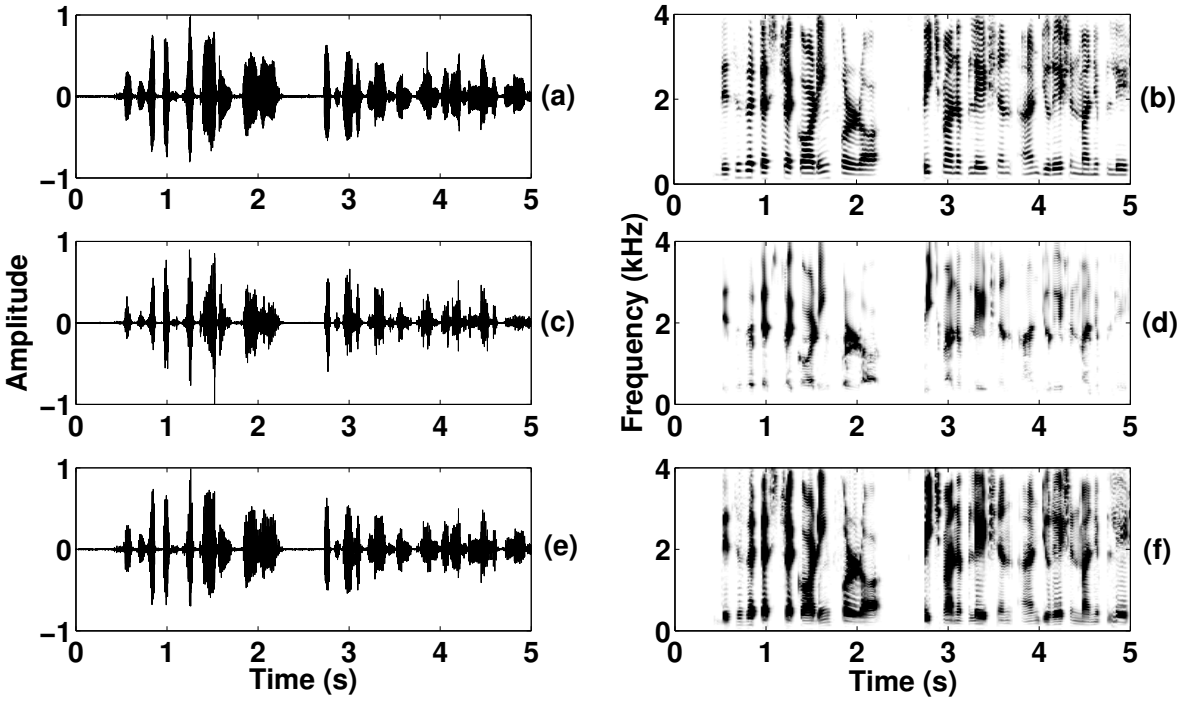

Figure 4.1: *Speech waveforms and their narrowband spectrograms for original speech ((a) and (b)), pitch modification by factor of 2.0 for EZFF-RM((c) and (d)) and EZFF-SM ((e) and (f)).*

Table 4.2: Computational time for prosody modification.

Method	Time for prosody modification
EGD-RM	6.22 sec
EZFF-RM	1.78 sec
EZFF-SM	0.93 sec

(EZFF-SM). The time taken for each method is tabulated in Table 4.2. The time for the proposed prosody modification (EZFF-SM) is significantly lower than the other two methods.

4.3.5 Subjective Evaluations

Performance of the proposed prosody modification (EZFF-SM) method is compared with EGD-RM, EZFF-RM, and also with the PSOLA method operating on the speech waveform called time domain (TD)-PSOLA . The TD-PSOLA method performs pitch and time-scale modifications of the speech waveform using pitch markers as anchor points. Perceptual evaluation was carried out by conducting subjective tests with 10 research scholars. Two sentences of Indian English accent (1 male and 1 female) and two sentences of American English accent (1 male and 1 female) are used for

Table 4.3: Ranking used for judging the distortion of the speech signal for different modification factors

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

Table 4.4: Mean opinion scores for different pitch modification factors.

Method	0.66	1.5	2.0
TD-PSOLA	2.31	2.97	2.39
EGD-RM	3.46	2.84	2.03
EZFF-RM	3.90	3.38	2.34
EZFF-SM	3.44	3.68	3.03

prosody modification. For each sentence the pitch period was modified by three factors: 0.66, 1.5 and 2. Similarly, the duration was modified by factors: 0.5, 1.5 and 2.5. After the required modification using the three methods, the filenames were coded to avoid bias toward a specific method. The tests were conducted by playing the speech signals through headphones. In the test, the subjects were asked to judge only the distortions present in the speech for various modification factors on a five-point scale given in Table 4.3.

The Mean Opinion Score (MOS) for each of the pitch period and duration modification factors are given in Tables 4.4 and 4.5, respectively. For moderate modification factor of 1.5, all the methods seem to provide at least fair quality speech, and among these the proposed method provides the best possible speech quality. For all the modification factors, the scores for the methods based on the knowledge of the instants of significant excitation are better than the TD-PSOLA. This demonstrates the significance of the instants of significant excitation for prosody modification. The TD-PSOLA method uses pitch

Table 4.5: Mean opinion scores for different duration modification factors.

Method	0.5	1.5	2.5
TD-PSOLA	2.75	3.17	2.05
EGD-RM	3.41	3.86	2.59
EZFF-RM	3.79	4.67	3.72
EZFF-SM	3.97	4.52	3.88

Table 4.6: Comparison of significance of differences in MOS scores of different methods with EZFF-SM for pitch modification and duration modification.

Pitch Modification			
Modification factors	EZFF-RM	EGD-RM	TD-PSOLA
0.66	<80	<80	>=99
1.5	<80	>=95	>=90
2	>=90	>=97.5	<80
Duration Modification			
0.5	<80	<80	>=99.5
1.5	<80	>=95	>=99.5
2.5	<80	>=95	>=99.5

markers computed by conventional pitch extraction methods like autocorrelation analysis. From the speech production and perception point of view, most of the speech signal characteristics are present in the samples around the instants of significant excitation. Thus preserving these samples using the knowledge of the instants of significant excitation results in better speech quality. Table 4.6 gives the level of significance obtained by student-t distribution of the difference in MOS of the proposed EZFF-SM from other methods for pitch and duration modifications [120]. From Table 4.6, it can be seen that difference in MOS score for TD PSOLA and EZFF-SM is significant as the level of confidence is more than 99. Also it is to be noted that the level of significance of difference between MOS scores of EZFF-SM and EZFF-RM is less than 80 both in the case of pitch and duration modification. Lower values of confidence level between the MOS of EZFF-SM and EZFF-RM indicates nearly the same perceptual qualities of speech produced from both the approaches. The difference between speech produced by ZFF based direct waveform prosody modification and GD based residual prosody modification is also significant as given in the Table 4.6.

It is interesting to note that the degradation in perception quality is less in duration modification compared to pitch period modification. This is because in duration modification the waveform in each pitch period is preserved. Only the number of pitch cycles are either reduced or increased depending on the modification factor. On the other hand, in pitch period modification the waveform is either truncated or stretched depending on the modification factor. Thus the degradation will be more for large pitch period modification factors. The residual modification seems to introduce higher distortion compared to waveform modification for large pitch modification factors as can be seen from column four in Table 4.4.

4.4 Dynamic Prosody Modification using Zero Frequency Filtered Signal

The epochs estimated using the ZFF method are used as the pitch markers for modifying the duration and pitch of the given speech signal. The modified epochs location for prosody modification are obtained by resampling of ZFFS. To achieve efficient prosody modification with reduced computation complexity, the prosody modification is performed on the speech waveform itself. There are mainly three steps in the prosody modification, namely,

(i) Finding the epochs location using ZFF method.

(ii) Deriving the modified epochs location according to desired prosody using ZFFS.

(iii) Reconstructing the speech waveform according to modified epochs location.

Section 4.4.1 and Section 4.4.2 give the detailed description of the proposed method for duration and pitch modification for dynamic modification factors, respectively.

4.4.1 Dynamic Duration Modification

The term epoch interval refers to the interval between successive epochs [11]. In dynamic duration modification, the duration of all the epoch intervals in the speech signal are modified according to the given time varying duration modification factor β_i. The β_i represents the duration modification factor for the signal samples present in the epoch interval starting from the i^{th} epoch. The steps involved in the dynamic duration modification are as follows:

(i) Find the epochs location using the ZFF method.

(ii) Find the positive integers P_i and Q_i such that $\beta_i = P_i/Q_i$, where β_i is the duration modification factor for the i^{th} epoch interval.

(iii) The ZFFS samples for the epoch interval starting from i^{th} epoch are resampled according to β_i.

(iv) For deriving the modified epochs location, the interval between every i^{th} and $i + Q_i^{th}$ positive zero crossings of the resampled ZFFS is divided into P_i equal intervals.

(v) The begin of each of these intervals is identified as the *modified epoch location*.

(vi) For each of the modified epoch location, the nearest original epoch location is found out and the speech samples in its epoch interval are copied.

(vii) The steps 3 to 6 are repeated for all the epoch intervals.

(viii) The final result is the *dynamic duration modified speech signal*.

Figure 4.2 illustrates the proposed dynamic duration modification method using a segment of voiced speech given in Figure 4.2(a). Figure 4.2(b) shows the ZFFS of this speech. Figure 4.2(c) is the resampled ZFFS for the dynamic duration modification. The resampled ZFFS is obtained by considering the original ZFFS segment from every epoch interval, resampling it according to the desired β_i and then concatenating. For this example, the duration modification factors are varied from 1.5 (initial region) to 0.5 (final region). In Figure 4.2(d), the original epochs location obtained from the original ZFFS are marked as 'O'. Figure 4.2(e) shows the positive zero crossings of the resampled ZFFS indicated by 'R'. Figure 4.2(f) shows the modified epochs location obtained for the desired dynamic duration modification. The corresponding original epochs are shown in Figure 4.2(g). The dynamic duration modified speech segment obtained by copying the speech samples using original epochs 'O' (Figure 4.2(g)) is shown in Figure 4.2(h).

Figure 4.3 demonstrates the duration modification applied for static duration modification for the same voiced segment of speech used in Figure 4.2. Static duration modification is achieved by keeping the duration modification factors same for all the epoch intervals. The Figure 4.3 demonstrates the static duration modification for β_i=1.5 for all i, where i represents the i^{th} epoch. Thus the existing static duration modification is indeed a special case of the proposed dynamic duration modification. As illustrated later, the static modification case helps in illustrating the effectiveness of proposed dynamic duration modification.

In the dynamic duration modification proposed above, all the epochs in the original signal are considered for the duration modification as performed in [11]. However, modifying the duration of the epochs in the unvoiced regions like fricatives and other unvoiced consonants causes unnaturalness to the duration modification [121]. Intuitively also this is true because when we increase or decrease the duration while speaking, most of the modification takes place only in the voiced regions. In particular, vowel-like regions [121–123]. Thus the original duration of the unvoiced regions needs to be preserved and the duration modification is to be performed only on the voiced regions.

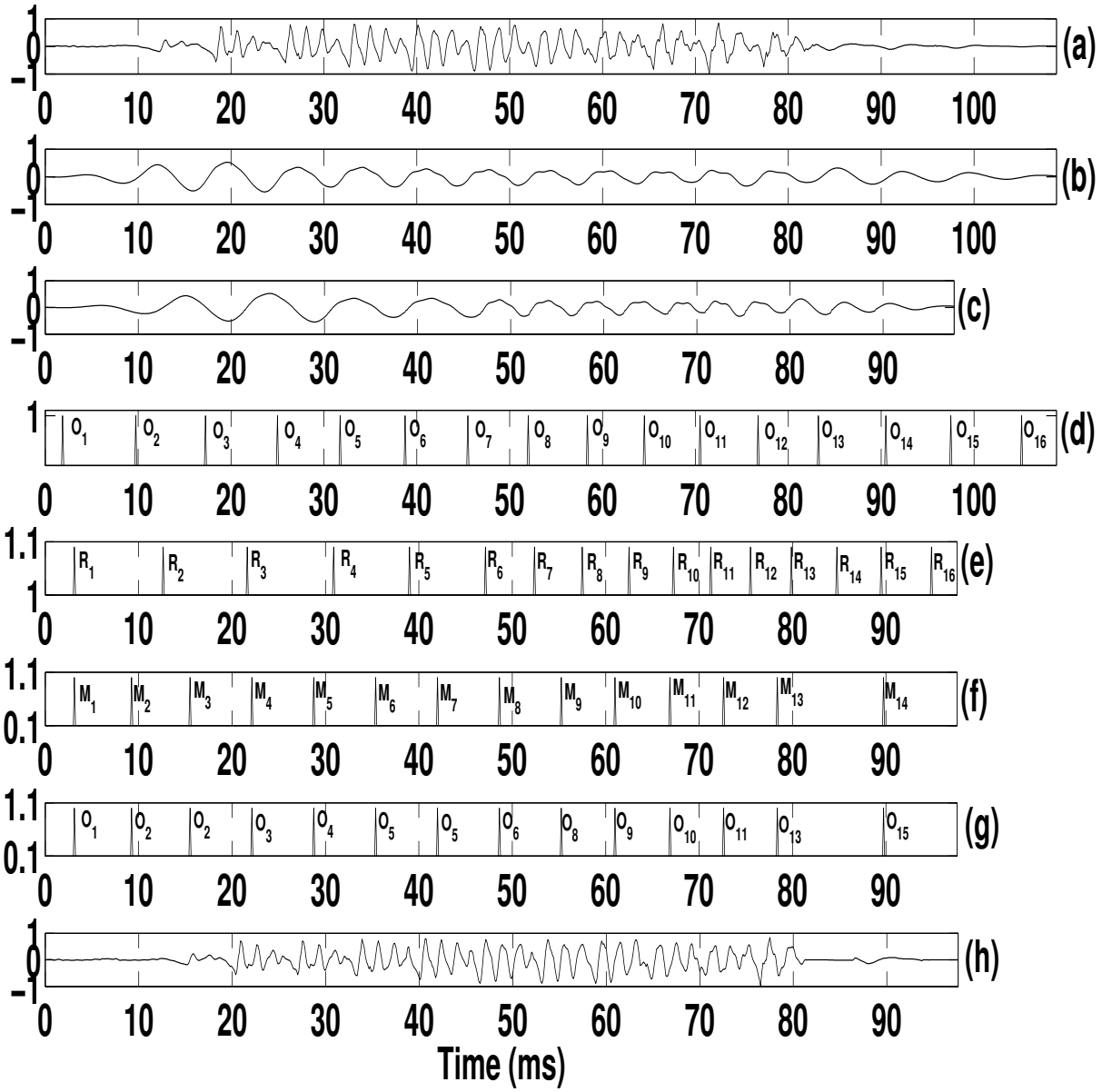

Figure 4.2: Deriving modified epochs location for dynamic duration modification with duration modification factors varied dynamically from 1.5 to 0.5. (a) A voiced speech segment of the original speech, (b) ZFFS from the voiced segment, (c) Resampled ZFFS according to β_i (d) Original epochs location (indicated by 'O') , (e) positive zero crossings from resampled ZFFS (indicated by 'R'), (f) modified epochs location (indicated by 'M'), (g) the mapped modified epochs location showing that the epoch intervals are repeated at the initial regions of the segment and some of the epoch intervals are deleted towards the final region of the segment, and (h) the corresponding dynamic duration modified segment.

