Glottal Activity Region based Processing for Speech Synthesis

NAGARAJ ADIGA

Glottal Activity Region based Processing for Speech Synthesis

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

NAGARAJ ADIGA

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

May 2017

Certificate

This is to certify that the thesis entitled "**Glottal activity region based processing for Speech Synthesis**", submitted by **NAGARAJ ADIGA** (11610235), 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.

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

To

My **parents Vasudeva Adiga, Rajeshwari Adiga,**

brother Sukesh Adiga

for their love, sacrifice, and support

&

My guide **Prof. S. R. M. Prasanna**

for his guidance and inspiration

Acknowledgements

I cherish all the experience I went through in the process of this graduate program, this has left me a little wiser for the day. Also, I feel I am blessed to have made it to this wonderful campus in the laps of mother nature. This thesis would not have been possible without the immense help and support of several people in various measures. I would like to convey my acknowledgment to all of them.

First and foremost, I express my sincere gratitude to my research supervisor, Prof. S. R. M. Prasanna for providing me an opportunity to work under his guidance. I am thankful for his continuous guidance in all aspects, constant motivation, and support throughout the doctoral studies. His sheer dedication, discipline, and hard work are the great sources of inspiration for me. It would be completely impossible for me to bring the research as well as the thesis to this form without the immense facilities provided by him in the EMST Laboratory and the freedom of work he has given to me. I also would like to sincerely thank him for providing me with the financial support for attending conferences and workshops.

I am thankful to my doctoral committee members Prof. S. Dandapat, Dr. R. Sinha, Dr. Ranbir Singh, and Dr. Priyankoo Sarmah for their encouragement and valuable suggestions on my work. I would like to thank faculty members and the office staffs of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their help in carrying out this research work. I am especially grateful to Prof. S. Dandapat and Prof. R. Sinha for insightful comments and constructive criticisms on the work to bring it to the current form. My special thanks to Dr. L. N. Sharma for maintaining the EMST laboratory smoothly.

I owe my special thanks to Dr. N R Ramesh, retired scientist F, National aerospace laboratory, and Sandeep Dabhade, my office colleague at Alcatel-Lucent for their constant motivation to do higher studies and join premier institute like IIT. Their philosophical life boosted my self-confidence and discussion with them encouraged me all the time.

I also would like to thank Dr. Chandra Sekhar Seelamantula for providing an opportunity to do research collaboration work on Riesz transform based speech synthesis in Spectrum Lab, IISc, Bangalore. I would like to acknowledge the help of Jitendra, Sunil, Harsha, and other members of Spectrum lab for the technical discussions and support during my stay.

I am grateful to all my EMST lab senior members Dr. Debadatta Pati, Dr. Govind, Dr. Gayadhar

Pradhan, Dr. Sumitra, Dr. Haris, Dr. Sunil, Dr. Syed, Dr. Deepak, and Ramesh sir for mentoring me during my research life in EMST lab.

I am thankful to my friends Deepak, Syed, Banri, Biswa, Padhy, Anurag, and Vikram for having a useful technical discussion to improve my knowledge and also their help in patiently correcting my papers and thesis. My sincere gratitude to my friends Banri, Biswa, Jiss, Padhy, Anurag, Rohan, Bhukya, Bhanupriya, Suman, Rajib, Subhasis, Bidisha, Himakshi, Vikram, Sisir, Akhilesh, Protima, Abhishek, Sarfaraz, Mrinmoy, Rajesh, Tilendra, Vineeta, and all other members of the EMST Laboratory for their help in various research work.

I thank my fellow project mates Sandeep Reddy, Anand, Bidisha, Gyanendro, Rajlakshmi, and all the members of TTS project for their help in building Assamese/Manipuri TTS systems and also their help in doing the subjective evaluation for the various TTS systems developed for publishing papers.

I would like to thank MHRD, Govt. of India, for providing me scholarship during my Ph.D. period. I also thank International Speech Communication Association (ISCA), Department of Science and Technology (DST), and Signal Processing Society (SPS) for providing me with financial support for attending conferences.

I would like to thank my friends Dheeraj Kumar Sinha (Dheeru), Satyabrata Dash (Das), Umesh Chaudhary (Umi), and Kashyap Kumar Prabhakar (Monu) without them my life at IIT Guwahati would have been incomplete for discussion related to non-technical things and keeping me motivated in research. Most importantly, none of this would have been possible without the selfless love, affection, and support of my parents and brother.

Finally, I attribute this achievement to my parents for their constant blessings, support, silent prayers for my success and moreover, making me stand in this position.

NAGARAJ ADIGA

Abstract

This thesis demonstrates the significance of glottal activity region based processing for speech synthesis. The glottal activity region is perceptually significant and comprises the majority of speech sounds. The distinct features derived from the glottal activity regions may therefore be used for speech synthesis. In particular, the thesis is focused on improving the voice quality of statistical parametric speech synthesis (SPSS) by glottal activity region based processing.

The naturalness of speech is mainly attributed to the source signal. In a conventional way, it is modeled by the impulse excitation for voiced sounds. It represents only one aspect of the voice source signal, namely periodic component. However, voice source signal consists of other attributes like aperiodic component and phase component. It may be difficult to derive all these aspects of voiced source signal using the conventional features based on the segmental processing of speech, which captures the average information of speech production system.

The intelligibility of speech, that is, message information is represented by the vocal tract system. The features used to represent the vocal tract system, like, linear prediction coefficients and Mel-cepstral coefficients (MCEP) are processed by segment wise and may not capture the coarticulation effect of the production mechanism efficiently. In addition, the excitation design of SPSS currently is based on a two-state model which depends on the accuracy of voicing decision. The failures like, voiced region classifying as a unvoiced region gives hoarseness to voice quality, whereas unvoiced region classifying as voiced region gives buzziness. Therefore, there need to be suprasegmental features which do voicing decision accurately.

In this thesis, the voicing decision is computed using the features such as the strength of excitation, normalized autocorrelation peak strength, and higher-order statistics, which depicts different attributes of glottal activity, namely, energy, periodicity, and asymmetrical

nature of the glottal signal, respectively. The combination of these three features outperforms state-of-the-art algorithms, demonstrating different aspects represented by each of these features for voicing decision.

To model the vocal tract system, a novel spectral feature based on the 2-D processing of spectrogram using Riesz transform is proposed. The Riesz transform is used for the demodulation of spectrogram for a longer patch of size 100 ms in the time domain and 600 Hz in the frequency domain. The smoothed spectral envelope computed from the spectrogram patches captures coarticulation mechanism effectively and it is compactly represented using MCEP. Further, demodulated carrier signal is used for the computation of voicing decision, aperiodicity map, and F0 estimation.

The source modeling is designed by processing integrated linear prediction residual (ILPR) in the glottal activity region. The nature of ILPR waveform resembles the glottal flow derivative signal and may preserve the speaker characteristics in a better way. The glottal activity regions constituted by periodic, aperiodic, and phase components. The magnitude spectrum of ILPR signal is modeled in the frequency domain by dividing the spectrum into two bands to characterize periodic and aperiodic components of the voice speech segment. The periodic components of ILPR signal below the maximum voiced frequency (f_m) are represented using residual Mel-cepstral coefficients (RMCEP), whereas aperiodic component above f_m is represented by white Gaussian noise modulated with pitch adaptive triangular noise envelope weighted by the strength of excitation (SoE). The phase component of ILPR signal is represented using the all-pass filter coefficients computed from cosine phase of ILPR signal with an iterative procedure.

For any practical application of speech synthesis system, along with analysis/synthesis framework (vocoder), proposed features have to be integrated into SPSS. Hence, voicing decision, vocal tract system features, and source features computed in the glottal activity region are applied individually to SPSS. The results show that the proposed features represent different aspects of speech production system which depicts the prosody, naturalness, and intelligibility of speech, respectively. Finally, all the proposed features are modeled together in SPSS. The experimental results presented in this thesis work shows that the glottal activity region based processing helps in improving the naturalness and prosody by

preserving speaker-dependent characteristics and better modeling of the vocal tract system to enhance the intelligibility of SPSS.

The primary contributions of this thesis are as follows:

- Glottal activity region detection using three glottal source features, namely, the strength of excitation (SoE), normalized autocorrelation peak strength (NAPS), and higher-order statistics (HOS).
- Showing the significance of glottal activity features for speech synthesis. Then using glottal activity region detection as a voicing indicator and improving the accuracy of voicing decision with the support vector machine classifier. Finally, applying voicing decision for speech synthesis in an SPSS framework.
- 2-D based processing of speech spectrogram using Riesz transform to get the smoothed vocal tract envelope. In addition, Riesz transform provides 2-D pitch map, voicing decision, and aperiodicity spectrum. Finally, modeling these Riesz parameters in SPSS and showing its importance for improving the quality of SPSS.
- Source modeling using different aspects of glottal activity region like epoch strength, aperiodic component, and phase component using ILPR signal. Finally, showing the significance of aperiodic and phase components in SPSS framework.
- Combining suprasegmental, source, and system features to improve the prosody, naturalness, and intelligibility of SPSS, respectively.

Keywords: Glottal activity region, Riesz transform, 2-D processing, integrated linear prediction residual, statistical parametric speech synthesis.

Contents

List of Figures xxi

List of Tables xxvii

List of Acronyms xxix

1 Introduction 1

1.1 Overview of Text-to-speech synthesis . . . 2

1.1.1 HMM based speech synthesis system . . . 3

1.2 Issues in the basic version of SPSS . . . 5

1.3 Analysis/Synthesis methods . . . 7

1.3.1 STRAIGHT vocoder . . . 7

1.3.2 Glott Vocoder . . . 10

1.4 Motivation for the present work . . . 13

1.5 Organization of the Thesis . . . 14

2 Acoustic Features for Statistical Parametric Speech Synthesis: A Review 17

2.1 Introduction . . . 18

2.2 Acoustic feature modeling using different SPSS techniques . . . 19

2.2.1 Hidden Markov model . . . 19

2.2.2 Deep learning model . . . 21

2.2.3 Hybrid Synthesis . . . 22

2.3 Suprasegmental features . . . 22

2.3.1 Multi-space distribution model . . . 23

2.3.2 Continuous distribution . . . 23

2.3.3 Acoustic features for F0 and voicing decision . . . 24

2.4 Vocal tract Spectral model . . . 25

2.4.1 Linear prediction model . . . 26
2.4.2 Cepstrum based methods . . . 26
2.4.3 MLSA filter . . . 27
2.4.4 STRAIGHT Spectrum estimation . . . 27
2.5 Different source models . . . 29
2.5.1 Mixed Band Excitation . . . 29
2.5.1.1 Mixed Excitation . . . 29
2.5.1.2 STRAIGHT based source model . . . 30
2.5.2 Residual modeling . . . 32
2.5.2.1 Closed loop training . . . 32
2.5.2.2 Pitch synchronous codebook . . . 33
2.5.3 Glottal source modeling . . . 34
2.5.3.1 LF model . . . 34
2.5.3.2 Glott model . . . 35
2.6 Other Signal Processing Models . . . 36
2.6.1 Sinusoidal modeling . . . 36
2.6.2 Harmonic-noise models . . . 37
2.6.3 Evaluation Parameters . . . 38
2.6.3.1 Naturalness . . . 38
2.6.3.2 Intelligibility . . . 39
2.7 Summary and Discussion . . . 40
2.7.1 WaveNet . . . 42
2.8 Organization of the Present Work . . . 43
3 Glottal Activity Region Detection 45
3.1 Motivation for detecting glottal activity region . . . 46
3.2 Features for characterization of glottal activity . . . 47
3.2.1 Strength of Excitation (SoE) . . . 48
3.2.2 Normalized autocorrelation peak strength (NAPS) . . . 50
3.2.3 Higher-order statistics (HOS) measure . . . 50
3.3 Robustness of features for characterizing glottal activity . . . 51

3.3.1 Zero-frequency filtered signal . . . 51
3.3.2 Integrated linear prediction residual signal . . . 53
3.4 Glottal activity region detection using different attributes . . . 54
3.5 Summary . . . 59
4 Glottal Activity Features for Speech Synthesis 61
4.1 Motivation for processing glottal activity region for synthesis . . . 62
4.2 Significance of Glottal activity features for Speech Synthesis . . . 64
4.2.1 Cosine phase as excitation . . . 64
4.2.2 Epoch and its location . . . 66
4.2.3 Epoch strength . . . 67
4.2.4 Voicing decision . . . 67
4.3 Glottal activity features for voicing decision . . . 68
4.3.1 Analysis of F0 for different voiced sounds . . . 69
4.3.2 Analysis of SoE, NAPS, and HOS for voicing decision . . . 70
4.4 Improvement in the detection of glottal activity region using classifiers . . . 72
4.4.1 Voicing classification . . . 74
4.4.2 Evaluation parameters . . . 75
4.4.3 Results . . . 75
4.5 Glottal activity features for Synthesis . . . 77
4.5.1 Integration of voicing decision in SPSS . . . 78
4.6 Experimental evaluation . . . 79
4.6.1 Subjective evaluation . . . 81
4.6.2 Objective evaluation . . . 83
4.7 Summary . . . 84
5 Riesz Transform for Speech Synthesis 85
5.1 Motivation behind the 2-D processing . . . 86
5.2 The Riesz transform based demodulation . . . 88
5.2.1 Hilbert Transform . . . 88
5.2.2 Riesz Transform . . . 89
5.2.3 Demodulation in 2-D using Riesz transform . . . 90