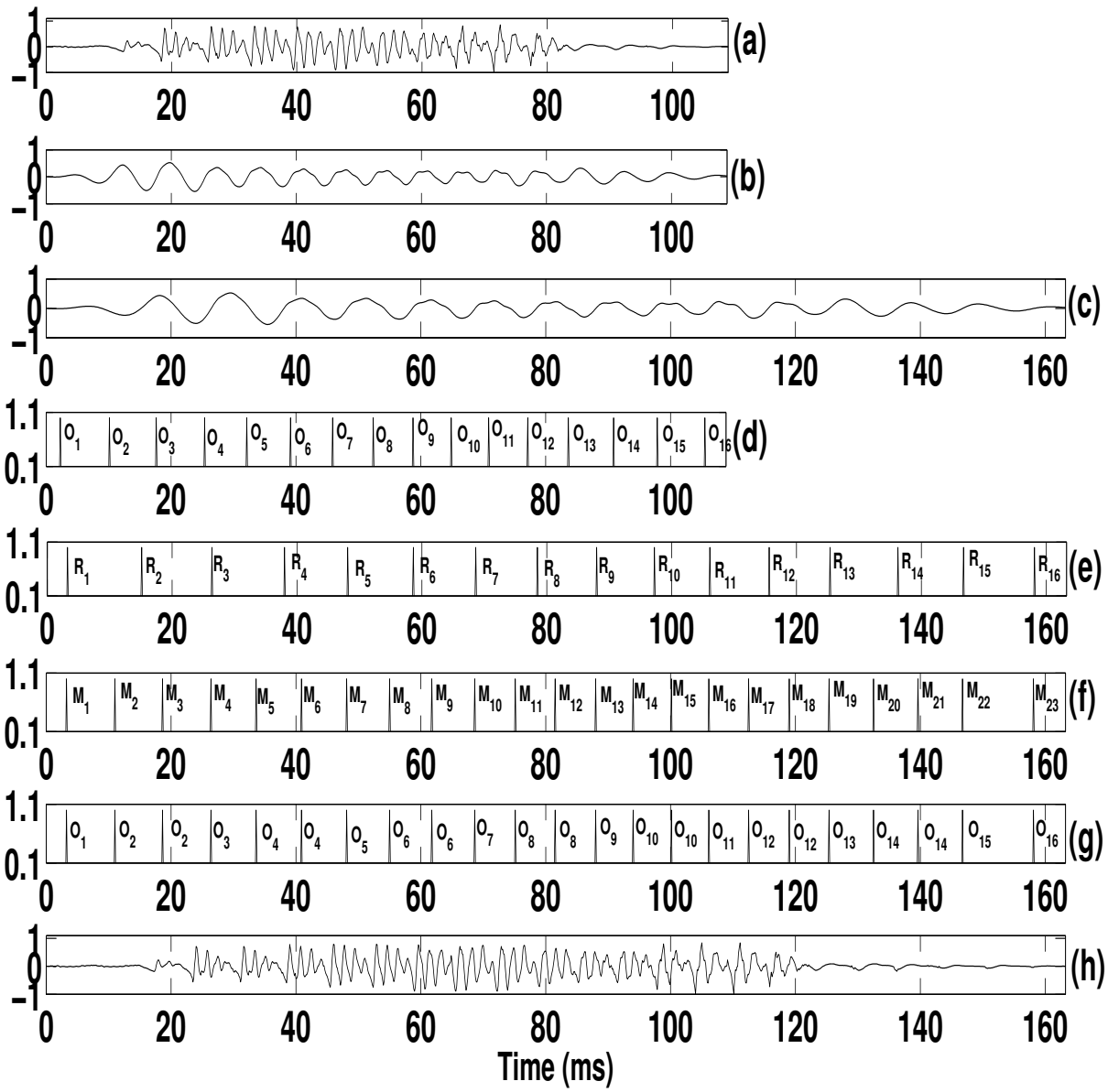

Figure 4.3: Demonstrating static duration modification as a special case of dynamic duration modification.(a) Voiced speech segment from the original speech, (b)corresponding ZFFS segment, (c) resampled ZFFS according to $\beta_i = 1.5$, (d) Original epochs location from original ZFFS segment (indicated by 'O'), (e) positive zero crossings from resampled ZFFS (indicated by 'R'), (f) modified epochs location (indicated by 'M'), (g) mapped modified epochs location showing the repetition of original epoch intervals and (h) the corresponding duration modified speech segment according to β_i.

For finding the epochs in the voiced regions, GA regions are defined. A method for finding the GA regions in the speech signal using ZFFS is described in [90]. The GA regions can be computed from the strength of excitation of the epochs, which is defined as the slope of the ZFFS around each epoch [90]. Figures 4.4(a) to (c) show the original speech, modified ZFFS using $\beta_i = 1.5$ for all i, and the strength of excitation of epochs derived from the resampled ZFFS, respectively. It can be observed that the GA regions have larger strength of excitation. These GA regions can be defined using a threshold on the strength of excitation. 30% of the average strength of excitation values, experimentally verified for different databases, is used as the threshold for defining the GA regions [118]. The GA regions derived from the strength of excitation are shown as dotted lines in the Figure 4.4(b). Then the final duration modified speech is generated by retaining the duration modified speech within the GA regions and replacing the samples in the non-GA regions from the original speech signal. The steps can be summarized as follows:

(i) Perform dynamic duration modification for all the epochs in both GA and non-GA regions.

(ii) Glottal activity regions are derived from the resampled ZFFS.

(iii) Retain the duration modified speech in the GA regions.

(iv) Replace the samples in the non-GA regions from the original speech signal.

Figure 4.4(d) plots the duration modified speech obtained using all the epochs, both from GA and non-GA regions. By comparing Figures 4.4(a) and (d), the length of duration modified speech using all the epochs is almost β_i times the length of original speech. Figure 4.4(e) shows the duration modification using GA detection. By comparing Figures 4.4(a) and (e), it can be observed that the durations of the non-GA regions of (a) (region around 0.6 sec of original speech) remains same as that of the duration modified speech (region around 0.75 sec of (e)) after GA detection. Because of this non-uniform duration modification, the overall length of the duration modified speech is less than β_i times the length of the original speech.

Figure 4.5 shows the spectrograms of original speech, duration modified speech using all the epochs and using GA detection, corresponding to the waveforms shown in Figures 4.4(a), (d) and (e), respectively. The overall spectro-temporal characteristics of duration modified speech follows that of the original speech, indicating no spectral distortions during the duration modification. In case of duration modification using GA detection, since there is no duration modification in the non-GA regions

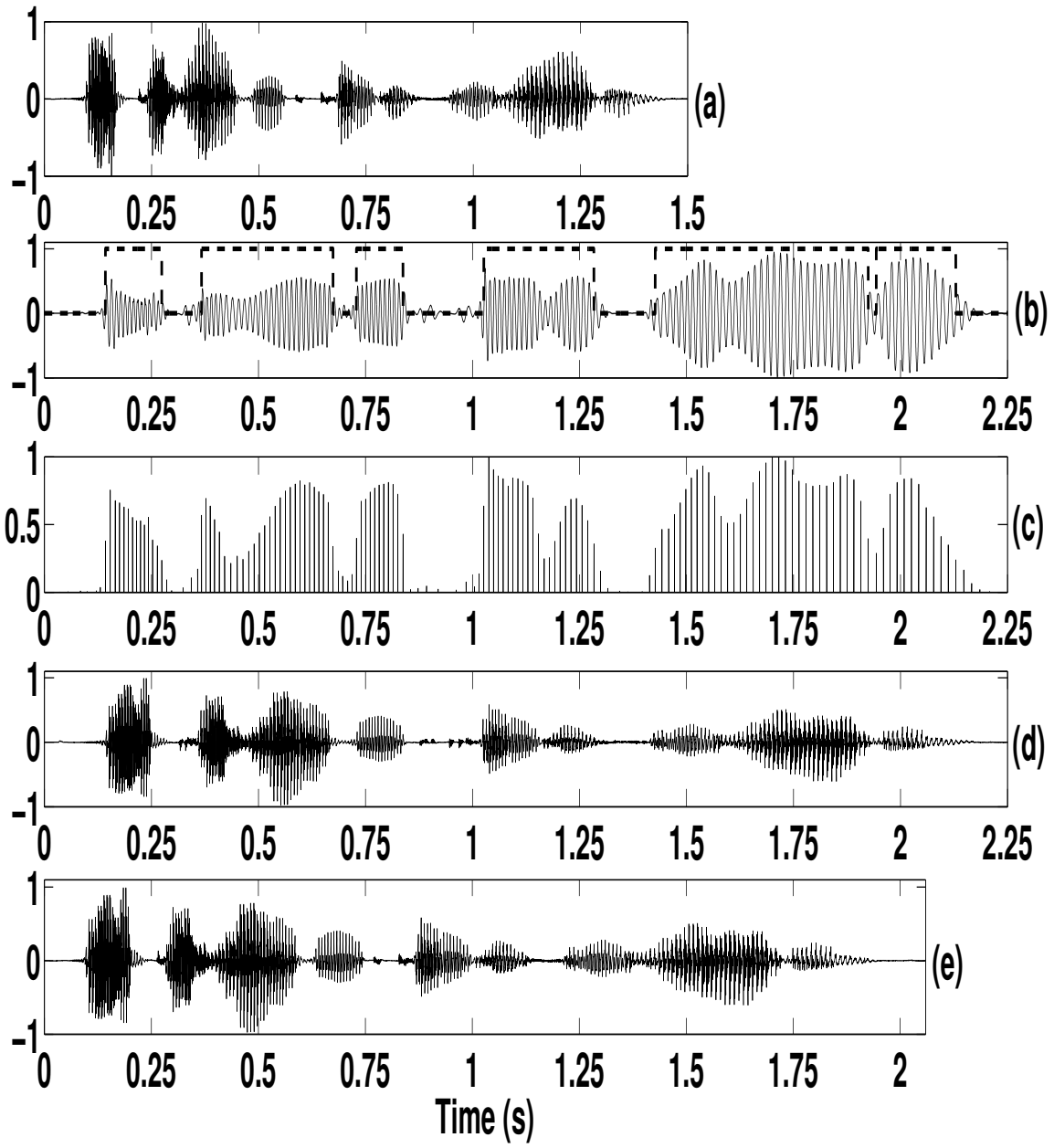

Figure 4.4: Demonstrating the duration modification for β_i=1.5. (a) Original speech, (b) modified ZFFS and GA regions (dashed lines), (c) strength of excitation derived from modified ZFFS , (d) duration modified using all the epochs and (e) duration modification after GA detection.

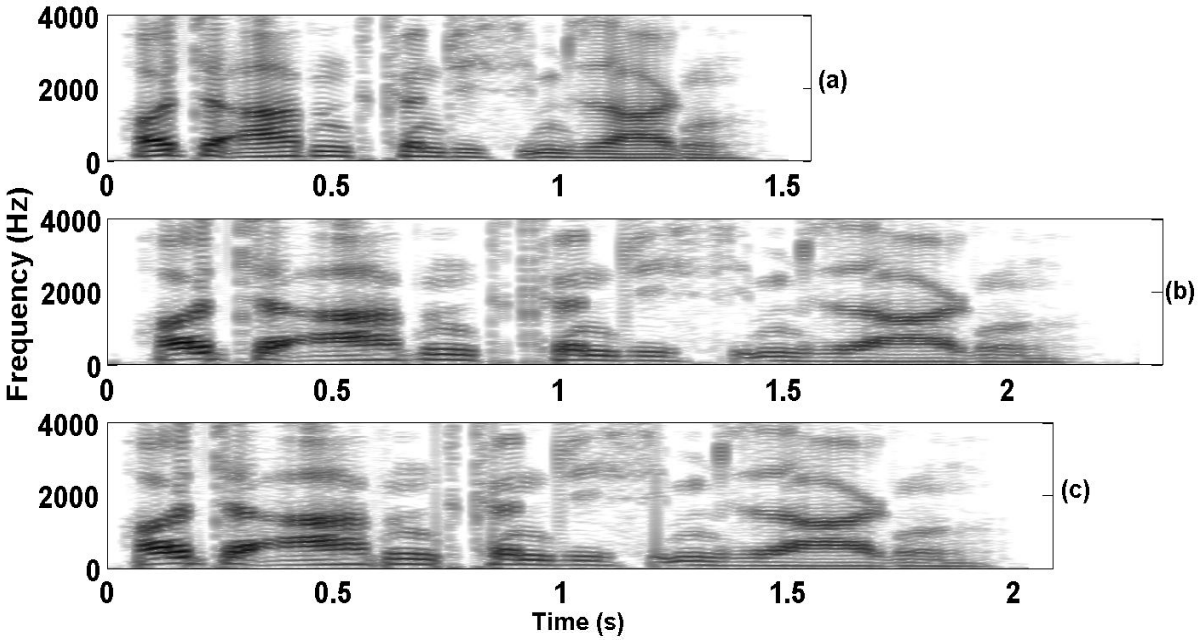

Figure 4.5: The Spectrograms of (a) original speech, the (b) duration modified speech with β_i=1.5 using all the epochs and (c) after GA detection.

(region around 0.75 sec), the spectro-temporal characteristics of these regions remain same as that of the original speech (region around 0.6 sec).

Figure 4.6(c) plots the waveform and spectrogram of dynamic duration modified speech by varying β_i from 3.0 to 0.5. Different regions are changed by different duration modifications that can be observed by comparing with the waveform of static modification given in Figure 4.6(b). For instance, the first GA region in Figure 4.6(c) is elongated in duration and the last GA region is compressed in duration. Where as, all the regions are uniformly duration modified in Figure 4.6(b). The duration of the non-GA regions are kept intact as in the original speech and hence their spectral characteristics remain same as in the original speech. This can be observed by comparing the region around 0.5 sec in original and dynamic duration modified speech spectrograms.

4.4.2 Dynamic Pitch Modification

In the pitch modification, the pitch periods are either reduced or increased according to the desired pitch modification factor. In dynamic pitch modification, such a variation is performed in a dynamic fashion. The pitch modification can also be achieved by reconstructing the modified ZFFS according to the dynamic pitch modification factor (α_i). The α_i represents the pitch modification factor for the

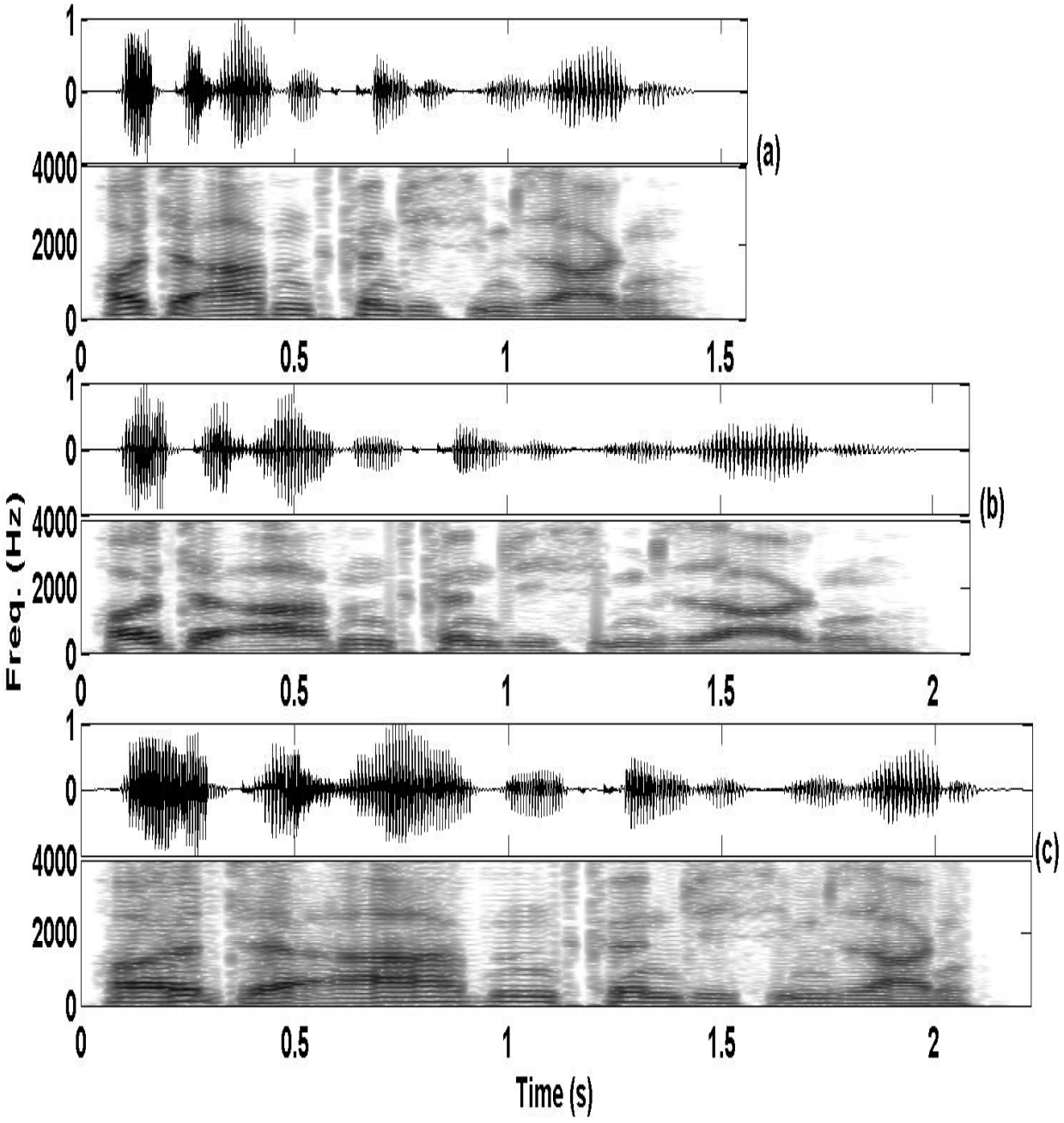

Figure 4.6: Dynamic duration modification. (a) Original speech signal and its spectrogram, (b) the static duration modified speech by a factor $\beta_i = 1.5$ and its spectrogram and (c) Dynamic duration modified speech with duration modification factor β_i varying dynamically from 3.0 to 0.5 and its spectrogram.

signal samples present in the epoch interval starting from the i^{th} epoch. The steps in the proposed dynamic pitch modification are as follows:

(i) Find the epochs location using the ZFF method.