5.3 Riesz transform based demodulation for speech spectrum . . . 91
5.3.1 AM envelope . . . 91
5.3.2 Carrier spectrogram . . . 93
5.4 Carrier spectrogram analysis . . . 94
5.4.1 Coherence map . . . 94
5.4.2 Voicing decision . . . 95
5.4.3 Pitch estimation . . . 96
5.5 Synthesis methodology . . . 98
5.5.1 Synthesis using STFT Phase . . . 99
5.5.2 Synthesis using F0 . . . 100
5.5.3 Synthesis using Random phase . . . 101
5.5.4 Synthesis using Aperiodicity . . . 101
5.6 Experimental validation . . . 102
5.6.1 Analysis and Synthesis framework . . . 103
5.6.1.1 Objective evaluation . . . 104
5.6.1.2 Subjective evaluation . . . 105
5.6.2 Statistical parametric speech synthesis . . . 106
5.6.2.1 Objective evaluation . . . 108
5.6.2.2 Subjective evaluation . . . 108
5.7 Summary . . . 109

6 Integrated Linear Prediction Residual for Source Modeling 111

6.1 Different components of source modeling . . . 112
6.2 Different components present in Glottal activity region . . . 114
6.2.1 Epoch based excitation model . . . 114
6.2.2 Experimental studies and discussion . . . 117
6.3 Periodic and Aperiodic modeling . . . 120
6.3.1 ILPR signal . . . 120
6.3.2 Residual MCEP representing harmonic component . . . 121
6.3.3 Noise component . . . 122
6.3.4 ILPR source modeling for HMM based speech synthesis . . . 123

6.3.5 Experimental evaluation . . . 124
6.4 Phase modeling . . . 128
6.4.1 Phase modeling using Integrated linear prediction residual . . . 128
6.4.2 Cosine phase modeling . . . 131
6.4.3 Experimental evaluation . . . 133
6.5 Summary . . . 136
7 Suprasegmental, System, and Source features for Speech Synthesis 139
7.1 Introduction . . . 140
7.2 Glottal activity region based processing . . . 141
7.2.1 Suprasegmental features . . . 141
7.2.2 System feature . . . 142
7.2.3 Source features . . . 143
7.3 Proposed analysis/synthesis framework . . . 144
7.4 Experimental evaluation . . . 145
7.4.1 Database . . . 146
7.4.2 HMM based speech synthesis system . . . 146
7.4.3 DNN based speech synthesis system . . . 149
7.5 Summary and Discussion . . . 151
8 Summary and Conclusions 153
8.1 Summary of the work . . . 154
8.2 Contributions of this thesis . . . 156
8.3 Directions for future work . . . 157
Bibliography 159
List of Publications 169

List of Figures

1.1 Block diagram of HMM speech synthesis system 4

1.2 Illustration of over-smoothing effect in SPSS: (a) and (b) showing the spectrogram view of natural and synthetic speech with rectangular box showing the difference in the transitions of formants in natural and synthetic speech, respectively. 5

1.3 Examples of voiced segment of the natural speech, generated trajectories from the HMMs 6

1.4 Examples of sequence of MCEP of the natural speech, generated trajectories from the HMMs 7

1.5 F0 extraction of STRAIGHT 1.5(c) and its comparison with conventional method 1.5(b) for speech segment 1.5(a) 8

1.6 Aperiodic component present in glottal and non-glottal region 9

1.7 Block diagram of STRAIGHT synthesis 10

1.8 Synthesis block diagram of Glott vocoder 11

1.9 Glottal pulse 1.9(b) and glottal flow derivative 1.9(c) extracted using IAIF for speech segment 1.9(a) 12

2.1 Synthesis block diagram of the pitch synchronous codebook method for source modeling 34

3.1 Nature of EGG (a) and DEGG (b) for glottal (0-0.05 s), glottal to nonglottal transition (0.05-0.08 s), and nonglottal (0.08-0.1 s) regions with glottal opening and closing marked in continuous and dashed arrow, respectively. 47

3.2 Characterization of glottal activity regions from source signal for a segment *Philip produced a coup*: (a) segment of source representations from DEGG, ZFFS, and ILPR, in three different columns of subplots; (b)-(d) three features SoE, NAPS, and HOS, respectively, obtained from each source representation (in three columns); reference marks are shown by dotted line in all the subplots. 49

3.3 Source signal representation for a speech segment consisting of glottal activity regions: (a) and (b) speech segment and its DEGG; (c) and (d) ZFFS and ILPR source signal derived from speech. 52

3.4 Scatter plot of DEGG vs. ZFFS, and ILPR source representation 54

3.5 DET curve of glottal activity region detection task from different methods: The DET curve is shown for ZFFS, ILPR and combined source signal in each subplot of Figure (a), (b), and (c), respectively, for clean, white, and babble noise cases at 0 dB. 56

3.6 Glottal activity region detection frame error (%) for all the methods in clean speech and noisy speech at 0 dB of SNR with two types of noise. 58

4.1 Excitation and synthesized speech for voiced segment: ((a)-(d)) A different types of excitation signal representing LP residual signal, cosine phase of LP residual, epoch based excitation signal, and strength weighted impulse signal, respectively; ((e)-(h)) corresponding synthesized speech from excitation signal obtained using LP residual, cosine phase of LP residual, impulse excitation signal, and strength weighted impulse signal 65

4.2 (a) SPSS synthesized speech for an English word "sleep"(/s/, /l/, /iy/, /p/,/sil/) using RAPT algorithm; (b) fundamental frequency with voicing decision; (c) spectrogram for the same word; ((d)-(f)) shows the SPSS synthesized speech for the same word using the proposed glottal activity features, fundamental frequency with voicing decision and spectrogram, respectively. 69

4.3 The distribution of different features present in glottal activity region for voicing decision: (a)-(d) relative frequency of feature values for F0, SoE, NAPS, and HOS, respectively 70

4.4 (a) Natural speech for a phrase "a big canvas"(/a/,/b/,/ih/,/g/,/k/,/ae/,/n/,/v/,/ah/,/s/) with reference voicing marking in dotted line; (b) voicing decision obtained from RAPT with amplitude 0.8, and normalized F0 evidence; (c)-(e) glottal activity evidence obtained from features, SoE, NAPS, and HOS, respectively; (f) proposed voicing decision (dotted line) using combined evidence of glottal activity features (continuous line) . . 71

4.5 Voicing decision using classifiers: (a)-(b) speech and DEGG segment with reference marking in dotted line; (c) signal processing combined evidence in continuous line (black color) and final nonlinear mapping in dotted line (red color); (d)-(f) the voicing evidence obtained from k-NN, DBN, and SVM, respectively, with likelihood probability in continuous line (black color) and mapping in dotted line. 73
4.6 Block diagram of SPSS 78
4.7 Integration of glottal activity features to the vocoder of SPSS 80
4.8 Average MOS of five different SPSS systems with RAPT, STRAIGHT, REAPER, Continuous, and GA model based voicing decision, respectively, for SLT (female speaker) and BDL (male speaker) 82

5.1 (a) Modulating signal: 2-D hamming window $V(\omega)$, (b) Carrier signal: 2-D cosine $\cos(\langle \mathbf{\Omega}_0, \boldsymbol{\omega} \rangle)$ at $\Omega_0 = 20\pi$ and $\theta_0 = \pi/4$, (c) Amplitude Modulated signal using hamming window, (d) Estimated envelope using CRT, (e) Estimated carrier using CRT, and (f) Error is envelope estimation. 91
5.2 Smoothed spectral envelope obtained from (a) Riesz transform; (b) STRAIGHT method 92
5.3 Carrier spectrogram computed for diphthong sound unit /ai/, where sound unit is present from 0.2 s to 0.85 s and nonharmonic component shown in rectangle dotted line around 0.5 s region. 93
5.4 Aperiodic spectrum computed for speech utterance taken from TIMIT database: (a) speech signal; (b) and (c) shows the aperiodic spectrum computed from STRAIGHT and Riesz transform, respectively. 95
5.5 Voicing decision with corrections for silence frames using Energy of frame: (a) speech signal taken from TIMIT database with voicing decision and transcription ; (b) coherence map showing the evidence of voicing region 96
5.6 (a) Carrier slice for a voiced frame and for a frequency sub-band from 0 to 350 Hz, (b) Fourier transform magnitude. 97
5.7 Pitch estimation using the coherence map: (a) speech signal along with the transcription; (b) carrier spectrogram obtained after demodulation of speech signal; (c) coherence map computed from carrier spectrogram using structure tensor method; (d) pitch map computed using carrier spectrogram and coherence map 98

5.8 F0 estimation from Riesz transform: (a) Speech utterance from TIMIT with transcription; (b) comparisons of F0 estimation from ZFF and Riesz coherence methods shown in red and black color, respectively. 99
5.9 Synthesis framework of proposed method 103
5.10 PESQ score of the male speakers for the proposed method in analysis/synthesis framework evaluated for 4 Indian languages 104
5.11 PESQ score of the female speakers for the proposed method in analysis/synthesis framework evaluated for 4 Indian languages 105
5.12 Block diagram of HMM based speech synthesis 107

6.1 Epoch based excitation signal derived from different methods: (a) a speech segment consist of voiced and unvoiced speech portions, (b) LP residual, (c) epochs location calculated from zero-frequency filtered signal, (d) epochs with their strength, ((e)-(g)) excitation signals derived around epochs by considering small portion of LP residual, Hilbert envelope, and Hilbert envelope + cosine phase, respectively. 116
6.2 Time domain waveforms of synthesized speech for word *I was*: (a) original speech signal, synthesized speech based on ((b)-(e)) strength weighted epochs, LP residual, Hilbert envelope, and Hilbert envelope + cosine phase, respectively. 118
6.3 Spectrogram view of synthesized speech for word *I was*: (a) original speech signal, synthesized speech based on ((b)-(e)) strength weighted epochs, LP residual, Hilbert envelope, and Hilbert envelope + cosine phase, respectively. 119
6.4 Source signal representation of a speech segment of voiced regions: (a) and (b) reference DEGG source signal and ILPR signal for a same speech segment, respectively. 121
6.5 Periodic and noise component of the ILPR source signal for voiced regions: (a) the ILPR source signal for a segment voiced speech; (b) and (c) periodic and noise component of the ILPR signal obtained by low-pass filtering and high-pass filtering the ILPR signal, respectively. 122
6.6 The work flow of the source modeling using ILPR signal 125
6.7 Average MOS of 4 HTS systems, impulse/noise, mixed, STRAIGHT and ILPR, respectively, for SLT and BDL speaker 127

6.8 Comparison of different type of excitation with its synthesized speech: ((a)-(d)) represents the residual, impulse, cosine phase, and group delay phase excitation , respectively; ((e)-(h)) represent synthesized speech for the excitation signal shown in ((a)-(d)), respectively 129
6.9 Integration of all-pass model to the proposed vocoder 133

7.1 Proposed Analysis framework for Glottal activity based processing for SPSS 144
7.2 Proposed Synthesis framework for Glottal activity based processing for SPSS 145
7.3 Average MOS of Proposed and STRAIGHT system for Assamese, Manipuri and English Database in HTS framework 149

List of Tables

2.1 Overall summary of different acoustic features for SPPS 41

3.1 *Pearson's correlation coefficient between DEGG vs. ZFFS and ILPR* 54
3.2 Glottal activity region detection performance in clean conditions represented in terms of EER for both DEGG and speech signal 55
3.3 Comparison of the proposed method with other methods in clean conditions represented in terms of glottal activity region detection frame error 57

4.1 PESQ for different types of excitation 68
4.2 Comparison of the different classifiers: represented in terms of the percentage (%) of voiced, unvoiced, and combined voicing error 76
4.3 Comparison of the frame level context information used in different classifiers: represented in terms of the percentage (%) of voicing error 77
4.4 Subjective evaluation results of MOS and PT with 95% confidence intervals from expert subject 81
4.5 Subjective evaluation results of MOS and PT with 95% confidence intervals from naive subjects 81
4.6 Objective evaluation results of PESQ, LSD, and VU_E with standard deviation 84

5.1 Performance of formant extraction in terms of detection rate using different method . 93
5.2 Pitch estimation from KEELE and CSTR database 97
5.3 PESQ score for Riesz and STRAIGHT spectral envelope with the different types of excitation 102
5.4 PESQ and SNR scores of Riesz and STRAIGHT methods for Indian TTS database . . 104
5.5 MOS and PT results for analysis/synthesis framework with 95% confidence interval . . 106

5.6 Comparison of Objective results using STRAIGHT and proposed analysis/synthesis framework for HMM based speech synthesis 108
5.7 MOS and PT results for SPSS with 95% confidence interval 109

6.1 MOS for different source modeling using epoch based excitation 120
6.2 Speech parameters used per frame for training the HTS 124
6.3 Subjective evaluation results of MOS and PT 126
6.4 Objective evaluation results of PESQ and LSD 127
6.5 PESQ for different types of excitation 131
6.6 Objective evaluation results of only phase modeling in proposed method 134
6.7 Objective results of proposed phase and aperiodicity modeling 134
6.8 Subjective evaluation results of only phase modeling in proposed method 135
6.9 Subjective results of proposed phase and aperiodicity modeling with 95% confidence interval 136

7.1 Assamese and Manipuri database showing the number of unique words, syllable and duration of each word 147
7.2 Speech parameters used per frame for the proposed Glottal activity region based system vs STRAIGHT 147
7.3 Comparison of objective results of STRAIGHT and proposed analysis/synthesis framework for HTS 148
7.4 Preference test (PT) results 149
7.5 Comparison of Objective results using STRAIGHT and proposed analysis/synthesis framework for SPSS 151

List of Acronyms

AM	Amplitude Modulation
APC	All-Pass filter Coefficients
BAP	Band Aperiodicity
CRT	Complex Riesz Transform
CTF	Complex Time Function
DBN	Deep Belief Network
DEGG	Differentiated Electroglottograph
DET	Detection Error Trade-off
DNN	Deep Neural Network
DR	Detection Rate
DSM	Deterministic and Stochastic Model
EER	Equal Error Rate
EGG	Electroglottograph
EM	Expectation-Maximization
FM	Frequency Modulation
GA	Glottal Activity
GCI	Glottal Closure Instant
GE	Gross Error
GLOTT	Glottal
GMM	Gaussian Mixture Model
GOI	Glottal Opening Instant
HE	Hilbert Envelope
HMM	Hidden Markov Model
HNM	Harmonic and Noise Model

HNR	Harmonic Noise Ratio
HOS	Higher Order Statistics
HTS	HMM based Speech Synthesis System
IAIF	Iterative Adaptive Inverse Filtering
ILPR	Integrated Linear Prediction Residual
k-NN	k-Nearest Neighbors
LF	Liljencrants-Fant
LP	Linear Prediction
LPC	Linear Prediction Coefficients
LSD	Log Spectral Distance
LSP	Line Spectral Pairs
LSTM	Long Short-Term Memory
MCEP	Mel-Cepstral Coefficients
MCD	Mel-Cepstral Distance
ME	Mean Error
MFCC	Mel-Frequency Cepstral Coefficients
MGCEP	Mel-Generalized Cepstral Coefficients
MLSA	Mel-Log Spectral Approximation
MOS	Mean Opinion Score
MSD	Multi-Space Distribution
NAPS	Normalized Autocorrelation Peak Strength
PCA	Principal Component Analysis
PESQ	Perceptual Evaluation of Speech Quality
PT	Preference Test
RAPT	Robust Algorithm for Pitch Tracking
RBM	Restricted Boltzmann Machine
REAPER	Robust Epoch And Pitch EstimatoR
RMCEP	Residual Mel-Cepstral Coefficients
RMSE	Root Mean Square Error
RN	Re-sampling and Normalizing

RNN	Recurrent Neural Network
SKR	Skewness to Kurtosis Ratio
SNR	Signal to Noise Ratio
SoE	Strength of Excitation
SPSS	Statistical Parametric Speech Synthesis
SRH	Summation of Residual Harmonics
STFT	Short Time Fourier Transform
STRAIGHT	Speech Transformations and Representations using Adaptive Interpolation weiGHTed spectrum
SVM	Support Vector Machine
TTS	Text-To-Speech
USS	Unit selection synthesis
V-UV	Voiced-Unvoiced
ZFF	Zero Frequency Filtering
ZFFS	Zero Frequency Filter Signal

1

Introduction

Contents

1.1 Overview of Text-to-speech synthesis 2
1.2 Issues in the basic version of SPSS 5
1.3 Analysis/Synthesis methods 7
1.4 Motivation for the present work 13
1.5 Organization of the Thesis 14

Speech is one of the important forms of communication among human being. The speech production mechanism is controlled by the brain, which involves the control and coordination of the neuromuscular activities associated with different speech production organs like lungs, larynx, and vocal tract [1, 2]. The lungs provide energy required for generating the airflow to the larynx. The larynx modulates the airflow from the lungs and provides either periodic air-puffs by the glottal vibration or provides airflow to the vocal tract constriction for generating turbulent noise [1]. The resulting source excitation signal is quasi-periodic in nature. The vocal tract spectrally shapes the source signal by articulatory movements to give different speech sounds, which are mainly classified as voiced and unvoiced sounds. Voiced sounds, such as vowels, semi-vowels, nasals, voiced stops, and voiced fricatives, are characterized by the glottal activity [3]. The frequency composition of these sounds is regular and can be described by the harmonic structure. These harmonics are emphasized near the resonance frequencies of the vocal tract are called as formants. Variations in the vocal tract shape, such as lips opening and tongue placement contribute to the differentiation between different types of speech sounds. Unvoiced sounds are generated either by creating a rapid flow of air through one or more constrictions at some point between the trachea and the lips or by forming a closure at the location of constriction and abruptly releasing it. The first action acts like a turbulent noise source excitation while the second action produces a transient excitation followed by a turbulent flow of air, such as the excitation of the stop consonants [1]. From long-time humans try to emulate these speech production mechanisms to generate "human-like" speech. For example, text-to-speech (TTS) synthesizers can generate speech sounds for a given input text. Despite the fact that the quality of the synthetic speech has yet to fully match the quality of human speech, these systems have been successfully used in daily applications like in-car navigation systems, e-book readers, voice-over functions for the visually impaired, spoken dialog systems, communicative robots, singing speech synthesizers, and speech-to-speech translation systems [4].

1.1 Overview of Text-to-speech synthesis

TTS synthesis is a technique for generating artificial production of human speech for a given input text. Typical TTS systems have two main components, text analysis and speech waveform generation [4]. In the text analysis component, the input text is converted into a linguistic specification consisting of elements such as phonemes. In the speech waveform generation component, speech

waveforms are generated from the produced linguistic specification [5]. Most commonly used TTS are rule based formant synthesizer, concatenative speech synthesizer using festival, and statistical parametric speech synthesis (SPSS) [6–8]. In rule based formant synthesis [6], each phonetic units are picked by the rule and then speech is synthesized based on the source-filter model which internally consists of formant synthesizer. In concatenative speech synthesis [7], speech waveform units are stored in the database and while synthesizing, units are picked from the database and concatenated to get the waveforms. Units may be word, syllable, diphone, phone or half phone. Unit selection synthesis (USS) is most common concatenative synthesis system [7]. In SPSS [9], parametric models are built using statistics to captures the distribution of parameter value found in the training data. During the synthesis, waveforms are generated from models using the source-filter model. Although any generative model can be used for modeling these parameters. In this thesis, hidden Markov model (HMM) based SPSS framework is used in most of the chapters. To develop the system, HMM based speech synthesis system (HTS) software is used. The main advantage of the SPSS is its footprint size is much smaller compared to USS system [10]. The statistical representation also allows transforming voice characteristics, speaking styles, and emotions [9]. Hence, SPSS is a popularly used TTS.

1.1.1 HMM based speech synthesis system

The general framework of HMM based synthesis system is shown in Figure 1.1. The two main steps involved in developing SPSS are training and synthesis. In the training part, spectral and excitation parameters are extracted from the speech data [5]. Spectral parameters include Mel-cepstral coefficients (MCEP) and their dynamic features (delta and delta-delta). Excitation parameters consist of fundamental frequency (F0) and its dynamic features, where F0 is modeled as logarithmic of fundamental frequency (logF0) [8]. Both excitation and spectral parameters are trained in HMM using the expectation-maximization (EM) algorithm [11]. Modeling in speech synthesis is similar to speech recognition except that along with vocal tract parameters, excitation parameters are also modeled [8]. In addition, linguistic and prosodic contexts are considered for training the acoustic parameters.

The context-dependent mono-phone models are trained using acoustic features and the time-aligned phonetic transcriptions. The basic unit considered for HMM synthesis system is the context-dependent quinphones. The context-dependent models are built with a set of context-independent mono-phone HMMs [5]. In this process, acoustically similar states are tied in order to reduce the

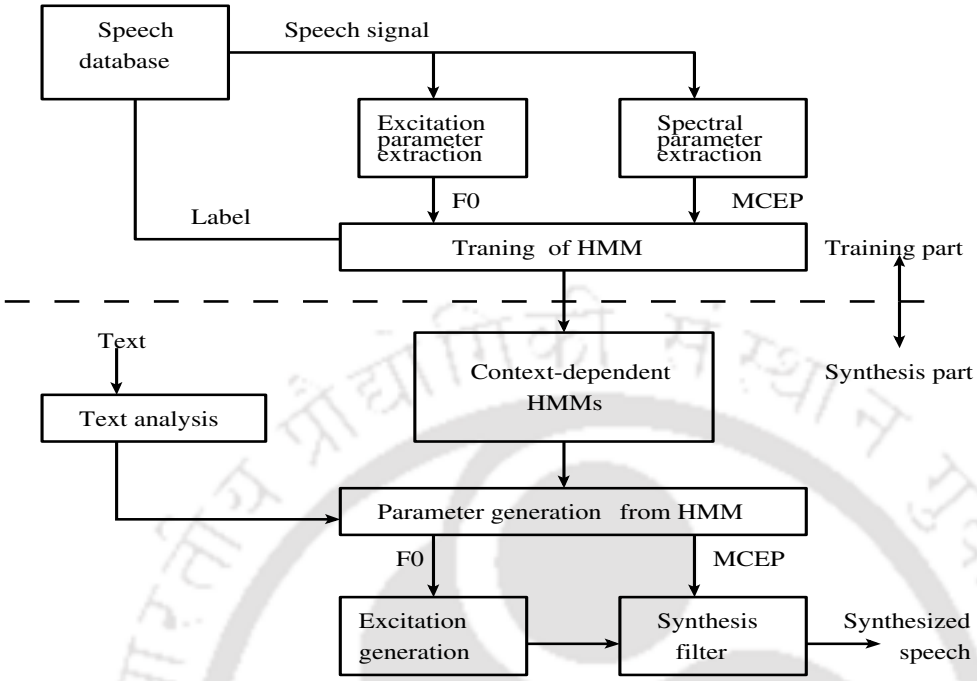

Figure 1.1: Block diagram of HMM speech synthesis system

total number of parameters without degrading the performance of the models [12]. Here, tree based clustering is used for state-tying.

In the synthesis part, the input text is converted into a context-dependent label sequence, and the utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Speech parameter generation algorithm generates the sequences of spectral and excitation parameters from the utterance HMM. Finally, a speech waveform is synthesized from the generated spectral and excitation parameters using a speech synthesis filter such as Mel-log spectrum approximation (MLSA) filter excited by the impulses [13]. Impulse spacing depends on the excitation parameter F0. In the base version of SPSS, impulse excitation is used for generation of voiced excitation. The MLSA filter is excited by periodic impulse excitation for voiced sounds, while white Gaussian noise excitation is used for generation of unvoiced sounds. Intelligibility of this system is good, however, naturalness is comparatively poorer than the USS system [14, 15].

1.2 Issues in the basic version of SPSS

- Acoustical source and spectral features include vocal tract and excitation features. The vocal tract features represent the spectral envelope of speech and compactly represented using MCEP [16]. The excitation features consist of fundamental frequency F0 of speech and glottal volume pulse parameters. Errors in pitch extraction can contribute to unnaturalness and degradation in voice quality. The pitch extraction algorithms include accurate voicing decision and pitch estimation. The conventional pitch estimation algorithms are based on the segmental analysis. These algorithms give a stepwise F0 trajectory. This can be seen in Figure 1.5 (b), where F0 is computed from robust algorithm for pitch tracking (RAPT) algorithm shows the stepwise trajectory of F0, however, in the natural speech its trajectory will be smooth [17, 18]. In addition, the decision regarding whether the speech segment is voiced/unvoiced can be erroneous when the speech signal is weakly periodic and lower amplitude [19, 20]. The MCEP computed from the short-term Fourier transform (STFT) spectrogram still consists of glottal source effect [18]. Moreover, due to fixed window analysis of speech, STFT spectrogram may not capture the transitions and coarticulation effect of articulatory movements, which is shown in Figure 1.2 (b) around 2.3 s with dotted rectangle box, whereas natural waveform is shown around 2.4 s for the same segment having proper formant transitions.

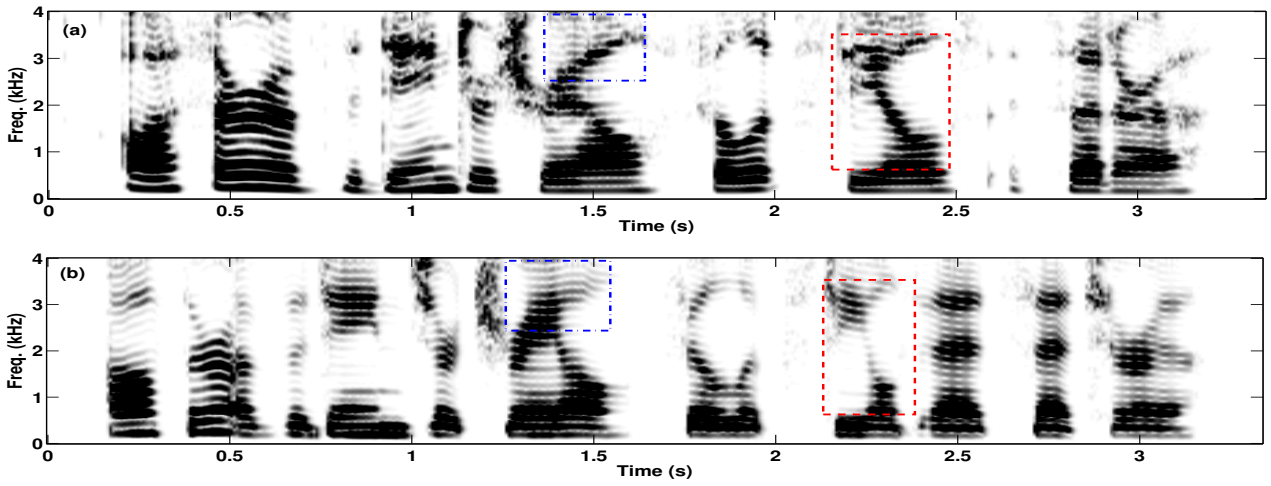

Figure 1.2: Illustration of over-smoothing effect in SPSS: (a) and (b) showing the spectrogram view of natural and synthetic speech with rectangular box showing the difference in the transitions of formants in natural and synthetic speech, respectively.

- Statistical modeling causes the over-smoothing of the generated parameters from the HMM [21].

The generated speech parameter trajectories from SPSS are over-smoothed when compared with the natural speech. The statistical averaging of the parameters for different phonemes in different contexts introduces smoothing. The natural variation in the original parameter trajectories cannot be reconstructed due to over-smoothing of the parameters. The over-smoothing effect is present in both the time and the spectral domain. These effects make the synthetic speech sound muffled [21]. Figure 1.3 shows the effect of smoothing in time domain for a voiced segment, where for the synthesized waveform variations are not captured well and it is smoothed out compared to the natural waveform. In addition, vocal tract response decays fast, when compared to natural speech. In the spectral domain, MCEP are used to represent the vocal tract response. The dynamic variations present in these parameters are less in the modeled parameters due to the Gaussian mixture modeling of MCEP. This is illustrated in Figure 1.4 (a) and (b) shows the second and third MCEP, respectively. The parameter shown in the figure is extracted from a sentence taken from ARCTIC database [22]. The MCEP are able to grossly capture the transitions of vocal tract response. However, the natural variations present in the original waveforms are absent in the case of synthesized speech.