(ii) For the i^{th} epoch, find the positive integers P_i and Q_i for α_i such that $\alpha_i = P_i/Q_i$.

(iii) The interval between every P_i original epochs location starting from i^{th} epoch is divided into Q_i equal intervals.

(iv) The begin of each of these intervals is identified as the *modified epoch location*.

(v) For each of the modified epoch location, the nearest original signal epoch location is found out.

(vi) *In case of decrease in pitch period*, all the speech samples in the original pitch period starting from the original epoch location are copied to the nearest modified epoch location in an overlap-add manner [99].

(vii) *In case of increase in pitch period*, all the speech samples in the original pitch period starting from the original epoch location are copied to the nearest modified epoch location and 10% of pitch period samples from the tail end are extrapolated according to the modified pitch period .

(viii) The steps 2 to 7 are repeated for the remaining epochs starting from $i + P_i^{th}$ epoch.

(ix) The final result is the *dynamic pitch modified signal* by processing all the epochs present in both GA and non-GA regions.

(x) GA regions are derived from the original ZFFS.

(xi) The original speech samples are copied to the non-GA regions to obtain the final dynamic pitch modified speech signal.

Figure 4.7 shows (a) segment of voiced speech, (b) its ZFFS, (c) original epochs location, (c) modified epochs location derived according to α_i varied from 0.6 to 1.5 and (e) the sequence of original epochs location that are close to the modified epochs location. In dynamic pitch modification demonstrated here, the pitch modification factor is varied from 0.6 to 1.5. The modified epoch locations for the first epoch ($i = 1$) of the given speech segment according to pitch modification factor $\alpha_i = 0.6$ ($P_i/Q_i = 3/5$) are obtained by inserting 5 (Q_i) equal intervals between 1^{st} and $4^{th}(P_i + 1^{th})$ original

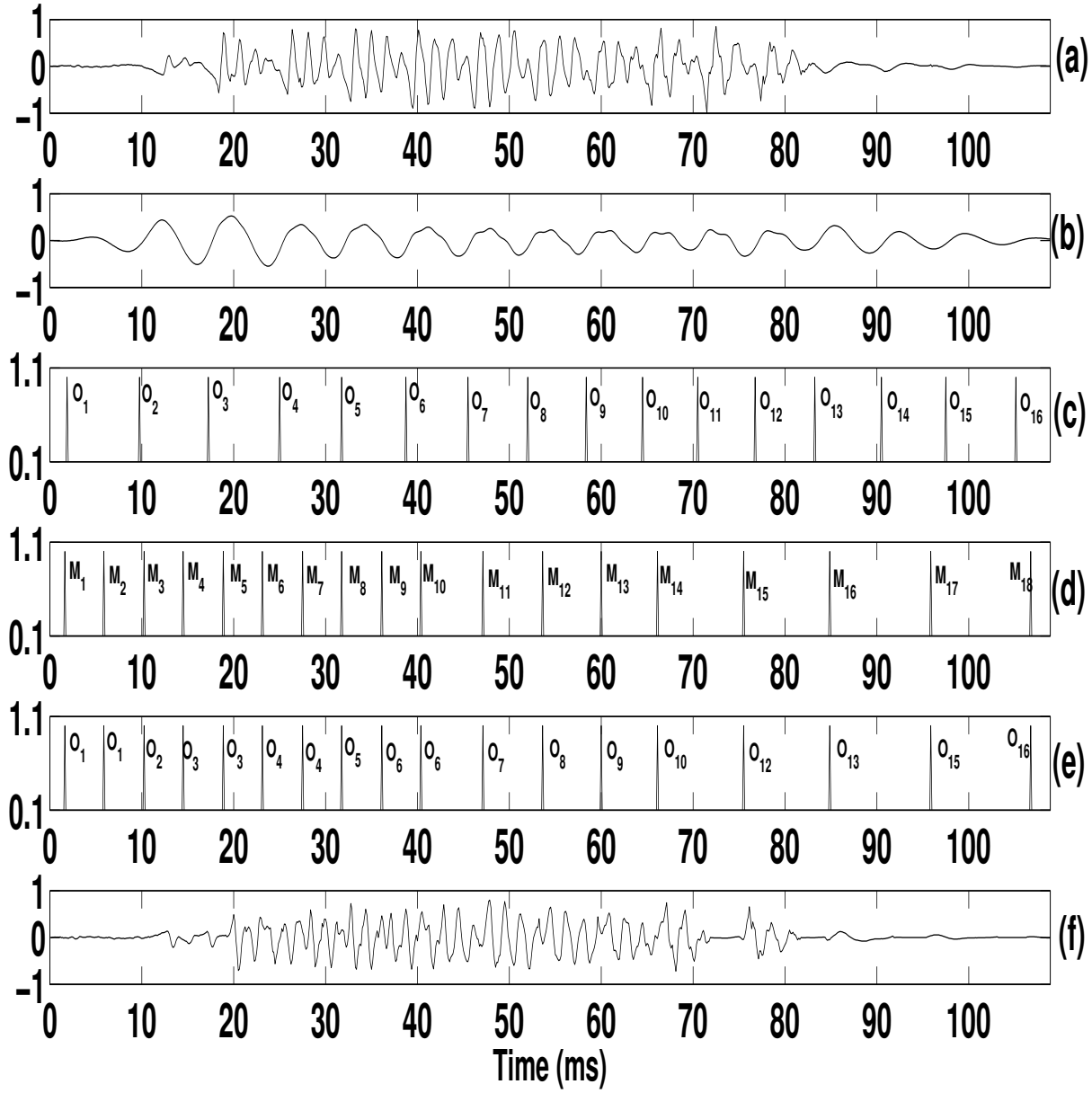

Figure 4.7: Deriving modified epochs location for dynamic pitch modification with α_i varying from 0.6 to 1.5. (a) A voiced segment of original speech, (b) ZFFS from the voiced segment of original speech, (c) Original epochs location from the positive zero crossings of original ZFFS segment (indicated by 'O') , (d) modified epochs location (indicated by 'M'), (e) the mapped modified epochs location shows the sequence of original epoch intervals that are near to the modified epoch intervals and (f) the dynamic pitch modified speech segment.

epochs location. Then the pitch modification factor corresponding to the 4^{th} epoch location ($i = 4$) is retrieved next. According to this pitch modification factor, again Q_i equal intervals are inserted between 4^{th} and $(4 + P_i)^{th}$ epochs location. This process is repeated until the last epoch. After deriving the modified epochs location, the original epoch location that are nearest to the modified epochs location are noted. Figure 4.7(f) shows the dynamic pitch modified speech segment obtained by using the original epochs give in Figure 4.7(e). Figure 4.8 shows the spectrograms of original speech,

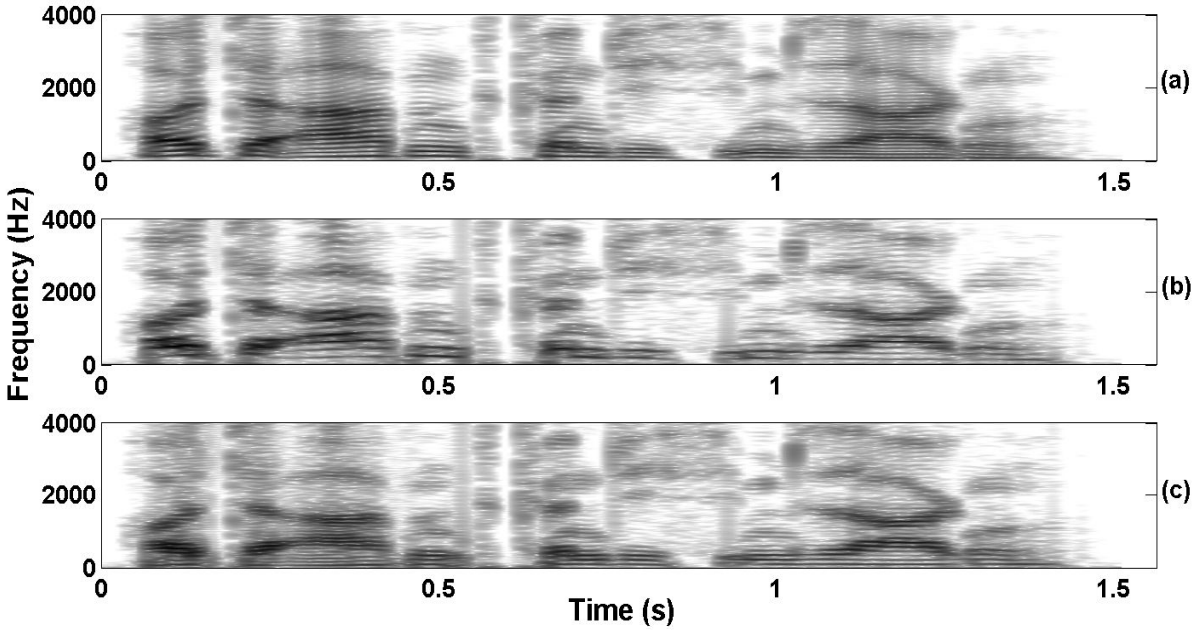

Figure 4.8: Dynamic pitch modification. The (a) spectrogram of the original speech signal, (b) spectrogram of the static pitch modified speech by a factor $\alpha_i = 0.6$ and the (c) spectrogram of the dynamic pitch modified speech with pitch modification factor α_i varying dynamically from 0.6 to 1.5.

pitch modified speech by a static modification factor, α_i=0.6 for all i and dynamic pitch modification by varying α_i from 0.6 to 1.5, respectively. The evidence of the dynamic pitch modification can be observed from the Figure 4.8(c). The interval between pitch and its harmonics is large in the beginning and reduces towards the end. Where as a constant upward shift $1/\alpha_i = 1/0.6$ times the original pitch can be observed from the spectrogram of static pitch modified speech.

4.4.3 Dynamic Excitation Strength Modification

In this work, the strength of excitation is defined as the amplitude of the impulse occurring around the epochs. The strength of excitation is measured using the slope of the ZFFS around the epochs as described in [90]. The strength of excitation is also an emotion specific parameter and hence its

modification is essential for effective neutral to emotional speech conversion . The dynamic excitation strength modification is achieved by scaling the speech samples in the epoch intervals according to the dynamic modification factor γ_i. The γ_i represents the excitation strength modification factor for the signal samples present in the epoch interval starting from the i^{th} epoch. The steps involved in the dynamic strength modification are as follows:

(i) Find the epochs location using the ZFF method.

(ii) The speech samples for the epoch interval starting from i^{th} epoch are scaled according to γ_i and copied to a new array.

(iii) The above step is repeated for all the epochs.

(iv) The final result is the *dynamic excitation strength modified speech signal*.

Figure 4.9 demonstrates the dynamic excitation strength modification with strength modification factor γ_i varying from 1 to 0.1. Figure 4.9(b) shows the strength of excitation at the epochs location for the speech signal given in Figure 4.9(a). Figure 4.9(c) shows the strength modified speech signal. Figure 4.9(d) plots the strength of excitation at the epochs locations derived from the dynamic strength modified speech signal and is found to be scaled according to γ_i. The Figures 4.9(e) and (f) show the spectrograms of the original speech and dynamic strength modified speech. The strength modification effect is reflected as the reduction in the intensity of gray scale values in the spectrogram as compared to the original speech spectrogram.

4.4.4 Dynamic duration, pitch and strength modification

We discussed about modifying duration, pitch and strength in an independent manner. However, the combined dynamic duration, pitch and strength modification is required for the neutral to target emotion conversion. This can be achieved as follows.

(i) Derive the epochs from the speech signal by ZFF method.

(ii) Derive the duration modified epoch locations according to β_i.

(iii) Derive the pitch and duration modified epoch locations by processing duration modified epoch locations according to α_i.

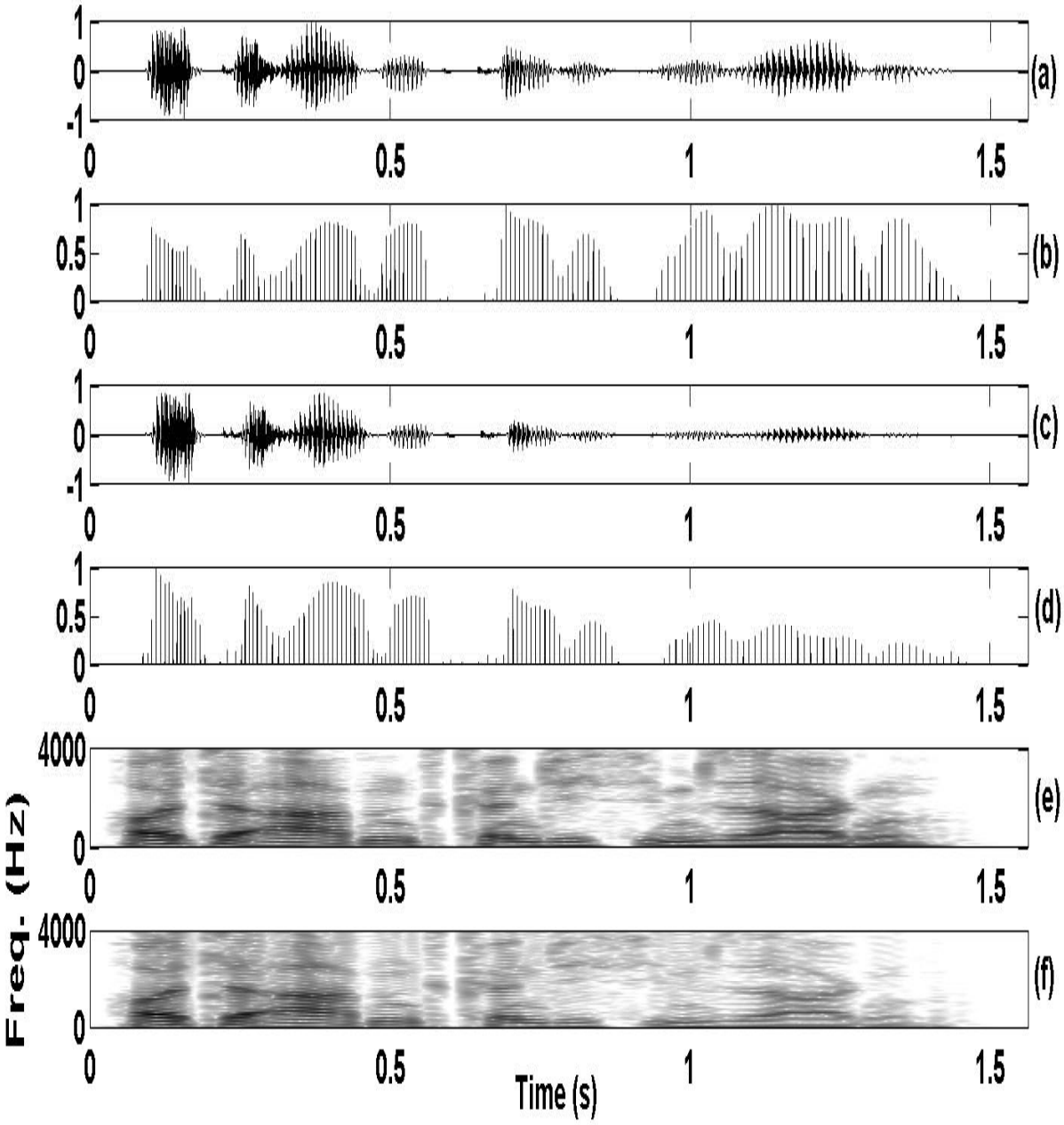

Figure 4.9: Dynamic excitation strength modification. The (a) original speech signal , its (b) strength of excitation derived from ZFFS, (c) excitation strength modified speech signal with γ_i ranging from 1 to 0.1, (d) the modified strength of excitation obtained from ZFFS of strength modified speech, (e) spectrogram of the original speech and (f) the spectrogram of the strength modified speech.

(iv) While deriving the speech signal for the duration and pitch modified epoch locations, the speech samples in the epoch interval of the i^{th} epoch are scaled according to γ_i.

(v) The result is the dynamic duration, pitch and strength modified speech.

4.5 Experimental Results and Discussions

The effectiveness of the proposed dynamic prosody modification method is evaluated by subjective studies. The subjective evaluations are carried out in three different ways, namely,

- Subjective evaluation for studying the effectiveness of GA regions
- Subjective evaluation for static prosody modification
- Subjective evaluation for dynamic prosody modification

4.5.1 Significance of GA detection for duration modification

The significance of GA detection for duration modification is studied by subjective evaluation. Four phonetically balanced utterances from 3 speakers (2 males and a female) of Arctic database are selected for the subjective evaluation. The files initially sampled at 32 kHz are down sampled to 8 kHz and used for the evaluation. The duration modified files are synthesized for various static duration modification factors using the proposed dynamic duration modification methods with and without GA detection. The synthesized duration modified speech files from both the cases are paired with the original file and used for the evaluation for each duration modification factor. The file names of the synthesized files are coded to avoid biasing of the subjects towards a particular method. 15 research scholars in the speech lab participated in the subjective evaluation. The subjects were instructed to give their opinion scores based on the perceptual quality (naturalness and intelligibility) with respect to the original file. The description for each of the scores are given in Table 4.7.

Table 4.8 shows the Mean Opinion Scores (MOS) obtained for each of the modification factors. For moderate modification factors like 0.7 and 1.5 the MOS scores obtained are nearly same. A significant improvement in the MOS scores for the proposed duration modification by GA detection compared to the case without using GA detection. This infers that only GA regions can be processed for prosody modification. The duration modified files can be accessed from the following link, *http://www.iitg.ac.in/eee/emstlab/demos/demo6.php.*

Table 4.7: Ranking used for judging the distortion of the speech signal for different modification factors.

Rating	Speech Quality	Justification for the ranking
1	Unsatisfactory	Very annoying and objectionable
2	Poor	Annoying but not objectionable
3	Fair	Perceptible and slightly annoying
4	Good	Just perceptible but not annoying
5	Excellent	Imperceptible

Table 4.8: Mean opinion scores for different duration and pitch modification factors.

Method	0.5	0.7	1.5	2.5
All Epochs	3.12	4.15	3.9	2.75
GA Region Epochs	3.93	4.45	4.1	3.55

4.5.2 Subjective evaluation of static prosody modification

The utterances used for the subjective study are same as before. The proposed dynamic prosody modification is initially evaluated for three different static duration and pitch modification factors. The proposed prosody modification method using ZFFS (ZFFS-SM) is compared with the speech files synthesized using TD-PSOLA, ZFF based LP residual prosody modification (ZFF-RM), ZFF based fast prosody modification (FAST), and using the epochs obtained by ESPS (ESPS). For all these methods, the epochs estimated using ZFF method are used as the pitch marks except for the speech synthesized using ESPS. In ESPS, the epochs estimated using RAPT algorithm which is available in ESPS toolkit is used. For all the methods the prosody modification is performed only in the GA regions obtained by ZFF method. The subjects participated in this study also remained same as before. Here the subjects were asked to rate the files according to the quality and distortion present in the file. A pilot test was given to all the subjects before the evaluation. Table 4.7 describes the significance of each of the scores used for the evaluation. The mean opinion scores (MOS) obtained in this subjective study are given in Table 4.9.

As we can observe from the Table 4.9, the MOS scores obtained for ZFFS-SM is comparable with the existing methods like TD-PSOLA, ZFF-RM and FAST. Because of the accuracy of the ZFF epochs compared to that obtained using ESPS, the speech synthesized using ZFFS-SM gets better scores compared to that of ESPS. For the duration modification, the MOS scores of ZFFS-SM, TD-PSOLA and FAST are comparable where as for pitch modification due to the truncation of the pitch cycle FAST gave lower MOS scores compared to ZFFS in case of decreasing pitch period. Since the waveform gen-

Table 4.9: Mean opinion scores for different duration and pitch modification factors.

Duration Modification			
Method	0.5	1.5	2.5
ZFFS-SM	3.59	4.32	3.75
TD-PSOLA	3.64	4.28	3.64
ZFF-RM	3.17	3.82	3.17
FAST	3.23	3.98	3.52
ESPS	2.60	2.07	1.75
Pitch Modification			
	0.6	1.5	2
ZFFS-SM	4.62	4.25	4.09
TD-PSOLA	4.40	4.15	4.09
ZFF-RM	4.33	4.12	3.96
FAST	3.59	4.03	4.09
ESPS	2.54	3.17	3

eration method for increasing pitch period modification is nearly same as that of FAST, the MOS scores for increasing pitch period is nearly same for ZFFS-SM and FAST (for the scale factors 1.5 and 2 in the Table 4.9). Since the waveform generation methods for TD-PSOLA and ZFF-SM are same for decreasing pitch period, the MOS scores for ZFFS-SM and TD-PSOLA are comparable. The synthesized files can be accessed from the following link: *http://www.iitg.ac.in/eee/emstlab/demos/demo5.php*

4.5.3 Subjective evaluation of dynamic prosody modification

The proposed dynamic prosody modification ZFFS-SM is compared with the TD-PSOLA, ZFF-RM, FAST and ESPS methods for prosody modification. The dynamic prosody modification is evaluated for both dynamic duration and pitch modification. The dynamic duration modification is evaluated for the duration modification dynamically varying from 2.5 to 0.5, and dynamic pitch modification is obtained by dynamically varying the pitch modification factors from 2 to 0.6. The proposed method is used for deriving the modified epochs locations according to desired dynamic prosodic parameters for TD-PSOLA, ZFF-RM and FAST also. For all these methods the epochs estimated using ZFF method are used as the analysis pitch marks. Whereas, the pitch marks locations estimated using RAPT, are used as the analysis pitch marks for the dynamic prosody modification given in ESPS.

The coded speech files synthesized using proposed ZFFS-SM, and TDPSOLA, ZFF-RM and FAST methods are presented to the same subjects who participated in the earlier subjective studies. The subjects were asked to judge the quality and the level of distortion present in the synthesized speech

Table 4.10: Mean Opinion Scores for dynamic duration and pitch modification.

Method	MOS
Duration Modification	
ZFFS-SM	4.0
TD-PSOLA	3.90
ZFF-RM	3.18
FAST	3.75
ESPS	2.56
Pitch Modification	
ZFFS-SM	4.25
TD-PSOLA	4.20
ZFF-RM	3.90
FAST	3.30
ESPS	3.10

files. The justification for each of the ratings are given in the Table 4.7. A pilot test was provided to the subjects before conducting the subjective evaluations. The objective of the pilot test was to demonstrate the dynamic prosody modification and to show how dynamic prosody modification is different from the static level modification.

The algorithm used for modifying the epochs location remained same for the all the methods. Table 4.10 shows the MOS scores obtained for the dynamic duration and pitch modification. The same trend can be observed in case of MOS scores obtained for dynamic prosody modification. ZFFS-SM and TD PSOLA provided nearly same scores for both dynamic pitch and duration modification. Like in the static modification case, in the dynamic case also ESPS shows degraded performance compared all other methods due to deviation from the accurate epoch locations. The distortions introduced due to sudden truncation of the pitch cycles causes the lowered MOS scores for FAST compared to ZFFS-SM, TD-PSOLA and ZFF-RM.

4.6 Summary

In this chapter, an improved perceptual quality in the existing epochs based prosody modification is achieved by using the accurate epochs location estimated from ZFF method. Also a computationally fast prosody modification is achieved by performing the prosody modification directly on the speech waveform. Then a dynamic prosody modification method is proposed using zero frequency filtered signal obtained during the epoch extraction using ZFF method. The dynamic prosody modification method developed in this work make use of the availability of ZFFS in the following ways for prosody

modification:

- Positive zero crossings of ZFFS for accurate estimate of the epochs locations
- The zero crossings of the resampled ZFFS used as the reference locations for prosody modification
- GA regions can be found based on the strength of excitation computed from the resampled ZFFS

The GA detection used in the duration modification was found to improve the naturalness of the duration modification than modifying the duration of all the epoch intervals without GA detection. The proposed method to generate synthesis pitch marks can also be used for achieving the dynamic prosody modification using epoch and PSOLA based techniques. The next chapter describes the neutral to expressive speech conversion using epoch based dynamic prosody modification.

5

Dynamic Prosody Modification for Neutral to Expressive Speech Conversion

Contents

5.1 Objective of Neutral to Emotion Conversion

The objective of the work presented in this chapter is to demonstrate the effectiveness of dynamic prosody modification in neutral to emotion conversion. In this work, the significance of dynamic prosody modification in neutral to emotion conversion is presented at various levels. The effectiveness of dynamic prosody modification is initially demonstrated for text dependant and speaker dependent emotion conversion system, where the dynamic prosody modification factors are derived directly from the available emotion speech in the target emotion. The experiments to study the effect of the dynamic prosody modification is then extended to text dependent and speaker independent case, where the dynamic prosody modification factors required for the neutral to emotion conversion are stored as patterns correspond to each GA region across various speakers in the database. Finally, the effectiveness of dynamic prosody modification is demonstrated for German emotion speech database for text independent and speaker independent case by comparing with the target emotion speech synthesized by the gross level prosody modification. The effectiveness of the dynamic prosody modification for neutral to emotion conversion in each level is confirmed by waveforms, spectrograms and subjective evaluations of the synthesized emotion speech using dynamic prosody modification.

5.2 Introduction

The neutral to emotion conversion deals with incorporating emotion specific parameters in neutral speech. The neutral to emotion conversion is achieved by modifying the emotion specific parameters of the neutral speech according to the desired emotion. The neutral to emotion conversion can be used as the back end post processing module of expressive speech synthesis system, where neutral to emotion conversion module converts the synthesized neutral speech from the text to the required emotion. The modification factors required for the neutral to emotion conversion are derived by analyzing different emotion speech databases. Most of the works presented in the literature achieve neutral to emotion conversion by modifying the emotion specific prosody parameters by fixed scaling factors across whole neutral speech utterance [8, 21]. However, the emotion specific parameters vary dynamically for various emotions. From the analysis of emotion speech databases, the instantaneous F_0, duration and strength of excitation are considered as the emotion specific parameters. Figure 5.1 plots the waveform, instantaneous pitch contour and strength of excitation of neutral and angry emotion speech utterances. The dynamic variations in the instantaneous F_0 of neutral and angry

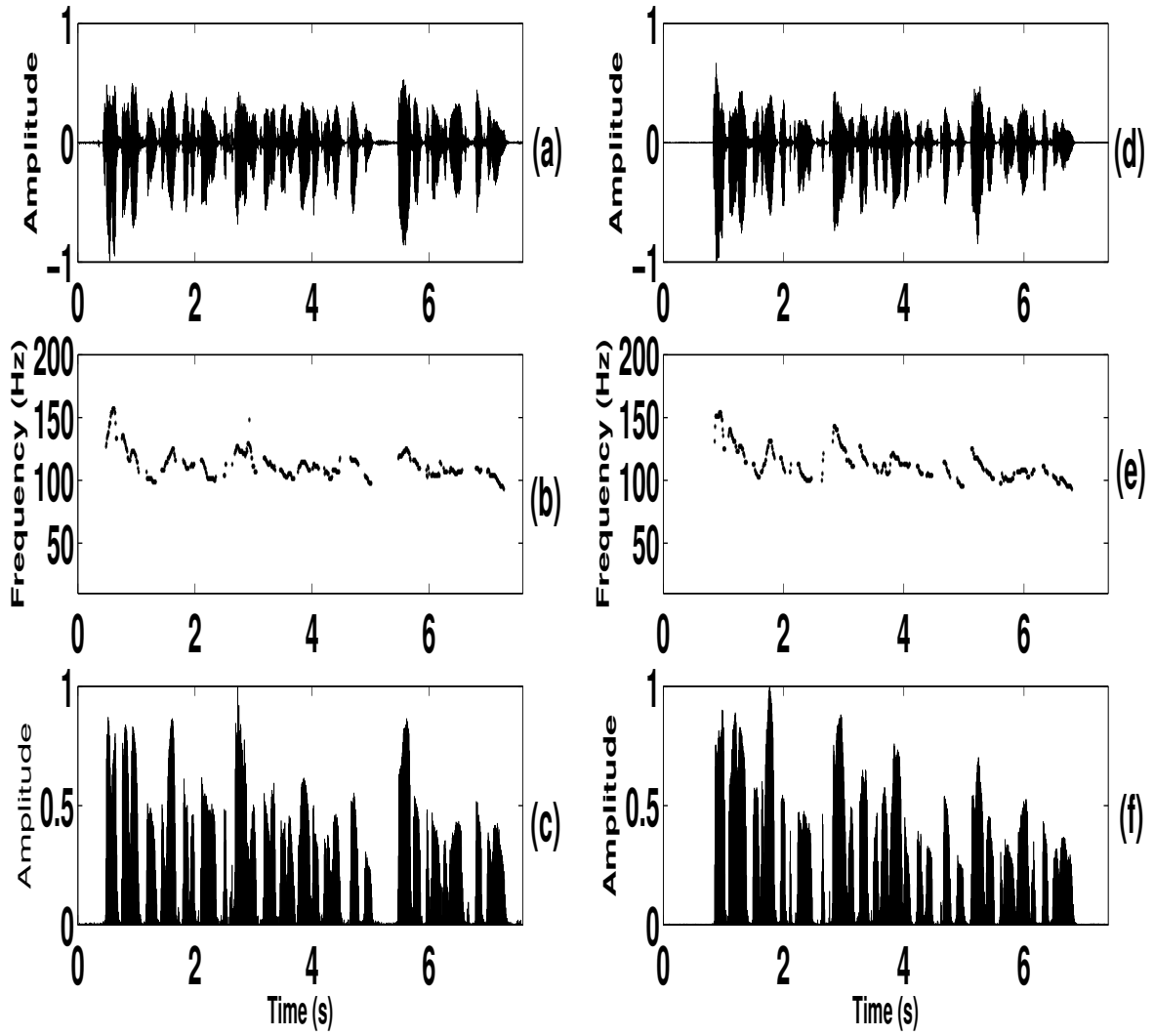

Figure 5.1: Dynamic variations in emotion specific prosodic parameters : The waveform, pitch contour and strength of excitation of neutral ((a)-(c)) and target angry emotion ((d)-(f)).

emotion speech utterances can be observed by comparing the Figures 5.1 (b) and 5.1(e) at various regions of the contours. The same dynamic variations in the strength of excitation also can be observed for the same emotions by comparing the corresponding regions in Figures 5.1 (c) and 5.1(f). Hence the incorporation of dynamic variations in these emotion specific parameters is essential for effective neutral to emotion conversion. The work presented in this chapter describes the significance of dynamic prosody modification for incorporating dynamic variations in instantaneous F_0, duration and strength of excitation parameters for neutral to emotion conversion. The effectiveness of dynamic prosody modification in emotion conversion is illustrated for the following cases:

- Text dependent and speaker dependent case
- Text dependent and speaker independent case
- Text independent and speaker independent case

A straight forward way to demonstrate the effectiveness of dynamic prosody modification is the emotion conversion for text dependent and speaker dependent case. Here the dynamic prosody modification factors required for the emotion conversion are derived directly from the glottal activity (GA) regions of target emotion speech. These dynamic prosody modification factors obtained for each GA region are used for modifying the prosodic parameters of the corresponding GA region of the neutral speech. Alternatively, the effectiveness of the dynamic prosody modification is illustrated by deriving the dynamic prosody modification factors from each syllable like unit of the target emotion speech. Here, the target emotion speech is synthesized by dynamically modifying the prosody parameters of syllable like units of neutral speech according to corresponding dynamic prosody modification factors. The emotion conversion is also achieved by dynamically modifying the prosodic parameters of each phoneme unit of the neutral speech according to that of the target emotion.

For demonstrating the effectiveness of dynamic prosody modification in text dependent and speaker independent case, the dynamic prosody modification factors corresponding to each GA region are derived by computing the average across the speakers and stored as dynamic prosody modification patterns in the database. For the emotion conversion, the prosodic parameters of the GA regions of the neutral speech are modified according to the dynamic prosody modification factors stored as patterns for the corresponding GA region.

Dynamic prosody modification is demonstrated for text independent and speaker independent neutral to emotion conversion where the dynamic prosody modification factors are derived for initial, middle and final regions of the utterances in the emotion speech database. The target emotion speech is synthesized by dynamically modifying the prosody parameters in initial, middle and final regions of the neutral speech. The effectiveness of the dynamic prosody modification is demonstrated by subjective comparisons with the text independent and speaker independent neutral to emotion conversion system by static prosody modification of neutral speech.

The rest of the chapter is organized as follows: Section 5.3 describes the effectiveness of dynamic prosody modification in text dependent and speaker dependent neutral to emotion conversion system. Section 5.4 describes the usefulness of dynamic prosody modification in text dependent and speaker independent neutral to emotion conversion system. The effectiveness of dynamic prosody modification in text independent and speaker independent neutral to emotion conversion is described in Section 5.5. The subjective evaluation for each case is given in Section 5.6. Finally Section 5.7 summarizes the works done for demonstrating the significance of dynamic prosody modification for neutral to emotion conversion.