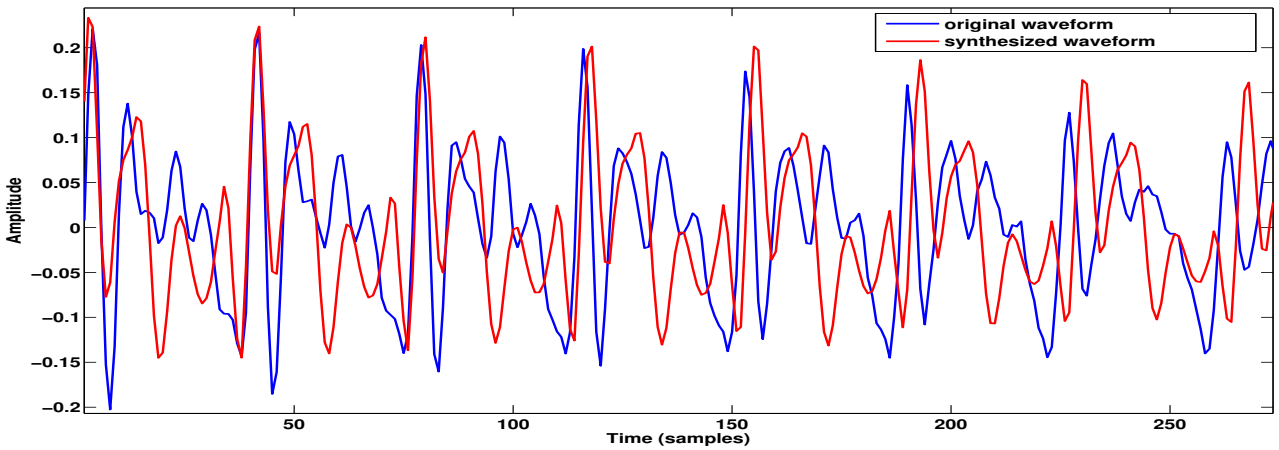

Figure 1.3: Examples of voiced segment of the natural speech, generated trajectories from the HMMs

- Over-simplified vocoder model is not able to generate the quality of speech present in natural speech [14, 23–26]. The speech synthesized by the basic version of SPSS sounds buzzy since it uses MCEP features representing a vocal tract function with a simple periodic impulse train or white Gaussian noise as excitation. However, in the speech production mechanism, the voice source excitation is represented by the impulse excitation, which represent only the periodicity

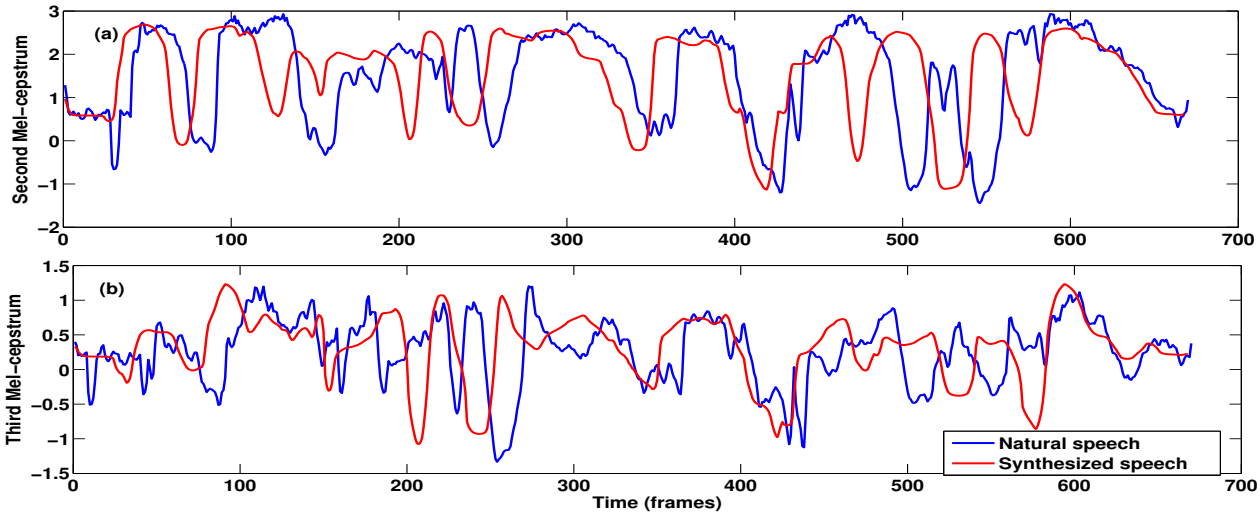

Figure 1.4: Examples of sequence of MCEP of the natural speech, generated trajectories from the HMMs

effect of the source signal. The other aspects of source signal like aperiodicity, phase delay, and variations in the strength of excitation are also important for the naturalness of speech. Further, the effect of impulse excitation is illustrated with spectrogram plot of the natural and synthesized waveform in Figure 1.2 for a sentence taken from ARTIC database [22]. The effect of impulse excitation can be seen throughout the duration of the synthesized waveform in this Figure 1.2 with regular harmonic structure even in higher frequency, particularly around 1.5 s region shown in a rectangle box, which is absent in natural waveform.

1.3 Analysis/Synthesis methods

In order to address the aforementioned issues present in the conventional analysis/synthesis framework, different attempts are tried in the literature. Among them, speech transformation and representation using adaptive interpolation of weighted spectrum (STRAIGHT) and Glott (glottal) vocoder are very popular and these models are also suitable for statistical framework [18, 27]. Hence, STRAIGHT and Glott vocoder are briefly explained here to get cues for improving naturalness and intelligibility of SPSS.

1.3.1 STRAIGHT vocoder

Kawahara *et al.* proposed the STRAIGHT vocoder [18]. In this framework, the speech signal is assumed as the convolution of a spectrally flat excitation by the spectral envelope of the speech

signal. During speech analysis, spectral envelope, F0, and aperiodicity parameters are computed from the speech signal. For the synthesis of voiced speech, a mixed multi-band excitation is an input to the synthesis filter defined by the spectral parameters. In case of unvoiced speech, the excitation is modeled as white noise. Original STRAIGHT parameters are represented as F0, Fourier transform magnitudes, and aperiodicty measurements, which are adapted to SPSS by Zen *et al.* [15].

F0 extraction

In STRAIGHT approach, the speech signal is assumed to be a nearly sinusoidal model, where speech waveform is the nearly harmonic sum of frequency modulation (FM) sinusoids modulated by amplitude modulation (AM). For flexible and high-quality modification of speech parameters, it is important to extract F0 trajectories which do not have any trace of interference caused by interaction between analysis window and the waveform of the signal. The conventional F0 extraction method uses interval measurements and provides stepwise F0 trajectories. This stepwise trajectory says that the signal is periodic. However, pitch period of speech is keep changing slowly for each glottal cycle. Hence, F0 is extracted using instantaneous frequency (a derivative of the instantaneous phase) in STRAIGHT. Extracted F0 from STRAIGHT method is plotted in Figure 1.5(c), which shows the smoothed F0 trajectories.

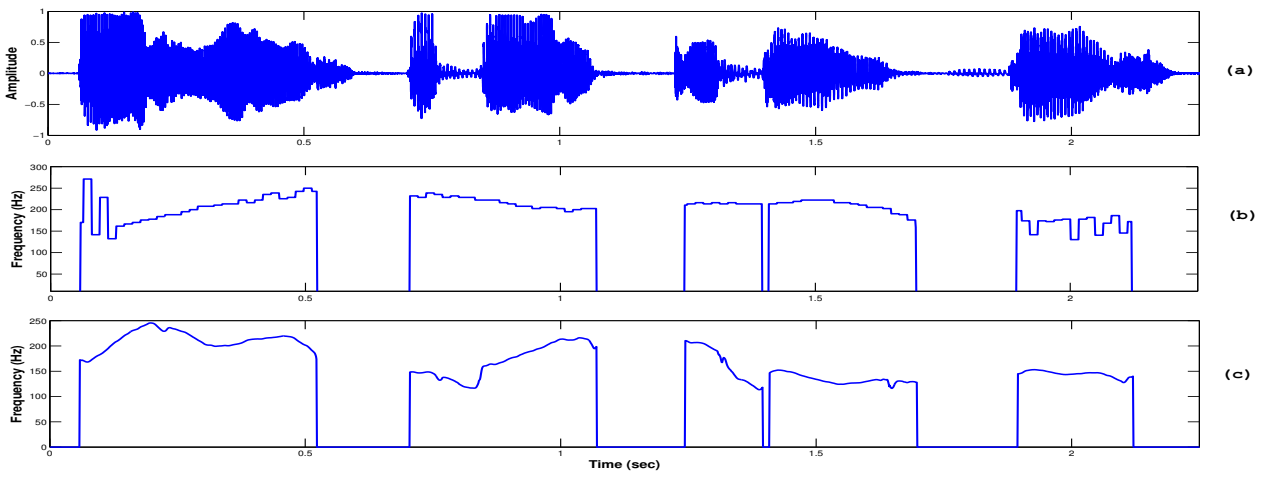

Figure 1.5: F0 extraction of STRAIGHT 1.5(c) and its comparison with conventional method 1.5(b) for speech segment 1.5(a)

Aperiodicity measurement

Aperiodicity measurement is introduced in the STRAIGHT to improve the buzziness present in the earlier version of STRAIGHT, which was based on only impulse excitation [25]. Aperiodicity measure is defined as the ratio between the lower and upper smoothed spectral envelope. Spectrogram of aperiodicity measurement computed from STRAIGHT method is plotted in Figure 1.6. From this figure, it can be observed that some aperiodic components are present even in the case of voiced sounds. The aperiodicity measurements estimate the amount of harmonic component in relation to non-harmonic component in the signal.

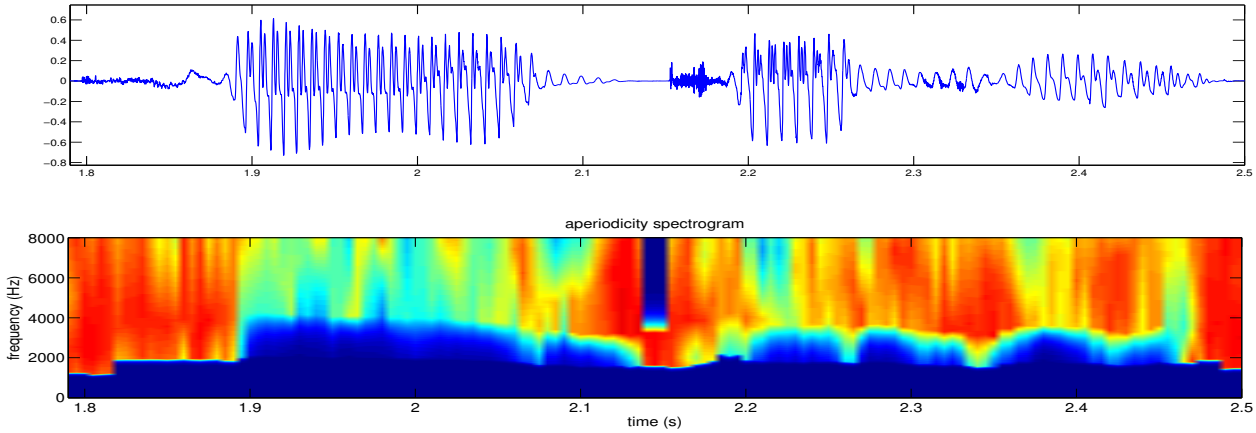

Figure 1.6: Aperiodic component present in glottal and non-glottal region

Smoothed spectral envelope

STRAIGHT spectrum of the speech signal is computed by using a pitch-adaptive STFT analysis. It uses two compensatory time windows to calculate the spectrogram. First, a convolution of the speech signal with a pitch-adaptive window is performed. The main objective of the convolution of the speech signal with this window is to smooth the spectrogram in the frequency domain. The periodicity of the speech signal in the time domain produces phase interference in the spectrogram and it is compensated by the compensatory window. Finally, the power spectrum of the speech signal is represented as a weighted squared sum of the power spectrum obtained from both the windows.

Speech re-synthesis

STRAIGHT synthesis is based on the frame-by-frame processing by designing a mixed excitation signal. The excitation signal is based on the F0 and aperiodicity measurements accounts the harmonic information. The harmonic impulse train is convolved with an all-pass filter having a fixed group-delay

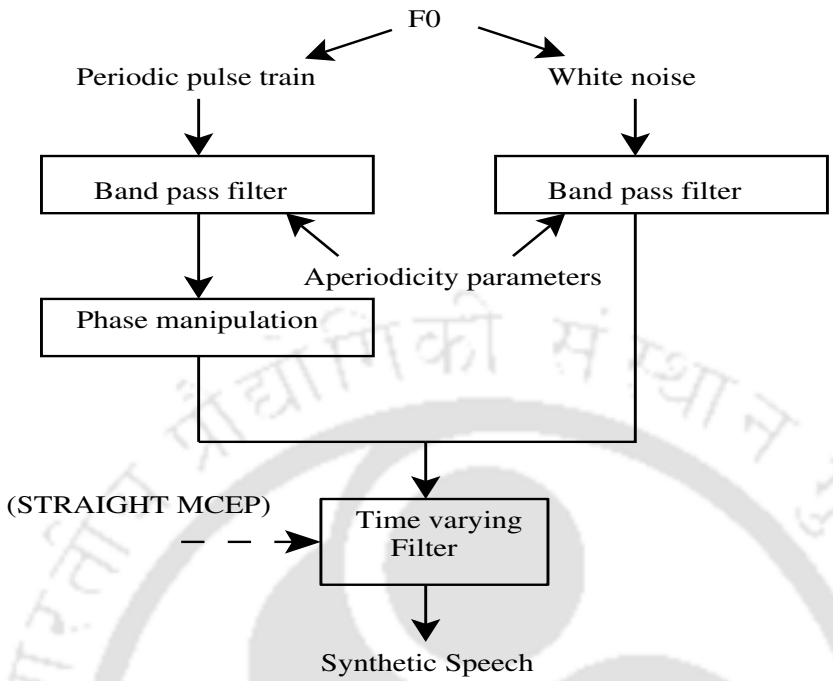

Figure 1.7: Block diagram of STRAIGHT synthesis

response to reduce the buzziness. The generated mixed excitation signal is convolved with the minimum phase MLSA filter derived from the MCEP features. Finally, the pitch-synchronous overlap-add algorithm is applied to the synthesized frames to get the final speech signal. The synthesis process is illustrated in Figure 1.7. The analysis and synthesis framework of STRAIGHT give flexibility in modifying various parameters like instantaneous F0 trajectories, aperiodic components, phase components, and smoothed vocal tract envelope without any degradation in synthesis quality.

1.3.2 Glott Vocoder

Alku *et al.* proposed iterative adaptive inverse filtering (IAIF) method. This method gives an approximated glottal source signal from speech [28]. The Glott vocoder is motivated by the fact that designing the excitation signal by using the approximated glottal pulse. Glott vocoder for SPSS is proposed by Raitio *et al.* in 2011 [27]. The advantage of using this method is to use the approximated glottal pulses in designing excitation signal for synthesis framework, which provides more natural synthesis quality compared to the impulse train excitation.

In this method, the speech signal is high-pass filtered with a cut-off frequency of 70 Hz, and energy is computed for each windowed signal. Next, IAIF algorithm [28] is applied to each frame. The algo-

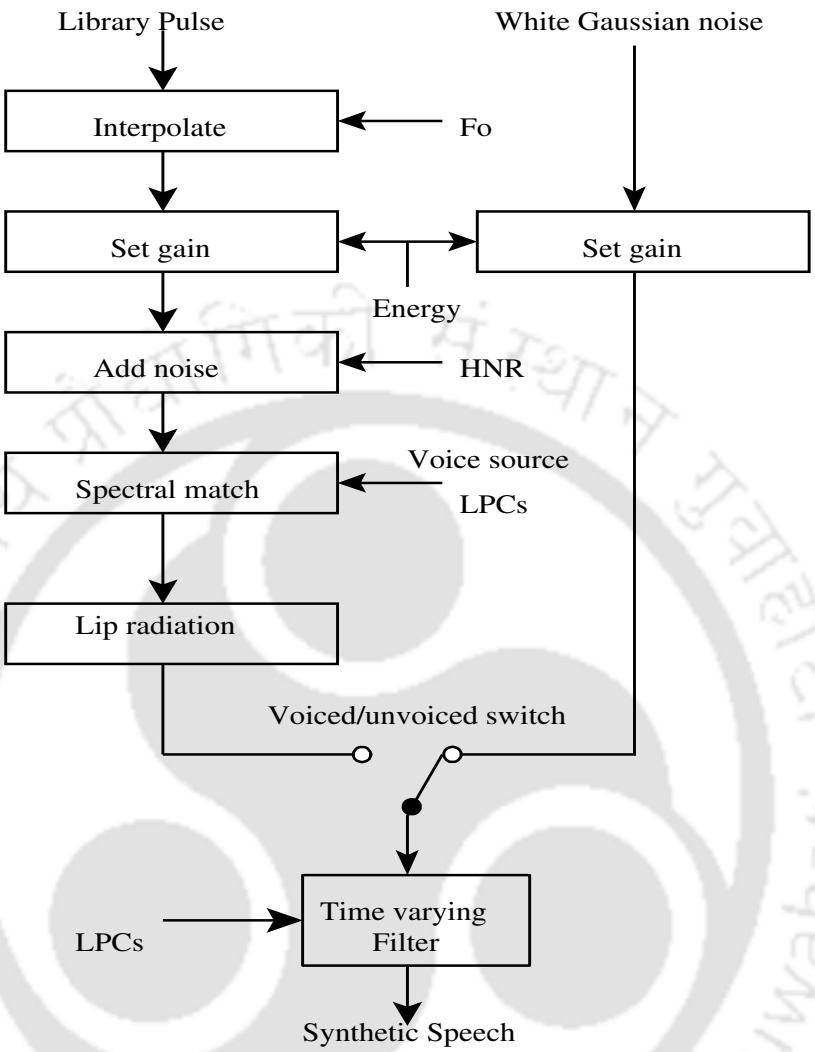

Figure 1.8: Synthesis block diagram of Glott vocoder

rithm gives linear prediction coefficients (LPC) for both the vocal tract spectrum and the waveform representation of the voice source signal. The LPC spectral envelope of both voice source and vocal tract is converted into a line spectral pairs (LSP) representation since LSP is more stable representation when compared to LPC [29]. The glottal flow waveform is used for the acquisition of the F0 value as well as the harmonic-to-noise ratio (HNR) values for a predetermined amount of sub-bands of frequency. In Glott vocoder, instead of conventional linear prediction analysis, weighted linear prediction is used for the estimation of the vocal tract filter response, since, weighted linear prediction analysis mitigates the effect of the harmonic peaks on the estimation of formants [30]. To model the glottal source parameters in SPSS, the glottal pulse is stored as pulse library. The construction of the pulse library is performed by taking a segment of speech and applying the IAIF algorithm for

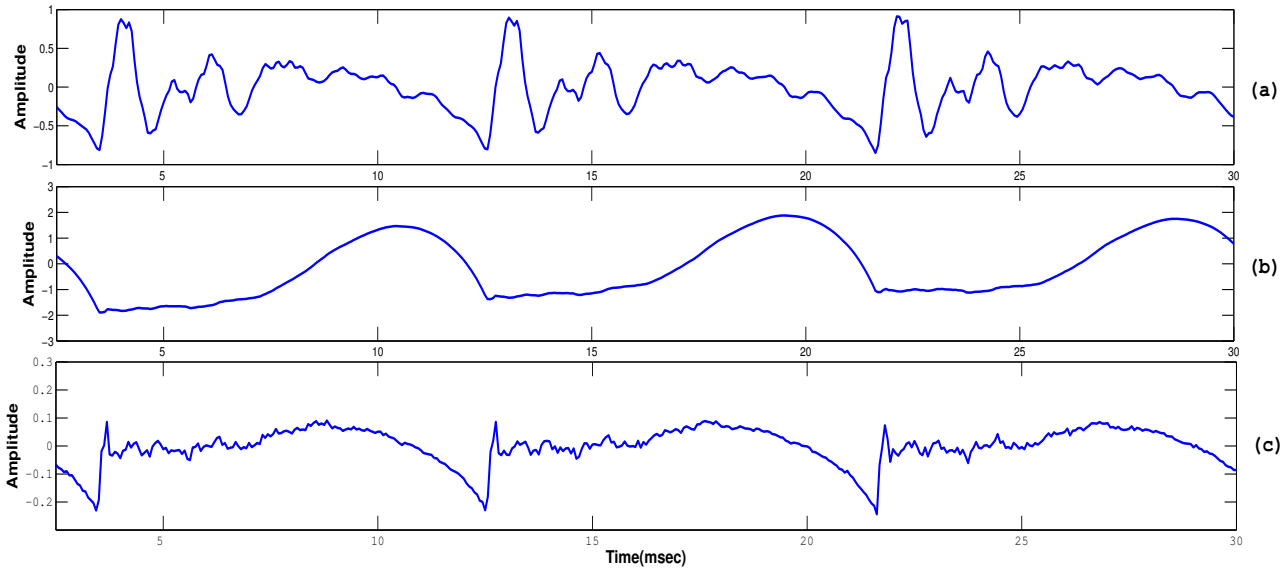

Figure 1.9: Glottal pulse 1.9(b) and glottal flow derivative 1.9(c) extracted using IAIF for speech segment 1.9(a)

each segment to get the glottal excitation signal. Figure 1.9(b) and (c) show the glottal pulse and glottal flow derivative obtained from IAIF for a three cycle speech segment. From this glottal pulse signal, the glottal closure instants (GCI) are detected, and GCI centered two-period long segments are extracted. The obtained glottal pulses are normalized in energy and saved in the pulse library with their voice source parameters.

The synthesis phase is shown in Figure 1.8. The glottal pulse is selected from the pulse library by minimizing the target and concatenation costs by using the Viterbi algorithm [31]. The target cost is composed of the root mean square error (RMSE) between the voice source parameters of the pulse and the parameters generated by the HMM. The concatenation cost is computed as the RMSE between the consecutive down-sampled pulse waveforms in each voiced frames. Minimizing the concatenation cost ensures that adjacent pulse waveforms do not differ substantially from each other. Once the pulse is selected, it is re-sampled and normalized according to F0 and log energy parameters. Further, noise is added based on HNR values to get mixed excitation. This excitation is passed through linear prediction (LP) filter to get synthesized speech. It was reported that the synthesis quality of the Glott vocoder showed an improvement when compared to the STRAIGHT version of SPSS [32]. The main reason for this is due to the using glottal pulses for generating excitation signal instead of using

impulse based excitation [32].

1.4 Motivation for the present work

Speech synthesis using SPSS is inevitable due to the generalized framework and its capability for adaptation of models [33]. However, shortcomings of SPSS is with the issues like vocoded synthetic quality due to the poor source-filter features, over-smoothing of the parameters in the HMM, and simplified vocoder model [9]. Based on the results of STRAIGHT and Glott vocoder work, naturalness and intelligibility can be improved by using better source-filter features and improved vocoder framework. Hence, in this thesis, emphasis is given to the better extraction of acoustic features, which represents the source-filter modules. Then model those acoustic features in the SPSS framework to get improved synthesis quality.

For addressing the issues mentioned in Section 1.2 for SPSS, the motivation for the present work in this thesis is the following:

- In general, most of the state-of-the-art algorithms use only one aspect of the glottal signal for voicing decision such as pitch or voicing strength. However, there are other parameters of voiced sounds such as asymmetrical nature (speed quotient), duty cycle (open quotient), sub-segmental variations. Hence, to characterize these aspects of speech production mechanism, different representation may be required. These representations along with existing parameters can improve the voicing decision.

- In SPSS, MCEP are computed from the STFT spectrogram to represent vocal tract. However, the STFT spectrogram still retains glottal source component due to fixed segmental analysis and windowing effect. Hence, in the STFT spectrogram, fluctuations in both time and frequency domain will be present and results in the lesser intelligibility of synthesized speech. Therefore, an alternative method is required to capture coarticulation aspects of speech production mechanism, which completely removes the effect of the source signal in vocal tract spectrum and gives smoothed representation of vocal tract. In addition, the dynamic range in MCEP features of synthesized speech is reduced due to statistical modeling, so there is a need for post-processing technique to improve the dynamic range of MCEP.

- The naturalness or speaker characteristics are mainly attributed by the glottal source modeling in speech synthesis. The naturalness of the current SPSS is lagging behind concatenative speech

synthesis mainly due to magnitude only representation of source signal by ignoring the phase and aperiodic nature of source signal. The codebook for source signal computed from linear prediction (LP) residual or glottal pulse can be used to preserve the glottal pulse shape, aperiodic, and phase components at a cost of higher memory. Hence, there needs to be a mechanism to compactly preserve the LP residual signal or an alternate mechanism to preserve the aperiodic and phase component to improve the naturalness.

Motivated by these observations, the primary objective of the work presented in this thesis is to demonstrate the significance of having a better suprasegmental, system, and source features to depict the speech production mechanism, which in turn improve intelligibility and naturalness of SPSS.

1.5 Organization of the Thesis

To address the issues mentioned in the previous section, this thesis work is organized into eight chapters. The content of each chapter is summarized as follows:

- Chapter 2 reviews several existing methods for improving the naturalness and intelligibility of SPSS. The review is broadly divided into three sections: different voicing detection algorithms, smoothed spectral envelope extraction from different methods, and different source modeling techniques, which are used in SPSS. Summary of the review and the scope of this thesis work are also discussed in this chapter.

- In Chapter 3, glottal activity region detection algorithm is proposed using the combination of glottal source features. The features used are the strength of excitation (SoE), normalized autocorrelation peak strength (NAPS), and higher order statistics (HOS), which represent the different aspects of the glottal activity. The proposed glottal activity regions detection is compared with other state-of-the-art algorithms.

- In Chapter 4,the significance of glottal-activity features for speech synthesis is shown in the analysis by synthesis framework. The evidence computed from SoE, NAPS, and HOS glottal activity features is further enhanced using the support vector machine (SVM) algorithm. Next, the HMM based speech synthesis system is developed using proposed glottal activity features and then compared with other state-of-the-art algorithms to demonstrate the merits of using glottal activity features for synthesis task.

- In Chapter 5, a 2-D based analysis for obtaining smoothed vocal tract envelope using Riesz transform is proposed. Compared to the conventional approach, the proposed 2-D based approach captures the coarticulation mechanism effectively. Further, the usefulness of Riesz transform for the computation of voicing decision, F0, and aperiodicity map is shown. Finally, the significance of different features computed from Riesz transform is shown for improving the naturalness and intelligibility of SPSS.

- In Chapter 6, integrated linear prediction residual based source modeling is proposed for SPSS. A method is proposed for showing the significance of epochs and the instants around the epochs for synthesis. Further, new representations for aperiodic and phase components present in the glottal activity region is explored and its significance is shown for SPSS.