5.3 Text Dependent and Speaker Dependent Neutral to Emotion Conversion

To test the effect of dynamic prosody modification in neutral to emotion conversion, we use three databases. The following subsection describes the information about the databases used for the neutral to emotion conversion

5.3.1 Databases

5.3.1.1 German Emotion Speech Database [74]

The speech of five different emotions (Neutral, Angry, Happy, Boredom and Fear) across 9 texts of 8 speakers (5 females and 3 males) from German emotion speech database are used for the present work. The emotions are elicited by professional artists and the utterances are segmented at the syllable level. The segmentation at the syllable-like unit level is carried out manually using electro-glottogram (EGG) and spectrograms of the recorded emotion speech utterances. All the utterances in the database are sampled at 16 kHz with 16 bit resolution per sample.

5.3.1.2 CSTR emotion speech database [36]

The CSTR emotion speech database consists of three emotions (Angry, Happy and Neutral) recorded in UK English. 400 sentences are used for recording each emotion of two speakers (1 male and 1 female). The female speaker in the database is a professional actress who worked in various TV shows and the other speaker is not a professionally trained actor. The sentences used for the recording are selected from UK newspapers and the sentences are selected for optimum phonetic coverage. The emotion speech is recorded at 16 kHz sampling rate and 16 bits per sample resolution.

5.3.1.3 Hindi emotion speech database

Hindi emotion database is prepared by recording four simulated emotions (Angry, Happy, Neutral and Boredom) of four speakers (2 males and 2 females). 10 Hindi sentences randomly selected from Hindi broadcast news database are used for the recording. The speech for each emotion is collected from every speaker in three sessions such that there are 3 examples for each emotion utterance for a given speaker. The speakers used for the recordings are students of the speech laboratory. The speech and EGG are simultaneously collected using an electro-glottograph for each speech utterance.

5.3.2 Hindi text dependent and speaker dependent neutral to emotion conversion

In text dependent and speaker dependent case, the F_0, duration and strength of excitation of each glottal activity (GA) region of target emotion are imposed on the corresponding GA regions of the neutral speech. The dynamic scaling factors are derived by scaling all the F_0 and strength of excitation values in the GA regions of target emotion with that of the neutral GA region. As the target and neutral GA regions have different lengths and have different number of F_0 and strength values, the interpolation of F_0 and strength values in each GA region is done. The interpolated F_0 and strength values in the GA regions are then resampled to match the length of the neutral and target emotion GA regions before deriving the dynamic scaling factors.

The dynamic pitch and strength of excitation scaling factors are derived by dividing the interpolated and resampled F_0, and strength of excitation contours of target and neutral GA regions. The duration scaling factors are derived by scaling the length of GA regions of target emotion with the corresponding GA regions of the neutral. The prosody parameters of the neutral speech signals are dynamically modified according to the dynamic prosody modification factors. Table 5.1 gives the prosody modification factors of different speakers (*Speaker-1* to *Speaker-4*). The dynamic scaling

Table 5.1: The dynamic prosody scaling factors derived for neutral to **angry** emotion conversion for 3 speakers from each of the GA regions.

Speaker	GA_1	GA_2	GA_3	GA_4	GA_5	GA_6	GA_7	Average
	Dynamic Pitch Modification Factors							Fixed Pitch Modification Factor
Speaker 1	1.6±0.3	1.7±0.1	1.7±0.2	1.5±0.2	1.4±0.6	1.9±0.2	1.3±0.3	1.6
Speaker 2	1.2±0.2	1.1±0.3	1.2±.04	1.3±0.1	1.1±0.1	1.6±0.2	1.2±0.3	1.2
Speaker 3	1.1±0.2	1.4±0.4	1.2±0.1	1.2±0.1	1.3±0.1	1.3±0.1	1.3±0.2	1.3
Average	1.3±0.2	1.4±0.3	1.4±0.1	1.3±0.1	1.3±0.3	1.6±0.2	1.3±0.3	1.4
Speaker 4	1.0±0.3	1.1±0.1	1.4±0.1	1.4±0.1	1.1±0.2	1.2±0.1	1.0±0.1	1.2
	Dynamic Duration Modification Factors							Fixed Duration Modification Factor
Speaker 1	1.0	0.7	0.8	0.7	0.6	0.7	0.7	0.7
Speaker 2	1.1	0.9	1.0	0.8	0.9	0.9	0.9	0.9
Speaker 3	0.8	0.8	0.7	0.9	0.9	0.8	0.7	0.8
Average	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Speaker 4	1.4	1.0	0.6	0.7	1.4	0.9	0.7	1.0
	Dynamic Strength Modification Factors							Fixed Strength Modification Factor
Speaker 1	0.8±0.3	1.0±0.4	1.2±.5	1.5±0.2	1.4±0.6	1.6±0.7	0.6±0.7	1.2
Speaker 2	0.9±0.6	0.8±0.9	0.9±.2	1.1±0.6	0.9±0.6	1.3±0.4	0.6±0.2	0.9
Speaker 3	1.1±0.4	1.2±0.6	1.8±0.7	1.1±0.8	1.3±0.3	1.5±0.4	0.9±0.7	1.3
Average	0.9±0.4	1.0±0.6	1.3±0.5	1.2±0.5	1.2±0.5	1.5±0.5	0.7±0.5	1.1
Speaker 4	0.9±0.5	1.1±0.5	1.1±0.5	1.2±0.7	1.5±0.9	1.2±0.4	1.1± 0.7	1.2

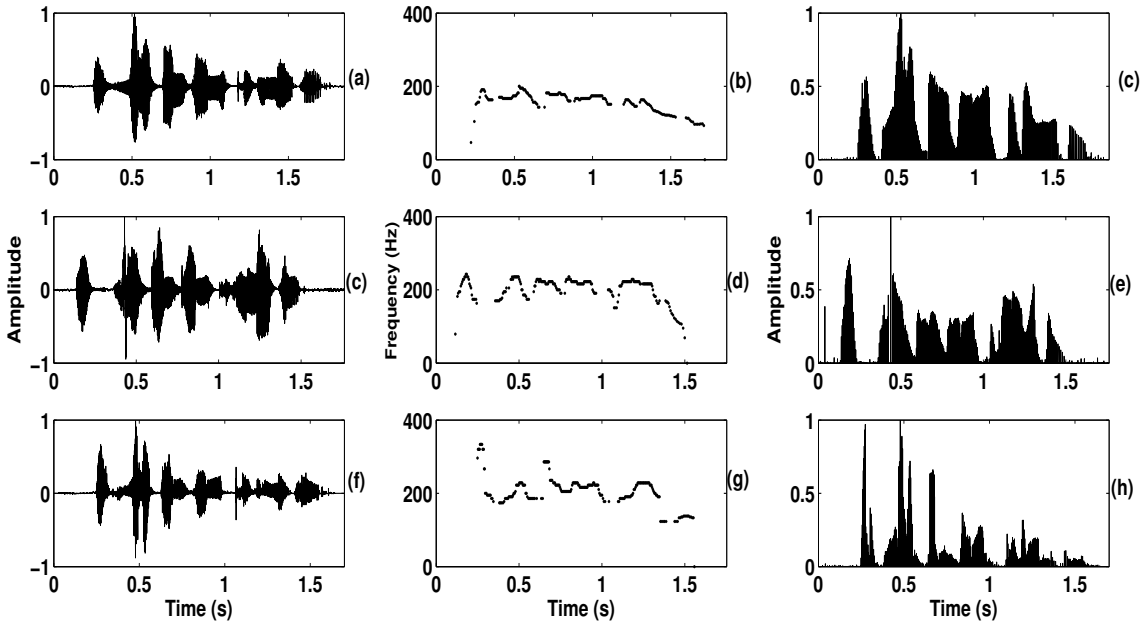

Figure 5.2: Text dependent and speaker dependent emotion conversion by dynamic prosody modification: The waveform, pitch period contour and strength of excitation of neutral ((a)-(c)), target angry emotion ((d)-(f)) and synthesized angry ((g)-(i)) emotion using prosodic parameters of the target emotion .

factors corresponding to instantaneous F_0 and strength of excitation of each GA region of angry expression are represented by the mean and variance values in the Table 5.1. Whereas, the dynamic duration modification factors are derived from the mean duration of each GA region. For instance, the dynamic pitch modification factors derived for the first GA region (GA_1) has a mean of 1.6 and a variance of 0.3. This means the pitch modification factors in GA_1 vary dynamically within 1.6±0.3. Figure 5.2 shows the dynamically prosody modified neutral speech according to the target emotion speech of *Speaker-4*. The pitch contour and strength of excitation plots of the synthesized speech are significantly different from that of neutral speech and also resemble to the target emotion.

5.3.3 Text dependent and speaker dependent neutral to emotion conversion in German emotion speech database

This section describes the text dependent and speaker dependent neutral to emotion conversion using dynamic prosody modification of syllable like units in German emotion speech database. Here the F_0, duration and strength of excitation of syllable like units are dynamically modified to synthesize the expressions. Each emotion utterance in German emotion speech database was labeled at the syllable level. In order to study the effectiveness of the dynamic prosody modification, the F_0, duration and strength of excitation from the syllable like units of target expressions are imposed on the same syllables of neutral speech by dynamic prosody modification. Figure 5.3 shows the waveform, pitch contour, excitation strength and spectrogram of the synthesized target emotion in this case. The instantaneous pitch and strength contours of the synthesized target emotion match closely with that of the reference target emotion. The narrowband spectrograms also indicate that the spectral distributions of the synthesized expressions are different from that of the neutral speech and confirm that the spectral distribution of the pitch and harmonic structure is not merely a scaled version of neutral speech.

5.3.4 Text dependent and speaker dependent neutral to emotion conversion in CSTR emotion speech database

To test the effectiveness of dynamic prosody modification for text dependent and speaker dependent neutral to emotion conversion in English CSTR emotion speech database, the prosodic parameters of the phonemic units of neutral speech is dynamically modified according to the target emotions. As the CSTR emotion speech database is sufficiently large, the utterances in the database are automatically segmented at the phoneme level by HMM based force alignment. The dynamic duration, pitch and strength modification factors are derived by comparing the parameters of neutral and the target

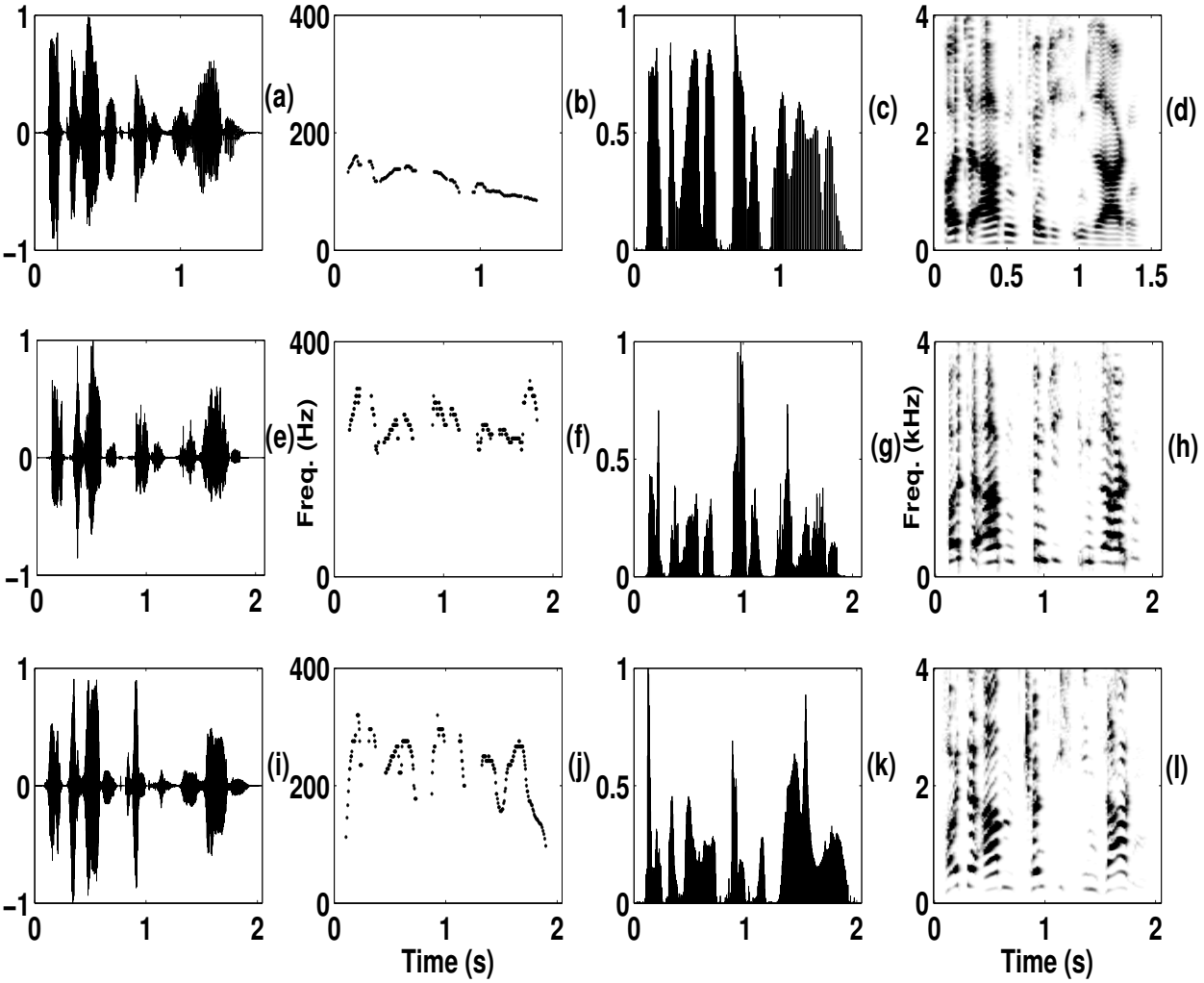

Figure 5.3: Text dependent and speaker dependent neutral to target emotion conversion by the dynamic prosody modification of syllable-like units of German emotion speech database. Speech waveform, pitch contour, excitation strength and narrow-band spectrogram of the neutral ((a)-(d)), synthesized target emotions by deriving the scale factors from the target emotion syllables ((e)-(h)), and original target angry emotion ((q)-(t)).

emotion phonemes. The duration scale factors derived for each phoneme unit is assigned to each of the epochs location of the phoneme units of the neutral speech for dynamic prosody modification. The pitch modification factors for each phoneme is derived by scaling the average F_0 of the target emotion phoneme units and the corresponding average F_0 of neutral phoneme units. Each F_0 scale factor is then assigned to the epochs location of the neutral phoneme unit. Similarly, all the epochs location of the neutral speech are assigned with the respective pitch modification factors of each phoneme. The strength modification factors for dynamic strength modification are also derived in the similar manner. The neutral speech is then dynamically prosody modified according to the these dynamic duration, pitch and strength modification factors. Figure 5.4 plots the waveform, F_0 contour and strength of excitation of original neutral, synthesized angry and original target angry emotions. The F_0 contour and strength of excitation of the synthesized emotion speech is estimated using the modified ZFF method by providing the the synthesized emotion speech as the input. We can observe from the Figure 5.4(e) that the dynamics of the F_0 contour of the synthesized angry emotion is more close to original angry emotion utterance given in Figure 5.4(h). The dynamics of the F_0 contour due to target emotion is confirmed by observing the region around 0.5 sec in Figure 5.4(e) and corresponding region in target angry emotion in Figure 5.4(h). Also the same observation can be found in strength of excitation case by comparing Figure 5.4(f) and Figure 5.4(i). Figure 5.5 plots the waveform, F_0 contour and strength of excitation of original neutral, synthesized happy and original target happy emotions. Here also dynamics due to the emotion can be confirmed by comparing the F_0 contours of the target and synthesized happy emotion utterances (Figure 5.5(h) and Figure 5.5(e)) and strength of excitation of target and synthesized happy emotions (Figure 5.5(i) and Figure 5.5(f)).