- In Chapter 7, a combination framework using suprasegmental, system, and source feature for SPSS is demonstrated for improvement in intelligibility and naturalness of synthetic speech. Further, proposed framework is tested for two Indian languages and also in deep learning based statistical model.

- Chapter 8 summarizes the work presented in this thesis, highlights the main contributions of the work and gives some directions for future research.

2

Acoustic Features for Statistical Parametric Speech Synthesis: A Review

Contents

2.1 Introduction 18
2.2 Acoustic feature modeling using different SPSS techniques 19
2.3 Suprasegmental features 22
2.4 Vocal tract Spectral model 25
2.5 Different source models 29
2.6 Other Signal Processing Models 36
2.7 Summary and Discussion 40
2.8 Organization of the Present Work 43

2.1 Introduction

Statistical parametric speech synthesis (SPSS) has rapidly overtaken the traditional unit selection synthesis (USS) both in intelligibility and flexibility [5,9]. USS approach is based on the concatenation of recorded speech units provides naturally sounding synthetic speech [7]. However, in USS method, modeling out of context acoustic space is challenging and gives poor quality for out of context words. It also requires a separate training mechanism to include those out of context acoustic space for accommodating speaking style and emotions [33]. SPSS using Hidden Markov models (HMM) and deep learning is extensively used in recent days. Statistical models using HMM is commonly implemented using software known as HMM based speech synthesis system (HTS) [5, 9, 34]. To build the deep learning based speech synthesis system, recently Merlin toolkit is introduced [35]. In Merlin system, linguistic features are taken as input and tried to predict acoustic features, which are then passed to a vocoder to produce the speech waveform.

It has overcome some of the problems present in USS. In SPSS, the speech signal is realized as a different set of features like fundamental frequency (F0), voicing decision, Mel cepstral coefficients (MCEP) for every phoneme. These features are input to HMM and they are trained together using phone specific context-dependent statistical models [36]. Speech is synthesized using the vocoder framework from the generated parameters for a given text. SPSS provides an edge over traditional synthesis approaches in terms of modeling broader acoustic space with significantly lower memory and good intelligibility. Further, SPSS gives flexibility to users to change the speaking style or emotions from modeled acoustic space. Nevertheless, one of the limitations of SPSS is that the quality of the synthesized speech is not natural when compared with USS [5].

There are multiple factors resulting in the degradation of synthesis quality of SPSS, which includes, inadequate representation of acoustic features, over-smoothing of acoustic features while modeling in a statistical framework, and a simplified vocoder model to generate synthetic speech. This review is focused on the role of various acoustic features and their influence on the perceptual quality of synthetic speech. This is because the errors in feature extraction lead to errors in the training process and subsequently leads to degradation in synthesis quality. The second motivation is to understand the role of different acoustic features for enhancing the prosody, naturalness, and intelligibility of SPSS. The main features employed for synthesis are broadly categorized as suprasegmental (F0 + voicing decision) features, excitation source features, and vocal tract system features, which depicts

prosody, naturalness, and intelligibility of speech, respectively. This classification is based on the feature extraction procedure and its usage in the analysis/synthesis framework. Hence, the exhaustive review is made individually on these three acoustic features.

In the recent literature, many state-of-the-art analysis/synthesis methods have been successfully integrated into the framework of SPSS. In this review chapter, their performance in terms of feature level is studied. Since the feature extraction can be added in three levels like suprasegmental, system, and source to vocoder framework, it would be beneficial to know the appropriateness of each of these three level features for SPSS modeling. Further, the aforementioned knowledge could lead to new refinements in analysis/synthesis framework. Moreover, to know the limitation and benefits of each of these features for the improvement in the perceptual quality of vocoder is studied. Also, a review of different modeling techniques present in SPSS like HMM, deep learning model, and hybrid model to add these acoustic features is given.

The structure of this chapter is as follows: a brief survey of modeling techniques present in SPSS is presented in Section 2.2. The different features suprasegmental, vocal tract, and source modeling are discussed in Section 2.3, 2.4, and 2.5, respectively. In Section 2.6, other signal processing models used in speech synthesis framework is explained. Section 2.7 presents the overall discussion on these three categories of features. The organization of the present work is given in Section 2.8.

2.2 Acoustic feature modeling using different SPSS techniques

SPSS has extensively researched speech synthesis method in the last decade. Despite the fact that SPSS does not give a speech with as natural sounding as that of USS method, its flexibility and robustness make it an attractive system for many applications. Further, the perceptual quality of SPSS has boosted a lot through the recent years. There are mainly three types of SPSS techniques for modeling acoustic features, the first is based on HMM, the second is based on deep learning models and the third is based on Hybrid model.

2.2.1 Hidden Markov model

HMM is a dominant statistical signal processing method for generation and discrimination of time series data [37]. HMMs have been successfully applied in speech processing applications such as speech recognition, enhancement, and synthesis. In HMM based speech synthesis, the speech features computed by analysis/synthesis framework are modeled using context dependent left-to-right phoneme

HMMs. The observation input vector for HMM will be continuous-valued speech features, and the output probabilities of each HMM state are modeled by single multi-variate Gaussian distributions.

Tokuda *et al.* integrated HMM into speech synthesis system [5]. The process of speech synthesis using HMM involves two steps: training stage and synthesis stage. In the training stage, spectral and excitation features are computed from the speech data. In the basic version of HTS, spectral features normally contain MCEP and their dynamic features (delta and delta-delta) [36]. Excitation features usually contain the pitch frequency (F0) and its dynamic features, where F0 is modeled as logarithmic of pitch frequency (logF0) [8]. Both the spectral and source features are trained in HMM by applying the expectation-maximization algorithm as follows [11]:

$$\lambda_{max} = arg_{\lambda} max\{p(O|\lambda, W)\} \tag{2.1}$$

where λ is produced HMM, O is a collection of training data, and W is word sequence analogous to O.

In the synthesis stage, the input text is transformed into a sequence of context-dependent phonemes. The speech parameter generation algorithm produces the vocal tract and source parameters from the HMM using maximization criteria [38]. The observation vectors o corresponding to different speech features formed for a text input w to be synthesized from the set of predicted models λ_{max} is given as

$$o_{max} = arg_{o} max\{p(o|\lambda_{max}, w)\} \tag{2.2}$$

Consequently, the speech waveform is generated from the produced vocal tract and source features by employing Mel-log Spectrum Approximation (MLSA) filter convolved with the excitation signal [13].

SPSS using HTS has a lot of advantages [5]. The footprint of the synthesis module is normally very low, implying fewer than 2 MB [10]. Further, HTS has a larger coverage of the acoustic space because speech is produced from statistical models. SPSS is somewhat simple to model viewing there are fewer tuning parameters than the USS. HTS is flexible due to the generating speech by an analysis/synthesis framework, which allows the exclusive control of the suprasegmental (F0, voicing decision, and duration), vocal tract, and source components. Lastly, language adaptation for low resource language can be dome using HTS [39]. Nevertheless, one of the foremost restriction of HTS is a lack of natural speech quality when compared to USS.

2.2.2 Deep learning model

Another current model in SPSS is the deep learning representation. In the last decade, the training of deep architectures in neural networks was supposed to be unsettled, however, the latest algorithms, improved computing capability, and large data have yielded extraordinary outcomes in speech recognition. Presently, similar methods are utilized for speech synthesis with encouraging outcomes [40–42]. There are deficiencies in the current decision-tree based clustering used in HMM. The data fragmentation happens with the decision-tree based clustering, and therefore it is ineffective for describing intricated dependencies among linguistic and acoustic parameters. On the other hand, deep learning approaches are more effective than HMM approach. Popular deep learning approaches, such as the deep belief network (DBN), the deep neural network (DNN), and the long short-term memory (LSTM) based recurrent neural network (RNN) have given encouraging outcomes in joining with HMMs or without [43–47]. Usually, deep learning is likely to result in over-fitting with small corpora, and thus a large amount of data is needed for successful training. The deep learning system is growing fast, and seemingly new approaches will get applications in speech synthesis. This gives enhanced synthesis quality and improved adaptability.

DNN based acoustic model provides better feature-cluster and cluster-feature mapping. The results showed that the DNN based spectral models show the improvement in the naturalness of SPSS. DNN based methods can be categorized into 2 types: cluster-to-feature mapping applying deep generative models and input-to-feature mapping applying deep joint models [48]. The cluster-to-feature mapping is similar to hidden Markov model-Gaussian mixture model (HMM-GMM) based approach, where inside the cluster deep learning models are implemented to get the generative features. In another case, deep learning models directly applied on the input linguistic or contextual features to get the output acoustic features. Finally, in both the methods, acoustic features are generated in the form of MCEP or spectral envelope. It was reported that synthesis quality from these generated features is better than HTS [48].

Source modeling using DNN was first introduced by Raitio *et al.* [49]. One of the advantages of this method is that it bypasses the difficulty of irregularly selecting improper glottal-pulses. Moreover, it allows the capability to modify the attributes of source signal [50]. This approach applies DNN to represent the context-dependent variation of the glottal pulse with a mapping from acoustic parameters to the relevant glottal pulse, where many acoustic parameters are modeled by an HMM. This technique

was proved to be similar in quality to a single glottal pulse based source model. Further, the work in [51] was proved to give reliable voice quality reproduction by synthesizing higher quality Lombard speech contrasted to a principal component analysis (PCA) based source model. However, the speaker identity was somewhat inferior opposed to the PCA based approach. In the original variant of the glottal pulse model, only speaker-dependent voice source DNN was developed. In the later version, a multi-speaker model also trained using DNN and utilized for the synthesis of multiple speakers by Suni *et al.* [52]. The perceptual quality of DNN based glottal pulse model presents an enhanced model of the source signal in terms glottal pulse shape and phase properties. However, the modeling of the aperiodic element in DNN still remains a hurdle. Further, the versatility of deep learning system is yet uncertain because the similar adaptation techniques used in the HMM cannot be applied directly. There are techniques for adapting DNNs, still, they have not been implemented extensively to speech synthesis.

2.2.3 Hybrid Synthesis

There are some works in SPSS by utilizing the advantages present in both statistical framework and USS approach to get natural speech and the approach was popularly known as Hybrid synthesis [53–55]. The main motivation of this modeling is to use the original speech segments from USS to get naturalness and statistical framework for selecting these segments to get the smoothed joining at the concatenation point. In the literature, there are some approaches which apply hybrid speech synthesis. One method was utilizing features from HMM for choosing the natural speech units or modifying the prosodical aspects of USS voice [54,56,57]. In another method, features from HMM are employed to smooth the joints within the speech units chosen from USS [55] to have enhanced speech trajectory. Tiomkin *et al.* [53] proposed another approach, where decision among natural segments and the parameters produced from HMMs are chosen using the Viterbi algorithm to have the least distortion among chosen units. Even though these methods generate natural speech quality, there are some complexity involved in combining two different approaches.

2.3 Suprasegmental features

The suprasegmental features refer to continuously distributed fundamental frequency (F0) and discrete voiced/unvoiced decision, which usually changes values for longer frames. In SPSS, to accommodate both continuous F0 and discrete voicing decision, Tokuda *et al.* introduced multi-space

distribution [58] model, where both F0 and voicing decision are modeled together. Recently, Yu *et al.* proposed continuous distribution model for suprasegmental features where both F0 and voicing decision are modeled independently [59]. Details of these methods are given below.

2.3.1 Multi-space distribution model

In the multi-space distribution (MSD) model, F0 is expected to be continuous in the voiced region and discrete in the unvoiced region. It is modeled for a state (S) as follows

$$P(\mathrm{F}_+ = \mathrm{F}|L,S) = \begin{cases} \mathcal{N}(\mathrm{F};\mu_S,\sigma_S), & L = \mathtt{V} \\ 0, & L = \mathtt{U} \end{cases} \tag{2.3}$$

$$P(\mathrm{F}_+ = \mathrm{NULL}|L,S) = \begin{cases} 0, & L = \mathtt{V} \\ 1, & L = \mathtt{U} \end{cases} \tag{2.4}$$

where $\mathcal{N}$ is a Gaussian density with mean μ_S and variance σ_S, $\mathrm{F} \in (-\infty, \infty)$ represents the real F0 value, and $L \in \{\mathrm{U},\mathrm{V}\}$ is the voicing label. According to the hypothesis of continuous F0, the voicing label V and the NULL symbol value cannot be observed at the same time. Similarly, the unvoicing label U and the real F0 value cannot occur together. The multi-space distribution model provides good quality speech when the voicing decision is accurate. However, if the voicing decision is not accurate, then it affect the F0 model and as well as voicing decision as both of these suprasegmental features are modeled together.

2.3.2 Continuous distribution

In the continuous distribution model, F0 and voicing decisions are modeled in two separate streams of HMM state [59]. Therefore, both the features do not depend on each other and results in better accuracy, particularly for voicing decision. There are some works proposed to improve the voicing decision using continuous distribution for SPSS [59–63]. In [59], F0 and voicing labels are modeled independently in separate streams to make the voicing decision independent of F0. In the globally tied distribution method [60], F0 samples for unvoiced frames are interpolated by resampling technique and smoothed by a low-pass filter. Both F0 and its derivatives are modeled in a single stream. Every state is modeled by a Gaussian mixture model (GMM) with two Gaussian distributions: one Gaussian is for voiced frames and another one is for unvoiced frames. The voicing decision in synthesis is

computed using voicing mixture weight. In the literature, many methods are used for measuring these suprasegmental features.