5.4 Text Dependent and Speaker Independent Neutral to Emotion Conversion

The more generic case is a speaker independent scenario. The proposed dynamic prosody modification can also be used for developing text dependent neutral to emotion conversion in speaker independent manner. The dynamic prosody modification factors derived for each GA region of each speaker as described above are averaged and used. For instance, the dynamic prosody modification factors derived for the neutral to angry emotion conversion obtained by averaging the three speakers data (*Speaker-1* to *Speaker-3*) are given in Table 5.1. The speech in the target emotion synthesized using the average dynamic prosody modification factors (obtained by the average of *Speaker-1* to

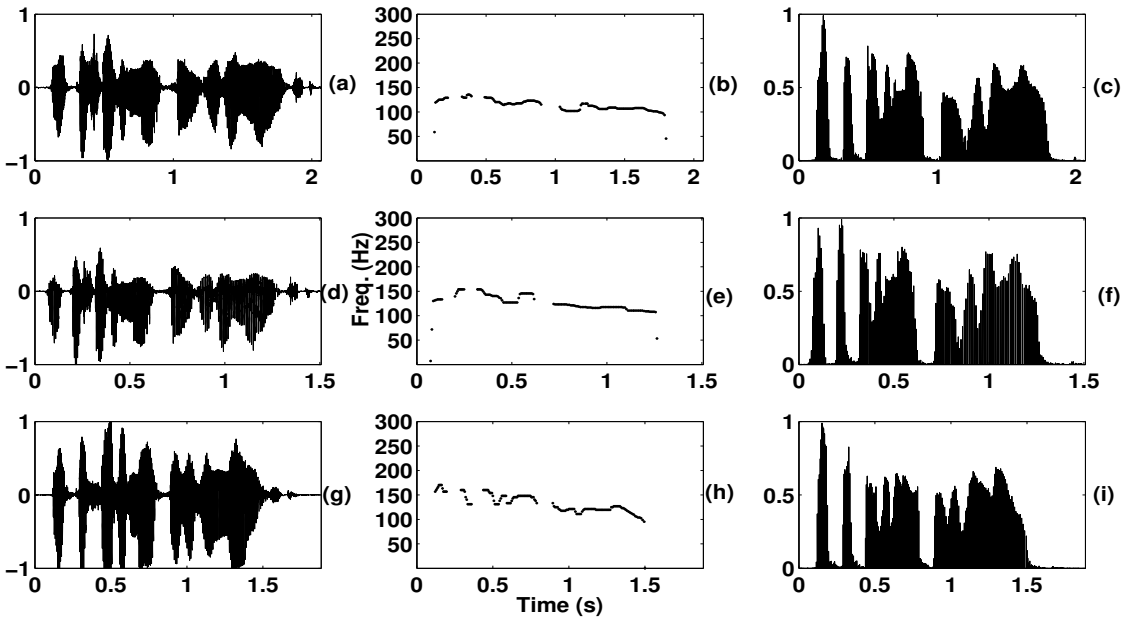

Figure 5.4: Text dependent and speaker dependent neutral to **angry** emotion conversion: Text dependent and speaker dependent neutral to target emotion conversion by the dynamic prosody modification of phonemes in CSTR emotion speech database. Speech waveform, pitch contour and excitation strength of the neutral ((a)-(c)), synthesized target angry emotion speech signals by deriving the scale factors from the target emotion phoneme ((d)-(f)), and original target angry emotion ((g)-(i)).

Speaker-3) is compared with the original target emotion of the Speaker-4. The Figure 5.6 shows the synthesized angry emotion speech of *Speaker-4* using these average factors. The pitch contour and the excitation strength plot of the synthesized angry emotion speech are significantly different to that of the neutral speech. This can be further improved using the average factors from a large number of speakers.

5.5 Text Independent and Speaker Independent Neutral to Emotion Conversion

The dynamic prosody modification for text independent and speaker independent neutral to emotion conversion is demonstrated for German emotion speech database. For demonstrating the dynamic prosody modification in the neutral to emotion conversion of German emotion speech database, the dynamic prosody modification parameters are derived from syllable like units in the initial, middle and final regions of the utterance. The text independent and speaker independent neutral to emotion conversion can also be developed by the gross level scaling of prosodic parameters of the neutral speech.

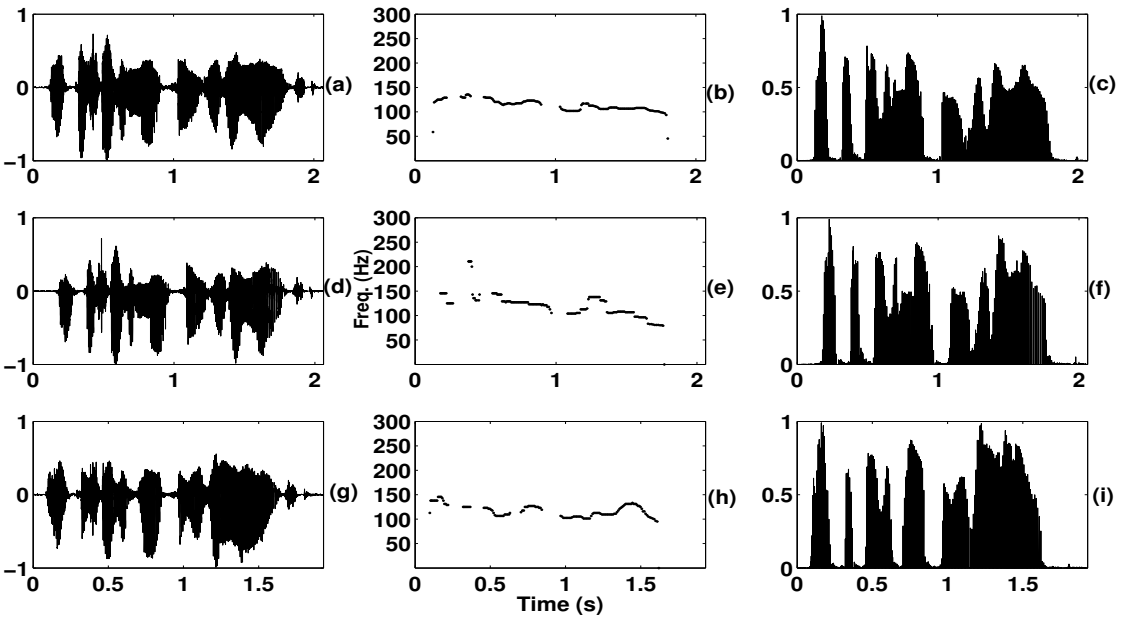

Figure 5.5: Text dependent and speaker dependent neutral to **happy** emotion conversion: Text dependent and speaker dependent neutral to target emotion conversion by the dynamic prosody modification of phonemes in CSTR emotion speech database. Speech waveform, pitch contour and excitation strength of the neutral ((a)-(c)), synthesized target happy emotion speech signals by deriving the scale factors from the target emotion phoneme units ((d)-(f)), and original target happy emotion ((g)-(i)).

The effectiveness of dynamic prosody modification over static prosody modification is demonstrated by comparison mean opinion score (CMOS) based subjective evaluation.

5.5.1 Text independent and speaker independent neutral to emotion conversion in German

A general observation is that the initial and final region syllables are relatively more affected compared to the middle region. A via media is therefore to average the effect of emotions across the initial, middle and final regions of the text. Two syllable-like units at the beginning and end of the text form the initial and final regions of the text, respectively. The syllable-like units between the initial and final regions constitute the middle region of the text. The average pitch, duration and strength of excitation modification factors are derived for each region of the neutral speech by comparing the corresponding regions of the target emotion. The F_0, duration and strength modification factors computed for 5 different emotions in 9 texts of 8 speakers are given in the Table 5.2. Variability of pitch, duration and strength modification factors across the initial, middle and final regions show the variable effect of the emotions across different regions in a given text. In order to convert neutral

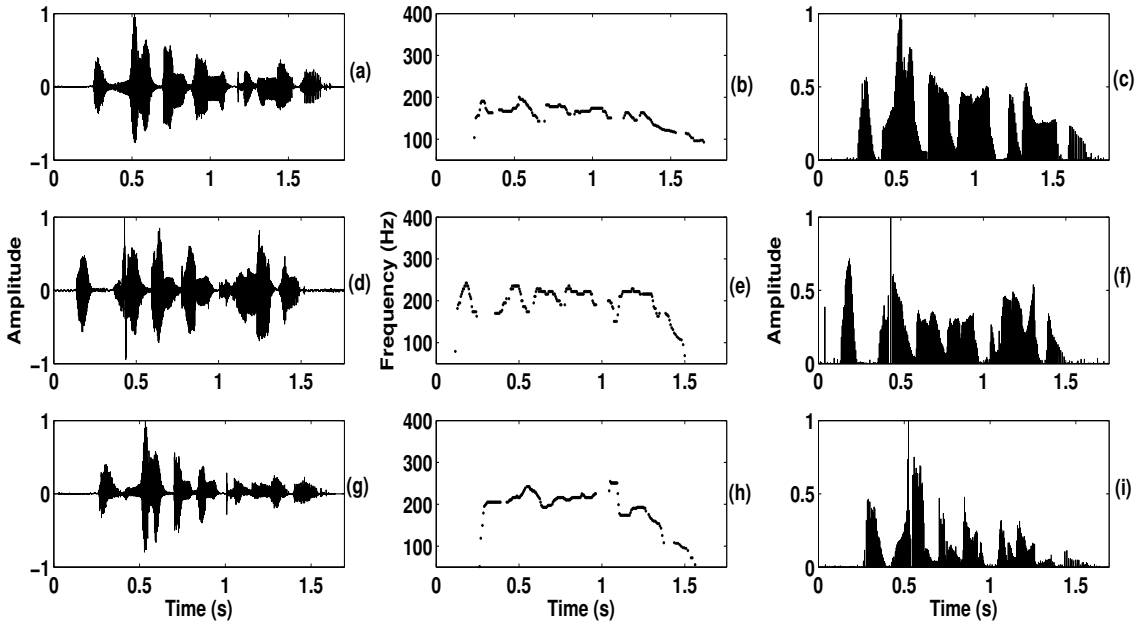

Figure 5.6: Emotion Conversion by dynamic prosody modification: The waveform, pitch period contour and strength of excitation of neutral ((a)-(c)), target angry emotion ((d)-(f)) and synthesized angry ((g)-(i)) emotion using the prosodic parameters of the target emotion.

to target emotion, the prosodic parameters of initial, middle and final regions of neutral speech are dynamically modified according to the scaling factors derived for each region. To demonstrate the effectiveness of the dynamic prosody modification, the emotion utterances synthesized using dynamic prosody modification, is compared with that synthesized by the static prosody modification. To convert the neutral to emotion speech by static prosody modification, the gross level modification of F_0 contour, strength of excitation and duration of the neutral utterances are carried out.

Table 5.2: Pitch, duration and excitation strength modification factors of initial, middle and final regions of sentences. I, M and F represents the initial, middle and final regions of the sentence, respectively.

Emotion	Pitch Mod.			Dur. Mod.			Strength Mod.		
	I	M	F	I	M	F	I	M	F
Angry	1.63	1.86	2.11	1.27	1.28	1.04	0.77	0.84	1.30
Happy	1.62	1.80	1.95	1.15	1.06	1.03	0.75	0.75	0.93
Boredom	1.19	0.94	0.98	1.24	1.25	1.19	0.99	1.04	1.13
Fear	1.38	1.53	1.83	1.11	0.98	0.90	0.67	0.84	1.33

Table 5.3: Average prosodic parameters parameters of different emotions estimated from the emotion utterances of German emotion speech database.

Emotion	Mean Pitch	Mean Dur. (s)	Mean Strength
Neutral	180.86	2.260	0.51
Angry	301.59	2.590	0.41
Happy	287.17	2.430	0.38
Boredom	175.60	2.700	0.52
Fear	249.10	2.240	0.42

Table 5.4: Pitch, duration and strength modification factors obtained by taking ratio of target emotion parameters with respect to the neutral emotion.

Target Emotion	Pitch Mod.	Dur. Mod.	Strength Mod.
Angry	1.67	1.15	0.80
Happy	1.59	1.08	0.76
Boredom	0.97	1.20	1.02
Fear	1.38	0.99	0.83

5.5.1.1 Neutral to emotion conversion by static prosody modification

The static prosody modification factors required for the gross level modification of the prosody parameters can be obtained by analyzing the different emotion speech signals taken from different speakers and texts of the German emotion speech database. We can have one modification factor for the entire sentence and use it for modification, termed as gross level modification or static prosody modification in this work. Table 5.3 presents the average instantaneous F_0, duration and sentence level duration and strength of excitation of emotion utterances from German emotion speech corpus. Table 5.4 shows the gross level scale factors for the neutral to emotion conversion by static prosody modification. For instance, in order to synthesize angry emotion, the F_0, strength and sentence level duration of neutral speech are modified by static scaling factors, 1.67, 0.80 and 1.15, respectively. For the gross level modification of prosody modification, the same dynamic prosody modification method is used where all the epochs location of the neutral speech are assigned with the same modification factor. For instance, for neutral to angry emotion conversion, all the epochs location in the neutral speech are assigned with a pitch modification factor of 1.6, strength modification factor of 0.80 and duration modification factor of 1.15.

Figure 5.7((e)-(h)) shows the relevant plots of gross level modification. Figure 5.7((i)-(l)) shows

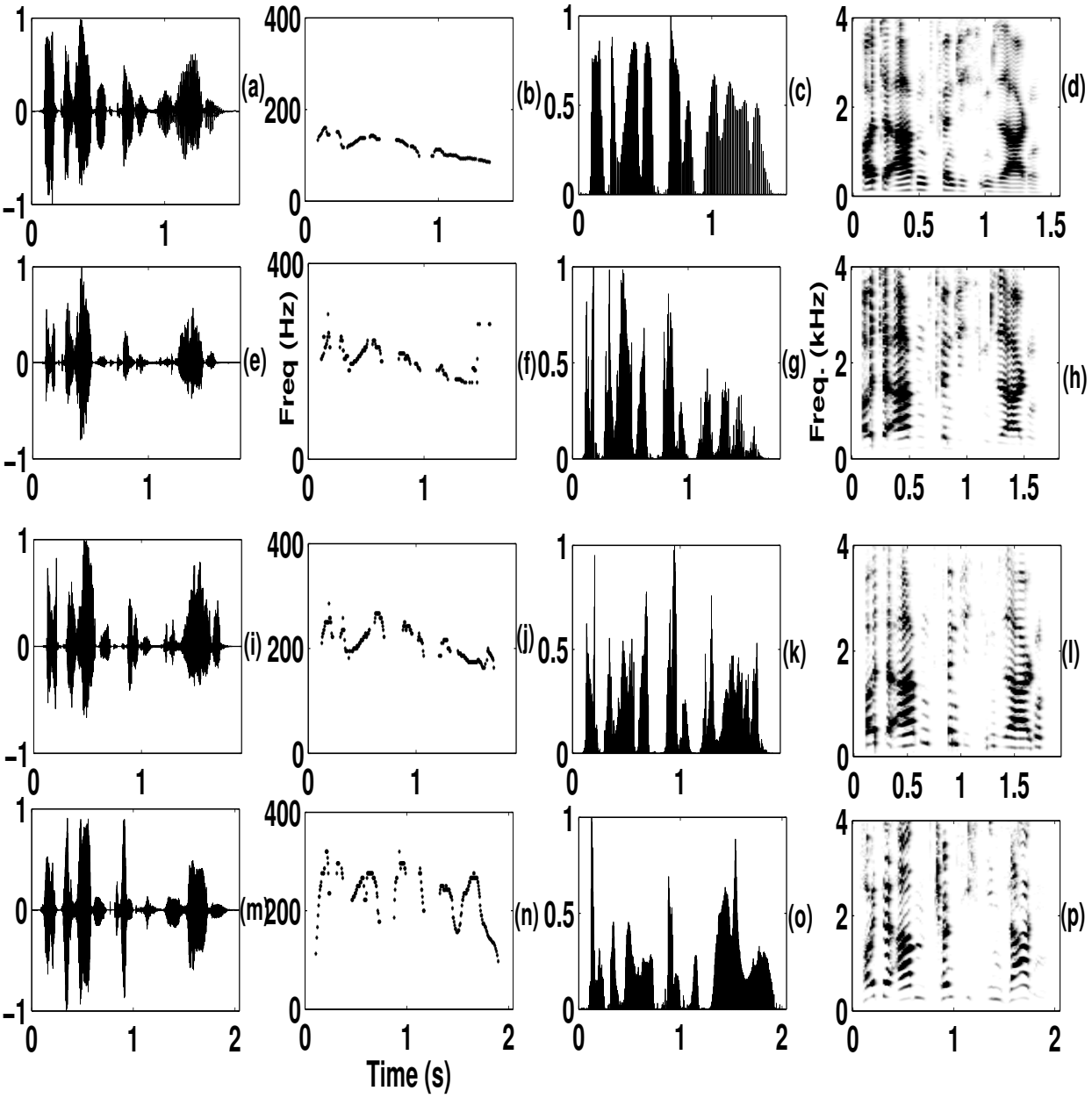

Figure 5.7: Neutral to target emotion conversion. Speech waveform, pitch contour, excitation strength and spectrogram of the neutral ((a)-(d)), by the gross level modification ((e)-(h)) and initial, middle and final region wise modification ((i)-(l)) and original target emotion ((m)-(p)).

Table 5.5: Ranking used in perceptual test to judge the similarity of the synthesized emotion with the target emotion.

Rating	Description for evaluating synthesized emotions
1	sounds exactly like neutral
2	sounds slightly different from neutral
3	sounds different from neutral
4	sounds more different from neutral
5	sounds exactly like target

the relevant plots for the dynamic prosody modification of the initial, middle and final regions. From the Figure 5.7((j)), it is to be noted that the shape of the modified F_0 contour matches more closely with the target emotion pitch contour than that of the neutral speech. Same trend can be observed in the strength of excitation plot also. Alternatively, in Figure 5.7(f), even though the range of the pitch values match with that of the target emotion, the gross trend of pitch contour matches closely to that of the neutral speech. Similar observation can be made with respect of excitation strength of gross level modification. Comparing the Figures 5.7((i)-(l)) and 5.7((e)-(h)), it may be noted that the shape of the pitch contour and strength of excitation of the target emotion are preserved better in the speech synthesized using region wise modification than the gross level modification. The spectrogram of the speech synthesized by modifying initial, middle and final region syllables of the neutral speech shows the spectral characteristics similar to that of the target emotion. The smooth transitions of spectral characteristics indicate that there are no spectral and temporal distortions present in all the three modification methods.

5.6 Subjective Evaluations

To demonstrate the effectiveness of dynamic prosody modification for neutral to emotion conversion subjectively, subjective evaluations were carried out. Here the the subjects were presented with the neutral speech, target emotion speech and synthesized emotion speech. The subjects were instructed to evaluate for the level of expressiveness or emotion content in the synthesized emotion by comparing the neutral speech and original target emotion speech. The subjects were told to provide their score in a five point scale. The significance of each score is given in Table 5.5.

5.6.1 Subjective evaluation for Hindi emotion speech database

The comparative subjective study conducted in this section is to develop the effectiveness of dynamic prosody modification over static prosody modification for text dependent and speaker independent neutral to emotion conversion. The average prosody modification factors of all the GA regions are used for static prosody modification based neutral to emotion conversion. The values of static F_0, duration and strength modification factors are given in Table 5.1. The stimuli required for subjective evaluations are synthesized from two speakers (1 male and 1 female) and for three emotions (Angry, Happy and Boredom) and for one text of the utterance from Hindi emotion speech database. A total of 24 files (2×4 ×3) are used for the subjective evaluation. The filenames of the emotion speech synthesized using and static and dynamic prosody modification were coded before presenting to the subjects. A pilot test was given to all the subjects. In the pilot test, the subjects were instructed to evaluate for the synthesized speech files by comparing the original neutral speech and original target emotion speech. A total of 15 research scholars of EMST laboratory participated in the subjective evaluations. The average value of the scores obtained for each file is then calculated as the comparison mean opinion score (CMOS). The Table 5.6 presents the CMOS obtained for emotion synthesized using static prosody modification and the dynamic prosody modification.

Table 5.6: Comparison mean opinion scores for the emotion speech synthesized by static and dynamic prosody modification

Emotion Conv. Method	CMOS		
	Angry	Happy	Boredom
Static	2.14	2.17	3.12
Dynamic	3.23	3.31	3.20

Table 5.6 shows a significant improvement in MOS scores obtained for synthesized emotions using dynamic prosody modification as compared to static prosody modification. This indicates the effectiveness of dynamic prosody modification in emotion conversion as compared to the existing techniques for emotion conversion using static prosody modification. Also a less variation in MOS scores for synthesized boredom emotion can be observed from Table 5.6. This may be due to lack of rapid variations as compared to angry and happy emotions. The synthesized files can be accessed from the following link: *http://www.iitg.ac.in/eee/emstlab/demos/demo7.php*

Table 5.7: Comparison mean opinion scores for the emotion speech synthesized by modifying parameters of each syllable, gross level and initial,middle and final regions of the neutral speech.

Method	CMOS			
	Angry	Happy	Boredom	Fear
Syllable	3.60	3.53	3.83	3.50
Static	2.78	2.42	3.44	2.53
Initial − Middle− Final	3.38	3.17	3.72	3.22

5.6.2 Subjective evaluation for German emotion speech database

The comparison subjective evaluations were performed to demonstrate the effectiveness of dynamic prosody modification for neutral to emotion conversion for German emotion speech database. Here, the effectiveness of emotion conversion by the dynamic prosody modification of syllable like units (text dependent and speaker dependent neutral to emotion conversion), dynamic prosody modification of initial, middle and final regions (text independent and speaker independent) and static prosody modification are evaluated subjectively by CMOS. The same subjects who participated for Hindi subjective evaluations, are participated in this case also. All the subjects are presented with original neutral and target emotions and synthesized emotions using the syllable, static and region wise methods. The subjects were asked to compare the synthesized emotions in the coded file with the original neutral and target files and rate them according to the descriptions given in Table 5.5. A total of 40 (5X4X2) speech files synthesized from a male and a female speakers of the German emotion speech corpus are used. The comparison mean opinion scores obtained for the each of the emotions are averaged to get the CMOS. The CMOS obtained for each emotion for all three methods are presented in Table 5.7.

It has to be observed that the emotions synthesized using region wise modification of the prosody parameters show higher CMOS than the static prosody modification of the parameters from neutral. The fear and happy emotions shows the lowest CMOS indicating the most confusable emotions with neutral. No significant difference in CMOS values are observed in case of boredom emotion. The higher CMOS obtained for prosody modification of each syllable like unit and, initial, middle and final regions reinforce the significance of dynamic prosody modification for neutral to emotion conversion over the static prosody modification of the neutral utterances. Some of the synthesized emotion speech samples are provided in the following link *http* : *//www.iitg.ac.in/eee/emstlab/demos/demo*4.*php* .

5.7 Summary

In this chapter, the effectiveness of dynamic prosody modification over the static prosody modification for neutral to emotion conversion is evaluated. The emotion speech utterances of three emotion speech databases in three languages are used for the present work. The significance of dynamic prosody modification to incorporate the prosodic parameters for text dependent and speaker dependent, text dependent and speaker independent and text independent and speaker independent neutral to emotion conversion is demonstrated in this work. The effectiveness of the dynamic prosody modification is confirmed for neutral to target emotion conversion by the subjective evaluations of synthesized emotion utterances from each emotion database. This indicates that the dynamic prosody modification predominantly incorporates the variation in the emotion specific prosodic parameters irrespective of the language in which the utterance is spoken. The comparative subjective studies show that the level of emotion information is more in the synthesized emotion speech by dynamic prosody modification as compared to the emotion speech synthesized by the static prosody modification. However, there are some distortions still present in the synthesized emotions using dynamic prosody modification. This is due to rapid variations in dynamic prosodic scale factors derived from the original target expressions. The future work should focus on eliminating this distortions introduced due to rapid variations in the dynamic scale factors.

6

Summary and Conclusions

Contents

6.1 Summary of Present Work

The objective of the present work is to demonstrate the effectiveness of the epoch based prosody modification in neutral to expression conversion for expressive speech synthesis (ESS) applications. The general approach used for the ESS is by explicitly incorporating the expression specific prosodic parameters by prosody modification. The works presented in this thesis focus mainly on the following stages of neutral to emotion conversion,

- Analysis and estimation of expressive parameters
- Development of epoch based dynamic prosody modification method for neutral to expression conversion
- Demonstrating the effectiveness of epoch based dynamic prosody modification in neutral to expression conversion

Accurate estimation of expressive parameters are essential for the expressive speech analysis. From the analysis of EGG of various emotions, the contour of the instantaneous F_0, strength of excitation and duration vary with emotions. These emotion specific parameters can be accurately computed by estimating accurate epochs location from emotional speech. As the ZFF method provides most accurate estimate of the epochs location from neutral speech signals, the conventional ZFF method is initially used for the epochs estimation from emotional speech. In ZFF method, the speech is passed through the cascade of two zero frequency resonators (ZFR) connected in cascade. A ZFFS is then derived by removing the trend in the ZFR output by the local mean substraction using average pitch period of the utterance as the window length. The zero crossing of the ZFFS are estimated to be the epochs location of the speech utterance. However, due to rapid pitch variations in emotional speech, the ZFFS of the emotional speech is found to show spurious zero crossings which in turn results in false estimation of epochs. Therefore a modified ZFF method is proposed to improve the epoch estimation performance from emotional speech. In the modified ZFF method, for trend removal of the ZFR output, the window length is updated for every 25 ms segments of ZFR output. These trend removed short segments of ZFR are further smoothed by low pass filtering using the cut off frequency as the average pitch. The modified ZFFS is then reconstructed from the trend removed short segments of ZFR. The epochs location are estimated as the positive zero crossings of the modified ZFFS. The instantaneous F_0 is computed from reciprocal of the successive epochs location and strength

of excitation at every epoch location is computed as the slope of the modified ZFFS around each epoch location. The sentence level duration characteristics are then computed from each emotion as the duration parameter. The German emotional speech database with five emotions (neutral, angry, happy, boredom and fear) and Hindi emotional database with four emotions (neutral, angry, happy and boredom) are considered for the analysis. According to the expressive speech analysis, the angry expressions found to have highest F_{0Avg} and boredom found to possess the lowest F_{0Avg} in both the databases. The average standard deviation of the F_0 contour indicates the variations in the F_0 contour. The highest average standard deviation obtained for angry emotion indicate the higher F_0 variation in the contour as compared to other emotions. In case of strength of excitation, the boredom and neutral emotions found to have highest average strength of excitation and angry and happy emotions found to have lowest average strength of excitation. In case of duration characteristics, the boredom emotion utterances have the highest duration compared to all other emotions considered.

Epoch based prosody modification is used for incorporating the variations in the F_0, duration characteristics and strength of excitation for effective neutral to emotion conversion. An improved perceptual quality in the existing epochs based prosody modification is achieved by using the accurate epochs location estimated from ZFF method instead of the group delay (GD) based epochs. The improved epoch estimation performance of the ZFF method as compared to GD method is the reason for the improved perceptual quality of the prosody modified speech. Since the algorithmic steps to estimate the epochs location in ZFF method is computationally cheap as compared to GD method, the use of ZFF based epochs also makes the epoch based prosody modification computationally fast. The computational complexity in the epoch based prosody modification is further reduced by performing the prosody modification directly on the speech waveform instead of the LP residual modification used in the existing epoch based prosody modification. The subjective evaluations showed an improvement in the perceptual quality of the prosody modified speech as compared to the existing prosody modification for static pitch and duration modification factors. As the prosodic parameters of various expressions vary dynamically, it is required to dynamically incorporate these dynamic variations for effective expressive conversion. Hence a dynamic prosody modification method is proposed using zero frequency filtered signal obtained during the epoch extraction using ZFF method. The dynamic prosody modification method developed in this work make use of the availability of ZFFS in the following ways for prosody modification:

- Positive zero crossings of ZFFS for accurate estimate of the epochs location
- The zero crossings of the resampled ZFFS used as the reference locations for prosody modification
- GA regions can be found based on the strength of excitation computed from the resampled ZFFS

The GA detection used in the duration modification was found to improve the naturalness of the duration modification than modifying the duration of all the epoch intervals without GA detection. The proposed method to generate synthesis pitch marks can also be used for achieving the dynamic prosody modification using epoch and PSOLA based techniques. The subjective evaluations show improved perceptual quality for the proposed dynamic prosody modification using ZFFS compared to PSOLA and existing PSOLA based techniques for both static and dynamic modification factors.

After developing method to dynamically incorporate prosodic parameters, the next step is to demonstrate the effectiveness of dynamic prosody modification for neutral to expression conversion. The effectiveness of the dynamic prosody modification is demonstrated for the following cases:

- Text dependent and speaker dependent case
- Text dependent and speaker independent case
- Text independent and speaker independent case

For demonstrating the effectiveness of dynamic prosody modification in text dependent and speaker dependent scenario, the prosodic parameters of the GA regions, syllable like units and phonemes are varied according to that of the target emotion speech of the same text and the same speaker. In the text dependent and speaker independent case, the prosodic patterns for each GA region is averaged across various speakers for the same text and stored in the database for each emotion. During the conversion, the prosodic patterns of each GA region corresponding to target emotion is retrieved from the database and dynamic prosody modification is performed on the corresponding GA regions of the neutral speech. For text independent and speaker independent neutral to emotion conversion, the average prosodic parameters corresponding to syllables in the initial, middle and final region are computed for all the emotion speech utterances and the neutral to emotion conversion is achieved by the dynamic prosody modification according to the modification factors in initial, middle and final regions. The effectiveness of the dynamic prosody modification in neutral to emotion conversion is confirmed from the waveforms and spectrogram plots and comparative subjective evaluations with the original neutral and target emotions.

6.2 Contributions of the present work

The contributions of the work reported in this thesis for the neutral to emotion conversion using epoch based prosody modification include,

- Identified instantaneous F_0, strength of excitation and duration as the expression specific parameters from the analysis of EGG of various emotions
- Identified a significant degradation in the epoch estimation performance for emotional speech using conventional ZFF epoch estimation method
- Proposed a modified ZFF method for improving the epochs estimation performance in emotional speech
- Epochs estimated using ZFF method are used as the analysis pitch marks in the existing epoch based prosody modification for improving perceptual quality of the prosody modified speech
- Computationally fast epoch based prosody modification method is proposed by performing prosody modification directly on the speech waveform
- General framework for deriving synthesis pitch marks for dynamic prosody modification is proposed using ZFFS
- The epoch based dynamic prosody modification using ZFFS is used for dynamic modification of pitch, duration and strength of excitation
- Performing the duration modification only in the GA regions found to improve the perceptual quality of the duration modification
- A text dependent and speaker dependent neutral to emotion conversion is demonstrated by dynamic prosody modification of GA regions, syllables and phonemes of the neutral speech
- A text dependent and speaker independent neutral to emotion conversion is demonstrated by storing the average prosodic parameters across various speakers for each GA region and for each emotion
- A text independent neutral to emotion conversion is demonstrated by dynamically modifying F_0, duration and strength of excitation of *initial*, *middle* and *final* syllable regions of neutral speech

- The effectiveness of the neutral to emotion conversion using dynamic prosody modification compared to the static prosody modification is demonstrated from the subjective evaluations.

6.3 Scope for future work

- To improve the accuracy for the estimation of expressive parameters like instantaneous F_0 and strength of excitation, we have proposed a modified ZFF method for extracting epochs location from emotional speech. Even though the proposed modified ZFF method provides improved epoch estimation performance for various emotions, the performance is not at par with that obtained for the neutral speech signals. Hence further refinements in the epoch estimation are required for the applications of converting one emotion to another emotion.

- Apart from prosodic parameters, methods have to be developed for the analysis and incorporation of vocal tract (VT) parameters for the effective neutral to emotion conversion

- In the proposed fast prosody modification, for extreme modification factors, other speech parameters such as spectral transitions and changes in loudness in different segments need to be incorporated to reduce perceptual distortion.

- There seems to be some reverberation in the proposed and other waveform modification methods for large prosody modification factors. Methods have to be developed to avoid these reverberation effects for extreme modification factors.

- Even though the effectiveness of the dynamic prosody modification is demonstrated by imposing the dynamic prosodic features of the target emotional speech, the challenging task will be to model these dynamic prosodic features for text independent emotion conversion task.

- In order to model the emotions, the prosody parameters of each emotion has to be analyzed for each sound unit from a large database and use it for neutral to emotion speech conversion using dynamic prosody modification

- Statistical modeling of F_0 and duration have to be done for each emotion and then the dynamic prosody modification has to be used for converting the neutral speech to target emotion speech according to the statistically predicted F_0 contours and duration parameters.

Bibliography

[1] K. R. Scherer, "Vocal affect expressions: A review and a model for future research," *Psychol. Bull.*, vol. 99, pp. 143–165, 1986.

[2] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text to speech synthesis system for American English," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, pp. 1099–1109, July 2006.

[3] J. P. Cabral, "Transforming prosody and voice quality to generate emotions in speech," Master's thesis, L2F -Spoken Language Systems Lab, Lisbon, Portugal, April 2006.

[4] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Commun.*, vol. 49, pp. 317–330, 2007.

[5] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, Sept. 2002.

[6] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G.Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Magazine*, vol. 18, no. 1, pp. 32–80, Jan 2001.

[7] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 1145–1154, Jul. 2006.

[8] J. P. Cabral and L. C. Oliveira, "Emo voice: a system to generate emotions in speech," in *Proc. INTERSPEECH*, 2006, pp. 1798–1801.

[9] C. E. Williams and K. Stevens, "Emotions and speech: Some acoustic correlates," *J. Acoust. Soc. Am.*, vol. 52, pp. 1238– 1250, 1972.

[10] J. E. Cahn, "Generation of affect in synthesized speech," in *Proc. American Voice I/O Society*, 1989, pp. 1–19.

[11] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 972–980, May 2006.

[12] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech and Language Process.*, vol. 16, no. 8, pp. 1602–1614, Nov. 2008.