2.3.3 Acoustic features for F0 and voicing decision

In SPSS, the F0 or pitch parameter was used to the model excitation signal representing the prosody of the speech. In the initial version of SPSS, to get the F0, robust algorithm for pitch tracking (RAPT) algorithm was used. The RAPT algorithm is based on the autocorrelation analysis [64, 65]. This conventional F0 extraction method provides stepwise F0 trajectories due to the frame based analysis, which is plotted in Figure 1.5(b). This stepwise structure says that signal is periodic over a frame size. However, pitch period of speech keeps changing slowly with each glottal cycle even within a voiced frame. Hence, Kawahara *et al.* proposed the TEMPO method [65] for F0 extraction. It is based on the concept of instantaneous frequency to get the smoothed trajectories F0 [18]. The instantaneous frequency is derived from the signal $s(t)$ by applying continuous wavelet transform.

The continuous wavelet transform is defined as:

$$D(u,\tau_0) = \frac{1}{\sqrt{\tau_0}} \int_{-\infty}^{\infty} s(t) g_{AG_\tau}\left(\frac{t-u}{\tau_0}\right) dt \tag{2.5}$$

where g_{AG_τ} is the wavelet function and τ_0 is a scaling factor of the wavelet, τ represents the different channel, and analyzing wavelet is determined by

$$g_{AG_\tau}(t) = g_\tau(t-\tau/4) - g_\tau(t+\tau/4) \tag{2.6}$$

$$g_\tau(t) = e^{-\pi\left(\frac{t}{\eta\tau}\right)^2} e^{-j\frac{2\pi t}{\tau}} \tag{2.7}$$

where $\eta > 1$ is a variable describing the frequency resolution of the wavelet function. The wavelet used in the TEMPO algorithm based on a Gabor function. Instantaneous frequency is located where the fundamentalness component is least. The fundamentalness map $M(t,\tau_0)$ is determined as

$$M(t,\tau_0) = -\log[\int_\Omega \left(\frac{d|D|}{du}\right)^2 du] + \log[\int_\Omega |D|^2 du] - \log[\int_\Omega \left(\frac{d^2 arg(D)}{du^2}\right)^2 du] + 2\log\tau_0 \tag{2.8}$$

The first and the third terms represent the magnitude of the AM and FM component, respectively. The second and the fourth term represents the normalized component of amplitude modulation (AM) and frequency modulation (FM), respectively. Extracted F0 from TEMPO method is plotted in

Figure 1.5(c), showing smooth F0 trajectories.

These methods provide voicing decision also along with F0 parameter. In addition, various refinements have been introduced for F0 computation from degraded speech recorded in realistic conditions. These cover extraction of pitch in adverse situations by utilizing autocorrelation of the Hilbert envelope of linear prediction (LP) residual [66, 67]. Further, some of the recent algorithms like yin, praat, get_F0, swipe *etc.* have also gives very good F0 estimation [17, 68–70]. Also, some of the algorithms focused on the instantaneous F0 estimation, which gives smoothed F0 trajectories compared to a conventional method based on the block based processing [18, 71].

Voicing decision methods are divided into time-domain and frequency-domain categories. Typically time-domain approaches estimate the acoustic nature of voiced sound such as energy, periodicity, zero crossing rate, and short-term correlation. The frequency-domain approaches measure harmonic components by decomposing speech signal using the Fourier transform or wavelet transform [20, 67, 72–74]. In both these approaches, voicing decisions are taken by comparing to a threshold chosen empirically. Hence, the performance of these methods depends critically on the threshold. Also, most of these measures are affected by noise and performances decline with a decrease in the signal to noise ratio. However, the main advantage of the signal processing based F0 and voicing decision is that these methods are data independent and do not require any training mechanism for extraction of these features.

To improve the accuracy of voicing decision and avoid using the threshold, statistical methods such as HMM, GMM, neural network model, and DNN are used for voiced/unvoiced classification [75, 76]. These methods do not depend on the threshold. However, they need discriminative features for training. The statistical methods such as GMM and multilayer-perceptrons based voicing decision was already attempted in SPSS [62, 63]. These classifiers helped in improving the accuracy of voicing decision. However, all these methods basically focus on better modeling of voicing label using existing features in HMM rather than using new features to improve the accuracy of voicing decision.

2.4 Vocal tract Spectral model

In literature, there are various methods proposed to model the vocal tract representation. Two mainly involved methods in SPSS are linear prediction (LP) and cepstrum based methods. In this section, an overview of these two methods, the procedure for adapting to SPSS, and new advancements

that have taken place are discussed.

2.4.1 Linear prediction model

LP analysis is a generally employed spectral estimation approach models the resonances of speech by the poles of the LP model [77]. The fundamental theory of the model is that most of the speech sounds have a resonance structure and can be modeled by the poles. This theory is adequate for most speech sounds, except for nasal sounds (consist of anti-resonance) or some fricative sounds. Nevertheless, by raising the LP order, anti-resonances may be estimated. Hence, it is necessary to choose a relevant order because the too lower order can not accurately model all the formant frequencies and too high order may model the harmonic parts of the excitation signal. The effect of the excitation signal harmonic component is critical particularly for high-pitched speech. To mitigate this influence, weighted linear prediction, and pitch synchronous analysis was employed in the literature [78, 79]. For instance, weighted linear prediction is employed for mitigating the biasing influence of the harmonics in shouted speech. In weighted linear prediction, more emphasis is given to the closing phase of the glottal cycle by a weighting function. This enables the effective estimation of the poles with regard to formants of the speech spectrum [78].

LP coefficients are not suitable for statistical modeling due to the stability issue, therefore they are transformed into other forms of representation. There are different representations that can be employed, of which the extensively applied is the line spectral pairs (LSP) representation [29]. This representation gives excellent interpolation and smoothing properties required for the modeling.

2.4.2 Cepstrum based methods

Another familiar vocal tract modeling based on the homomorphic processing of speech is Cepstral analysis. The cepstrum based analysis provides modeling of both poles and zeros present in the speech spectrum [80]. Normally, Mel-cepstrum is employed instead of cepstrum to improve the frequency resolution resembling that of human perception. Synthesis by Mel-frequency cepstral coefficients requires the approximation of filter response. Hence, Imai *et al.* proposed an iterative procedure to approximate the filter response and it is known as MLSA filter. The coefficients obtained from this technique is called as MCEP [81]. Further, this iterative method of cepstral analysis is generalized to get unified approach and coefficients computed from this method is known as Mel-generalized cepstral coefficients (MGCEP) [16]. However, in this thesis, MLSA filter is used. Hence, detailed review of

MLSA filter is given in next section.

2.4.3 MLSA filter

The vocal tract response from the MCEP was approximately estimated using MLSA filter [81]. The true spectrum of MLSA filter for M^{th} order MCEP $\tilde{c}[m]$ were given by

$$H[z] = exp\left(\sum_{m=0}^{M} \tilde{c}[m]\tilde{z}^{-1}\right) \tag{2.9}$$

where

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, |\alpha| < 1 \tag{2.10}$$

is the Mel-warped frequency characteristics of all-pass function. Here, α is a parameter used to get frequency resolution similar to human ear.

The above representation is in Mel-scale. Accordingly, MCEP cannot be utilized directly for synthesis and frequency unwarping has to be done from Mel-scale to normal scale. The approximated filter response is obtained through the linear transformation of MCEP [13]. The above filter equation is not fractional and not realizable. Hence, in the MLSA filter, modified filter coefficients are obtained using the Pade' approximation. The modified filter coefficients decay in the same order as in the case of MCEP. This implies that filter coefficients can be quantized and its statistical properties remain same as that of MCEP. Therefore, it is utilized as a parametric representation of speech.

Usually, MCEP of 20-60 coefficients are utilized for HTS, variations in the coefficients depend on the sampling rate. Higher coefficients normally result in enhanced quality, however, this may also lead to modeling the harmonic components. Hence, special methods are required to counter the biasing influence of the harmonic peaks. The most common method to prevent the harmonic effect both in time and frequency is using pitch adaptive time-frequency smoothing method like Speech Transformations and Representations using Adaptive Interpolation weiGHTed spectrum (STRAIGHT) [18].

2.4.4 STRAIGHT Spectrum estimation

STRAIGHT spectrum was proposed by Kawahara *et al.* [18, 65] in 1997. This method is driven by the requirement for a flexible and high-quality analysis-synthesis approach. STRAIGHT mainly involves removal of periodicity influences in the spectral analysis by employing the adaptive windows. The speech signal is usually processed by fixed window length, and length of the window for analysis is equivalent to the integer multiple of the fundamental period. Yet, this procedure is not useful for

natural speech signal. This is because the F0 of the speech signal keeps changing every time, and the fine discontinuities with the windowing make the analysis very susceptible to such lesser variations.

In STRAIGHT, periodicity effect present in the spectrum was eliminated using two processes. First, by utilizing the pitch-adaptive window ω_p to remove the temporal interference around peaks and then using a compensatory time window ω_c to remove the temporal interference at the valley in the spectrogram. The window ω_p is given by

$$\omega_p(t) = e^{-\pi(\frac{t}{\tau_0})^2} * b(t/t_0) \tag{2.11}$$

where t is the time variable, t_0 is the fundamental period, and $b(t)$ is the second order cardinal B-spline function defined as follows

$$b(t) = \begin{cases} 1 - |t|, |t| < 1 \\ 0, otherwise \end{cases} \tag{2.12}$$

The compensatory window ω_c is given by:

$$\omega_c(t) = \omega_p(t) \sin\left(\pi \frac{t}{t_0}\right) \tag{2.13}$$

where $\omega_c(t)$ is sinusoidally modulated window and gives maxima, wherever the initial spectrogram has a valley. Smoothing in the frequency domain is obtained by the employment of the second order cardinal B-spline function, which interpolates the spectral samples within the magnitude spectrum.

The initial and compensatory magnitude spectrograms $P_o(\omega, t)$ and $P_c(\omega, t)$ are computed using the windows ω_p and ω_c, respectively. The spectrograms are added into the final spectrogram $P_r(\omega, t)$ given by

$$P_r(\omega, t) = \sqrt{P_0^2(\omega, t) + \xi P_c^2(\omega, t)} \tag{2.14}$$

where ξ is blending factor to be chosen empirically.

STRAIGHT method provides smooth spectral envelope extraction and preserves both peaks and valleys of the spectrum well, by getting rid of the spectro-temporal variations using pitch adaptive and sinusoidal windows. Zen *et al.* integrated the STRAIGHT spectrum into HTS with MCEP and synthesis quality improved drastically [15]. However, there are issues in the STRAIGHT method such as over-smoothing of the spectral envelope, and the need for an accurate pitch estimation algorithms for extracting smoothed envelope. Moreover, the procedure to extract these parameters are complicated and needs nonlinear transformation techniques and many tuning parameters to get better performance.

Recently, Kawahara *et al.* [82] proposed TANDEM-STRAIGHT to alleviate some of the issues present in the legacy STRAIGHT implementations. In the TANDEM-STRAIGHT, pitch adaptive tandem windows are used for eliminating periodic temporal fluctuations. The computed spectrum from the tandem windows is temporally stable design for periodic signals and provides synthesized speech that is almost equal to original speech. Nevertheless, even for TANDEM-STRAIGHT, pitch adaptive windows are required. Hence, an alternate spectral envelope estimation method is needed, which gives accurate spectral envelope without any prior pitch estimation.

2.5 Different source models

Source modeling is one of the main modules of the vocoder to enhance the naturalness of the SPSS. In the base version of the HTS vocoder, Tokuda *et al.* used periodic impulse train located around the glottal closure instants (GCI) as the source signal. In this model, only F0 and MCEP parameters are used, which represent, excitation source signal and vocal tract response, respectively. These parameters are trained in HMMs. During synthesis, the source-filter model is used where voiced frames are excited by the impulse positioned according to the F0 interval. The unvoiced frames are excited by white-Gaussian noise. The generated excitation signal is convolved with MLSA filter response to get the synthetic speech. The quality of this simple excitation source model is affected by a buzziness owing to the same excitation strength and constant pitch period over a frame. Consequently, different approaches reported in the literature, that can be broadly classified into mixed band excitation, residual modeling, and glottal source model.

2.5.1 Mixed Band Excitation

In order to address the aforementioned problems present in impulse/noise excitation, several approaches have been reported in recent years and mixed multi-band excitation is one of the method. It uses extra parameters including the F0 parameter, to create a realistic source signal and to overcome buzziness present in the pulse excitation model. The mixed excitation source model and STRAIGHT based source model are the two primary approaches here.

2.5.1.1 Mixed Excitation

Yoshimura *et al.* first introduced mixed-excitation to SPSS [14]. The main idea of the mixed excitation source model is to have excitation signal similar to LP residual. Wherever the source

signal is represented with periodic impulses, voice quality sounds buzzy. Likewise, if the source signal is modeled by white Gaussian noise, voice quality sounds hissy. Hence, an accurate mixture of periodic and noise elements in the excitation signal will show an improvement in the perception of the synthesized speech.

During the analysis stage of mixed excitation source model, F0, voicing strengths in each band, and magnitude of Fourier spectrum are computed from speech data. Next, the voiced speech is separated into five sub-bands, with pass-bands of 0-1, 1-2, 2-4, 4-6, and 6-8 kHz [14], respectively. In every sub-band, strength is measured using normalized autocorrelation peak value around the pitch period. The voicing strength (VS) in each band around pitch (τ) is represented by

$$VS_\tau = \frac{\sum_{n=0}^{N-1} s_n s_{n+\tau}}{\sqrt{\sum_{n=0}^{N-1} s_n s_n \sum_{n=0}^{N-1} s_{n+\tau} s_{n+\tau}}} \tag{2.15}$$

where s_n and N represents the windowed speech signal and the window frame size, respectively. The first ten pitch harmonics are computed from the magnitude spectrum of the residual signal. The analysis vector computed for every frame consists of one F0 parameter, five band-pass voicing strengths, and ten Fourier magnitudes. These parameter are then trained using HMM.