[13] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech and Language Process.*, vol. 17, no. 4, pp. 614–625, May 2009.

[14] M. Bulut and S. Narayanan, "On the robustness of overall f0 only modifications to the perception of emotions in speech," *J. Acoust. Soc. Am.*, vol. 123, pp. 4547–4558, 2008.

[15] J. Vroomen, R. Collier, and S. J. L. Mozziconacci, "Duration and intonation in emotional speech," in *Proc. EUROSPEECH*, 1993, pp. 577–580.

[16] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoust. Sci & Tech.*, vol. 26, no. 4, pp. 317–325, 2005.

[17] N. Campell, W. Hamza, H. Hog, and J. Tao, "Editorial special section on expressive speech synthesis," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, pp. 1097–1098, July 2006.

[18] D. H. Klatt, "Review of text to speech conversion for English," *J. Acoust. Soc. Am.*, vol. 82, pp. 737–793, 1987.

[19] ——, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, pp. 971–995, 1980.

[20] S. King, "An introduction to statistical parametric speech synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, 2011.

[21] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for story telling applications," *IEEE Trans Audio, Speech and Language Proc.*, vol. 14(4), pp. 1099–1108, July 2006.

[22] J. Pittermann, H. Meng, and W. Minker, "Towards an emotion-sensitive spoken dialogue system - classification and dialogue modeling," in *Proc. IET Int. Conf. Intelligent Environments*, 2007.

[23] M. Schroder, "Expressive speech synthesis: Past, present and possible futures," *Affective Information Processing, Springer*, vol. 2, pp. 111–126, 2009.

[24] I. R. Murray and J. L. Arnott, "Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Am.*, vol. 93, pp. 1097–1108, 1993.

[25] F. Burkhardt and W. F. Sendilmeier, "Verification of acousical correlates of emotional speech using formant synthesis," in *Proc. ISCA Workshop on speech & emotion*, 2000, pp. 151–156.

[26] J. M. Montero, J. Gutierrez-Arriola, J. Colas, E. Enriquez, and J. M. PArdo, "Analsysis and modelling of emotional speech in spanish," in *Proc. ICPhS*, 1999, pp. 671–674.

[27] M. Schrder, "Emotional speech synthesis- a review," in *Proc. Eurospeech*, 2001, pp. 561–564.

[28] G. Fairbanks and L. W. Hoaglin, "An experimental study of pitch characteristics of voice during the expression of emotion," *Speech Monographs*, vol. 6, pp. pp. 87–104, 1939.

[29] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion by rule in synthetic speech," *Speech Commun.*, vol. 16, pp. 369–390, 1995.

[30] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Proc. INTERSPEECH*, 2006.

[31] C. Drioli, G. Tisato, P. Cosi, and F. Tesser, "Emotions and voice quality: Experiments with sinusoidal modeling," in *proc. ITRW VOQUAL03*, 2003, pp. 127–132.

[32] S. P. Whiteside, "Simulated emotions: An acoustic study of voice and perturbation measures," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. pp. 699–703.

[33] E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri, "Towards emotional speech synthesis: a rule based approach," in *proc. ISCA SSW5*, 2004, pp. 219–222.

[34] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for hmm-based speech synthesis," in *Proc. Eurospeech*, 2003, pp. 2461–2464.

[35] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "A speech synthesis system for assisting communications," in *ISCA Workshop on Speech & Emotion*, 2000, pp. 167–172.

[36] G. Hofer, K. Richmond, and R. Clark., "Informed blending of databases for emotional speech synthesis," in *Proc. INTERSPEECH*, 2005.

[37] R. Fernandez and B. Ramabhadran, "Automatic exploration of corpus specificproperties for expressive text-to-speech: A case study in emphasis," in *Proc. ISCA Workshop on Speech Synthesis*, 2007, pp. 34–39.

[38] N. Campbell, "Conversational speech synthesis and the need for some laughter," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 17, no. 4, pp. 1171–1179, July 2006.

[39] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control techniques for hmm-based speech synthesis," in *Proc. ICSLP*, 2004.

[40] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to spech synthesis with diverse voices and styles," in *Proc ICASSP*, 2007, pp. 1233–1236.

[41] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Commun.*, vol. 52, no. 5, pp. 394–404, May 2010.

[42] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. Systems*, vol. E 88-D, no. 3, pp. 1092–1099, 2005.

[43] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for hmm-based expressive speech synthesis," *IEICE Trans. Inf. Systems*, vol. E 90-D, no. 9, pp. 1406–1413, 2007.

[44] J. Gauffin and J. Sundberge, "Pharyngeal constrictions," *Phonetica*, vol. 35, pp. 157–168, 1978.

[45] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, vol. 53, pp. 1070–1082, Nov 1973.

[46] S. Maeda, "An articulatory model of the tongue based on a statistical analysis." *J. Acoust. Soc. Amer.*, vol. 65, pp. S22–S22, 1979.

[47] D. Beautemps, P. Badin, and G. Bailly, "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-films data for a reference subject, and articulatory-acoustic modeling," *J. Acoust. Soc. Amer.*, vol. 109, no. 5, pp. 2165–2180, 2001.

[48] N. S., A. A, and H. K, "An articulatory study of fricative consonants using magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 98, no. 3, pp. 1325–1347, 1995.

[49] P. Palo, "A review of articulatory speech synthesis," Master's thesis, Helsinki university of technology, 2006.

[50] J. M. Heinz and K. N. Stevens, "On the derivation of area functions and acoustic spectra from cineradiographic films of speech," *J. Acoust. Soc. Amer.*, vol. 36, no. 5, pp. 1037–1038, 1964.

[51] O. Engwall, "Modeling of the vocal tract in three dimensions," in *Proc. EUROSPEECH*, 1999.

[52] G. Fant, *Acoustic theory of speech production*. s-Gravenhage, Netherlands: Moutan & Co, 1960.

[53] H. K. Dunn, "The calculation of vowel resonances, and an electrical vocal tract," *J. Acoust. Soc. Amer.*, vol. 22, pp. 740–753, 1950.

[54] J. Kelly and C. Lochbaum, "Speech synthesis," in *Proc.International Congress on acoustics*, 1962.

[55] P. Badin and G. Fant, "Notes on vocal tract computation," STL-QPSR, Tech. Rep., 1984.

[56] H. Dudley, "The vocoder," Bell laboratories, Tech. Rep., 1939.

[57] G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[58] R. Carlson, T. Sigvardson, and A. Sjlander, "Data-driven formant synthesis," TMH-QPSR, Tech. Rep., 2002.

[59] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP, pp. 373-376*, 1996.

[60] J. P. Olive, "Rule synthesis of speech from dyadic units," in *Proc. ICASSP*, 1977.

[61] A. W. Black and N. Campbell, "Optimising selection of units from speech databse," in *Proc. Eurospeech*, 1995.

[62] J. L. Courbon and F. Emerald, "A text to speech machine by synthesis from diphones," in *Proc. ICASSP*, 1982.

[63] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, pp. 452–467, 1990.

[64] P. Taylor, *Text to Speech Synthesis*. Cambridge university press, 2009.

[65] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum , pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999.

[66] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005," *IEICE Trans. INF & SYST.*, vol. E90-D, pp. 325–333, 2007.

[67] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, pp. 1039–1064, 2009.

[68] T. Yoshimura, "Simultaneous modeling of phonetic and prosodic parameters and characteristic conversion for hmm-based text-to-speech systems," Ph.D. dissertation, Nagoya Institute of Technology, 1999.

[69] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from hmm using dynamic features," in *Proc. ICASSP*, 1995, pp. 660–663.

[70] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructering speech representations using a pitch adaptive time-frequency smoothin and an instantaneous frequency based f0 extraction: Possible role of repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3-4, pp. 187–207, 1999.

[71] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc ICASSP*, 1983, pp. 93–96.

[72] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "Hmm-based speech synthesiser using the lf-model of the glottal source," in *Proc. ICASSP*, 2011.

[73] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Feature-space transform tying in unified acoustic-articulatory modelling of articulatory control of hmm-based speech synthesis," in *Proc. Interspeech*, 2011.

[74] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlemeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[75] G. F. Banks and L. W. Hoaglin, "An experimental study of duration characteristics of voice during the expression of emotion," *Speech Monographs*, vol. 8, pp. 85–90, 1941.

[76] T. Jhonstone and K. R. Scherer, "The effects of emotions on voice quality," in *Proc. Int. Congr. Phoetic Sciences*, San Fransisco, 1999, pp. pp. 2029–2031.

[77] N. Campbell, "Developments in corpus -based speech synthesis: Approaching natural conversational speech," *IEICE Trans*, vol. 87, pp. 497–500, 2004.

[78] C. T. Ishii and N. Campbell, "Analysis of acoustic- prosodic features of spontaneous expressive speech," in *Proc. I st International Congress of Phonetics and Phonology*, Kobe, Japan, 2002, pp. pp. 85–88.

[79] W. L. Johnson, S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. Labore, "Limited domain synthesis of expressive military speech for animated characters," in *Proc. IEEE Speech Synthesis Workshop*, Santa Monica, CA, Sep. 2002.

[80] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, *LDC Emotional Prosody Speech Transcripts database*, univeristy of pennsylvania, Linguistic data consortium, 2002.

[81] Y. Hashizawa, S. T. M. D. Hamzah, and G. Ohyama, "On the differences in prosodic features of emotional expressions in Japanese speech according to the degree of the emotion," in *Proc. Speech Prosody*, 2004, pp. pp. 655–658.

[82] D. Erickson, T. Schochi, C. Menezes, H. Kawahara, and K.-I. Sakakibara, "Some non-f0 cues to emotional speech: An experiment with morphing," in *proc. Speech Prosody*, 2008, pp. 677–680.

[83] W. Hess, *Pitch determination of speech signals*. Berlin: Springer-Verlag, 1983.

[84] J. R. Deller, J. G. Proakis, and J. H. L. Hanson, *Discrete-time processing of speech signals*. New York: Mcmillan, 1993.

[85] J. D. Markel, "The sift algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, 1972.

[86] M. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-22, pp. 353–362, 1974.

[87] S. R. M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. ICASSP*, Montreal, Canada, May 2004.

[88] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 4, pp. 325–333, Sep.1995.

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

[90] K. S. R. Murty and B. Yegnanarayana, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469–472, June 2009.

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

[92] M. Farrus and J. Hernando, "Using jitter and shimmer in speaker verification," *IET signal process.*, vol. 3, no. 4, pp. 247–257, 2009.

[93] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. on speech and audio process.*, vol. 7, no. 6, pp. 609–619, 1999.

[94] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Am.*, vol. 47, pp. 634–648, Feb 1970.

[95] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, pp. 637–655, 1971.

[96] B. Yegnanarayana, "Formant extraction from linear-prediction spectra," *J. Acost. Soc. Am.*, vol. 63, no. 5, pp. 1638–1641, May 1978.

[97] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *Speech Commun.*, vol. 10, no. 3, pp. 209–221, 1991.

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

[99] E. Mourlines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, pp. 175–205, 1995.

[100] E. Hardam, "High quality time scale modification of speech signals using fast synchronized overlap add algorithms," in *Proc. IEEE*, 1990.

[101] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," in *Proc. ICASSP*, 1997, pp. 1303–1306.

[102] R. Muralishankar, A. G. Ramakrishnan, and P. Prathibha, "Modification pitch using dct in the source domain," *Speech Commun.*, vol. 42, pp. 143–154, 2004.

[103] K. S. Rao and B. Yegnanarayana, "Prosodic manipulation using instants of significant excitation," in *IEEE Int. Conf. Multimedia and Expo*, 2003.

[104] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *IEEE Signal Processing Letters*, vol. 14, pp. 762–765, Oct. 2007.

[105] D. Ruinskiy and Y. Lavner, "Stochastic models of pitch jitter and amplitude shimmer for voice modi cation," in *Proc IEEEI 2008*, 2008, pp. 489–493.

[106] K. S. Rao and B. Yegnanarayana, "Voice conversion by prosody and vocal tract modification," in *Proc. ICIT*, Bhubaneswar, 2006.

[107] K. S. Rao, "Voice conversion by mapping the speaker-specific features using pitch synchronous approach," *Computer, Speech And Language*, vol. 24, no. 3, pp. 474–494, July 2010.

[108] M. A. Joseph, M. H. Reddy, and B. Yegnanarayana, "Speaker-dependent mapping of source and system features for enhancement of throat microphone speech," in *Proc. INTERSPEECH 2010*, 2010, pp. 985–988.

[109] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text to speech synthesis system for American English," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, pp. 1099–1109, July 2006.

[110] *M Brookes Voicebox: A Speech Processing Toolbox for MATLAB 2006. [Online] Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox /voicebox.html.*

[111] B. Yegnanarayana, S. R. M. Prasanna, and G. Seshadri, "Study of robustness of zero frequency resonator method for extraction of fundamental frequency," in *ICASSP*, May 2011.

[112] M. R. Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Trans. Acoustics, Speech, Signal Process*, vol. ASSP-29, pp. 374–390, Jun 1981.

[113] T. F. Quatieri and R. J. McAulay, "Shape invariant time scale and pitch modification of speech," *IEEE Trans. on Signal Process.*, vol. 40, no. 3, pp. 497–510, Mar 1992.

[114] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Application of the dypsa algorithm to segmented time scale modification of speech," in *Proc. European Signal Processing Conference*, 2008.

[115] H.-Y. Gu and W.-L. Shiu, "A mandarin-syllable signal synthesis method with increased flexibility in duration, tone and timbre control," *Proc. Natl. Sci. Counc. ROC(A)*, vol. 22, no. 3, pp. 385–395, 1998.

[116] H.-Y. GU, "Notes for the syllable-signal synthesis method: Tipw," in *in proc. ISCSLP*, 1998.

[117] M. P. Pollard, et.al, "Enhanced shape-invarient pitch and time-scale modification for concatenative speech synthesis," in *proc. ICSLP*, 1996.

[118] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.

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

[120] W. Chase and F. Bown, *General Statistics*. Newyork : John Wiley and sons, 2000.

[121] K. S. Rao and B. Yegananarayana, "Duration modification using glottal closure instants and vowel onset points," *Speech Commun.*, vol. 51, no. 12, pp. 1263–1269, Dec. 2009.

[122] J. L. Flanagan, *Speech Analysis Synthesis and Perception", Springer-Verlag, Berlin*, Berlin, Ed. Springer-Verlag, 1972.

[123] K. N. Stevens, *Acoustic Phonetics*, U. Cambridge, MA, Ed. MIT Press, 1999.

List of Publications

Journals

1. D. Govind and S. R. M. Prasanna, "Dynamic Prosody Modification using Zero Frequency Filtered Signal", Int. Journal of Speech Technology (IJST), Springer [Available: DOI 10.1007/s10772-012-9155-3].

2. D. Govind and S. R. M. Prasanna, "Expressive Speech Synthesis : A Review", Int. Journal of Speech Technology (IJST), Springer [Available: DOI 10.1007/s10772-012-9180-2].

Conference Papers

3. D. Govind and S. R. M. Prasanna, "Epoch extraction from emotional speech", accepted for publication in SPCOM 2012, July 2012

4. D. Govind, S. R. M. Prasanna and B. Yegnanarayana, "Significance of glottal activity detection in duration modification", in Proc. Speech Prosody 2012, May 2012

5. D. Govind, S. R. M. Prasanna and B. Yegnanarayana, "Neutral to emotion speech conversion using source and supra-segmental information", in Proc. INTERSPEECH, Aug. 2011, pp. 2969-2972

6. D. Govind, S. R. M. Prasanna and Debadatta Pati, "Epoch extraction in high pass filtered speech", in Proc. INTERSPEECH, Aug. 2011, pp. 1977-1980

7. S. R. M. Prasanna and D. Govind "Analysis of excitation source information in emotional speech", in Proc. INTERSPEECH, Sep. 2010, pp. 781-784

8. S. R. M. Prasanna, D. Govind, K. S. Rao and B. Yegnanarayana "Fast prosody modification using instants of significant excitation", in Proc. Speech Prosody, May 2010.

9. D. Govind and S. R. M. Prasanna, "Expressive speech synthesis using prosody modification and dynamic time warping", in Proc. NCC, Jan. 2009, pp. 329-333

CURRICULUM VITAE

1. **NAME:** D. Govind

2. **DATE OF BIRTH:** 20 APRIL 1983

3. **EDUCATIONAL QUALIFICATIONS:**

   - April-2005 B.Tech
   - April-2007 M.Tech
   - July-2013 Ph.D

4. **PERMANENT ADDRESS:**

s/o P. G. Divakaran

"Gayathri"[Pukkattu House], Nanthyattukunnam, N. Parur

Eranakulam (District)

Kerala (State), India

683513

Ph: +91-484-2445731