During the synthesis stage, the voiced and unvoiced frames of the source signal are formed independently. The voiced frame is formed by creating a periodic pulse train with a pitch frequency of F0 and with the spectral properties of the initial ten harmonic component derived from the Fourier spectrum. The computed signal is transferred through band-pass filters. The band-pass filters gain is dependent on the band-pass voicing strengths. Both the filtered excitations are then combined to produce the mixed excitation signal. This signal is filtered by the MLSA synthesis filter, producing the synthesized speech. This technique has been shown to improve the naturalness of the synthesized speech. The improvement is essentially occurring owing to the mixed excitation, where noise is superimposed to various frequency bands of voiced excitation.

2.5.1.2 STRAIGHT based source model

As already discussed previously, STRAIGHT method gives smoothed spectral envelope and is widely utilized in SPSS. In addition, STRAIGHT method also gives other representation of speech like aperiodicty measurements. In the first version of the STRAIGHT, aperiodicity parameter is not

used [25]. Aperiodicity measurement is introduced in the later version of STRAIGHT to reduce the buzziness present in the initial version. Aperiodicity measure is determined as the ratio within the lower and upper smoothed spectral envelope. The aperiodicity component determines the amount of noise component in association to the periodic component present in the signal. Zen *et al.* integrated the STRAIGHT parameters into SPSS and synthesis quality improved drastically [15]. In order to model the aperiodicity spectrum in SPSS, its spectrum is averaged over the five frequency bands (0-1, 1-2, 2-4, 4-6, and 6-8 kHz). Hence, due to aperiodicity, the number of parameters is increased to 5 per frame for training in HMM.

The synthesis procedure of STRAIGHT model involves generations of voiced excitation using the F0 and aperiodicity measurements, which accounts for harmonic information. The impulse train is weighted by the aperiodicity spectrum and added with a random phase to scale down the buzziness present in the synthetic speech. The phase is obtained using an all-pass filter designed with a fixed group delay. The generated mixed excitation of each frame is convolved with the minimum phase MLSA filter response and synthetic speech $y(t)$ equation is given by

$$y(t) = e(t) \odot v(t) \tag{2.16}$$

$$v(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} V(\omega, t)\Phi(\omega)e^{j\omega t}d\omega \tag{2.17}$$

where $e(t)$ represents the excitation. Here, $\Phi(\omega)$ is an all-pass filter function used to add the random phase. The vocal tract filter response $V(\omega, t)$ is given by

$$V(\omega, t) = exp\left(\frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} h_t(q)e^{j\omega q}dq\right) \tag{2.18}$$

$$h_t(q) = \begin{cases} 2c_t(q) & q > 0 \\ c_t(q) & q = 0 \\ 0 & q < 0 \end{cases} \tag{2.19}$$

where $c_t(q)$ and q represents cepstral coefficients and quefrency, respectively.

The perceptual quality of STRAIGHT speech is better than the mixed excitation source model both in terms of naturalness and intelligibility. The gain in quality comes from two factors *i.e.*,

parametrization of the aperiodic component present in the voiced signal with aperiodicity parameters and introducing phase information with random phase using an all-pass filter, which is absent in both the impulse and the mixed excitation. Hence, STRAIGHT excitation is one of the most widely used source modelings in SPSS.

2.5.2 Residual modeling

Mixed-band excitation does not use the speech production knowledge to generate an excitation signal. Hence, the computed excited signal from mixed excitation is far away from the glottal flow signal. In the literature, some works are reported to mimic the actual glottal flow signal using error signal obtained from the LP analysis. The residual signal comprises of phase knowledge and nonlinear effects, which are not described by the mere strength and F0 parameters. To model the residual signal in literature, many attempts are done: out of which the closed-loop training and the pitch-synchronous residual codebook are two important methods.

2.5.2.1 Closed loop training

This approach was introduced by Maia *et al.* in 2007 for HTS [23, 24]. The concept is finding the optimal voiced and unvoiced filter parameters for generating the excitation signal of each HMM state.

In analysis stage, the training of voiced filter $H_v[z]$ and unvoiced filter $H_u[z]$ is performed for each state of HMM. During training, voiced source signal is deducted from the target source signal to get the unvoiced source signal. The resultant signal is filtered with the inverse unvoiced filter $1/H_u[z]$ to get the white noise error signal. The filters $H_v[z]$ and $H_u[z]$ are expressed as Mth order FIR and Lth order IIR filters, respectively, as

$$H_v[z] = \sum_{l=-M/2}^{M/2} h[l] z^{-l} \tag{2.20}$$

$$H_u[z] = \frac{1}{G[z]} = \frac{G_0}{1 - \sum_{l=1}^{L} g[l] z^{-l}} \tag{2.21}$$

where G is the gain parameter for the unvoiced filter. The above procedure is performed in an iterative fashion by adjusting the filter coefficients and optimizing position and amplitude of pulse train estimation and vice versa till the algorithm converges or the number of iterations is reached to a predefined value. The strengths and locations of impulse train are corrected from estimated filter coefficients.

During synthesis stage, filters $H_v[z]$ and $H_u[z]$ are defined according to filter coefficients generated from each HMM state. For voiced frames, impulse train is passed through the voiced filter $H_v[z]$, and for unvoiced frames, white Gaussian noise is passed through unvoiced filter $H_u[z]$. The combined mixed excitation is filtered by the MLSA filter to get the synthetic speech. The basic difference between closed-loop training method compared to the mixed excitation method is that here residual signal itself is learned through state dependent voiced and unvoiced filter parameters. The perceptual quality of speech is better than pulse excitation model and almost equal to the STRAIGHT based source model. However, this method involves an iterative procedure for learning the filter parameters, and accuracy of the excitation signal generation mainly depends on this iterative procedure.

2.5.2.2 Pitch synchronous codebook

In this method, the residual signal itself is preserved as codebook and it was proposed by Drugman [83, 84]. The pitch-synchronous analysis is applied on residual to get the codebook of excitation frames. The development of the residual codebook is done as follows. Initially, residual signal and GCI are estimated from the training database by inverse filtering of the speech signal. The residual signal is divided into frames of two pitch period with GCI as the center. Frames are condensed by the means of re-sampling and normalizing (RN) of residual frames to 20 samples. K-means clustering is applied on the RN-modified residual frames to get the RN codebook. The RN-modified frames are profoundly correlated and not usable in statistical modeling. Hence, PCA is applied to de-correlate the residual frames. The HMMs for each phoneme are trained with these PCA-transformed variants of the residual frames.

The synthesis stage of pitch synchronous codebook is shown in Figure 2.1. The PCA-transformed RN residual frames are transformed back to their initial RN form by matching to the frames in the RN codebook. The residual frame with the least euclidean length is chosen as the synthesis frame. After selecting the frame, its fundamental frequency and strength are transformed according to the target frame. Finally, synthetic signal is computed by employing the pitch synchronous overlap and add to each residual frame, and convolving resulting signal with the MLSA filter. The synthesis quality of pitch synchronous codebook for the male speaker is better than the mixed excitation model, but for female speaker, synthesis quality is not up-to the level of mixed excitation model. In addition, the residual codebook of large size has to be preserved in this method. Hence, Drugman *et al.* [84] applied PCA to the residual frames. This results in eigen-residuals and it was reported to give better excitation

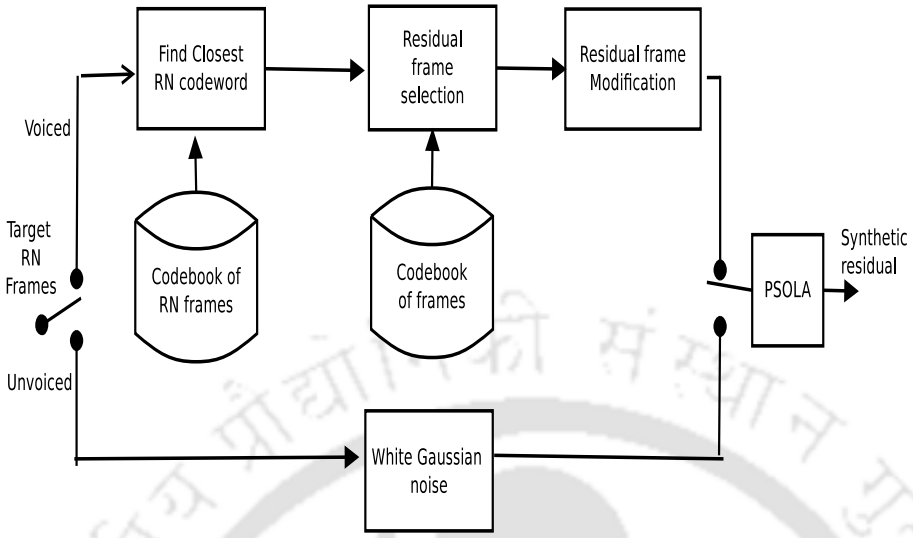

Figure 2.1: Synthesis block diagram of the pitch synchronous codebook method for source modeling

signal. However, the main disadvantage of this method is that separate memory is required to store the codebooks of residual signals. Further, the complex residual selection mechanism is required to pick the matching residual signal from the codebook.

2.5.3 Glottal source modeling

This method is driven by the fact that glottal pulses are used in the computation of the source signal [85, 86]. The Liljencrants-Fant (LF) model [87] and glottal (Glott) model [27, 32, 88] were two important methods in this category.

2.5.3.1 LF model

The LF model was integrated into the HTS STRAIGHT system by Cabral [87] in 2008 to model the voiced excitation rather than simple impulse excitation. This model represents the glottal flow derivative waveform approximately by employing seven parameters that are determined from the LP residual signal.

This model is separated into three components, represented in analytical form as:

$$e(t) = \begin{cases} A_0 e^{(\alpha_1 t)} \sin(\omega_g t), & 0 \le t \le t_e \\ \frac{A_e}{\varepsilon t_a}[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}], & t_e < t \le t_c \\ 0, & t_c < t \le T_0 \end{cases} \tag{2.22}$$

where $\omega_g = \frac{\pi}{t_p}$, t_0 is the vocal fold opening instant, t_p is the highest airflow instant, t_e is the maximum negative amplitude (A_e) instant, t_a is the duration from t_e to the instant where a tangent to the exponential at $t = t_e$ connects the time axis, t_c is the end of the exponential part instant, and T_0 is the fundamental period. The parameters A_0, α_1, and ε were computed from LF equation. The vectors for HMM training consists of the normal STRAIGHT parameters such as F0 and aperiodicity measures, and newly added logarithm of the inverted LF-model parameters.

During the synthesis stage, for every frame, voiced excitation is created by LF-model waveform of two pitch periods with the parameters gathered from HMM. Then, LF waveform is weighted with aperiodicity measures. In addition, to generate aperiodic component present in voiced frames, white Gaussian noise weighted by the aperiodicity measure is used. For unvoiced frames, only white Gaussian noise is employed as unvoiced excitation. Every excitation frame is filtered by the MLSA filter to generate the synthesized speech signal. The perceptual quality of this source model has proved to be better than impulse excitation source modeling, however, it still lags behind the STRAIGHT model. The main advantage of this model is its adaptability in regulating the different glottal source parameters coming from the LF model for use in voice transformation.

2.5.3.2 Glott model

Glott model for SPSS was proposed by Raitio *et al.* for HTS in 2011 using glottal pulses extracted by iterative adaptive inverse filtering (IAIF) algorithm [27,28]. The benefit of employing this approach for the synthesis is that approximated glottal pulses were utilized as the source signal and it gives added naturalness to the synthesis quality opposed to that of pulse train excitation.

The construction of the pulse library was implemented by using the IAIF algorithm. The glottal pulse and glottal flow derivative signal is obtained from IAIF algorithm. This glottal pulse signal is converted into frames of two-period with GCI as its center point. The collected frames are normalized and stored as pulse library with their source parameters as its cost function. During the synthesis stage, a glottal pulse is chosen from the pulse library with least target and concatenation costs by employing the Viterbi algorithm. The target cost is the root mean square error (RMSE) within the source features of the pulse and the features formed by the HMM. The concatenation cost is estimated as the RMSE within successive glottal pulses in every voiced frame. Once the glottal pulse is chosen, it is re-sampled and normalized according to F0 and log energy, also noise is added to get mixed excitation. This excitation is passed through an LP filter to generate synthesized speech.

The naturalness of the Glott model showed an improvement and better than the STRAIGHT version for a low-pitched male voice and as good as STRAIGHT for female speaker [27]. The improvement is mainly due to the use of natural excitation obtained from IAIF method. Yet, the speaker identity was slightly lower compared to the other method. Further, in this method separate algorithm is required to select the appropriate pulses for each frame from the library of pulses.

2.6 Other Signal Processing Models

The algorithms discussed so far are based on the source-filter analogy. However, there are some signal processing algorithms such as sinusoidal model and Harmonic-noise models where source-filter separation is explicitly not done. This section reviews these methods and its usage in the SPSS.

2.6.1 Sinusoidal modeling

Sinusoidal modeling was accommodated to HTS by Abdel *et al.* [89, 90] in 2006. Here, instead of separating the speech signal into source-filter modeling, spectral envelope information is captured by harmonics and these are modeled by the HTS along with the F0 parameter. The analysis stage of sinusoidal model consists of three stages: F0 estimation, vocal tract amplitude extraction, and sub-band voicing estimation. F0 calculation needs to be performed carefully since the vocal tract envelope calculation is reliant on it. The vocal tract envelope was estimated by determining the root mean square power of the harmonic component from the short-time Fourier transform (STFT) spectrum for every frame. In analytical form, the calculation of the harmonic components is formulated as:

$$A_j = \sqrt{\sum_{m=(j-0.5)k+0.5}^{(j-0.5)k+0.5} {s_m}^2} \tag{2.23}$$

where A_j is the amplitude of the j^{th} harmonic component, s_m is the m^{th} STFT sample with k is the number of STFT samples per frame, computed by $k = \frac{2\text{F0}N}{F_s}$, N is the number of samples and F_s is the sampling frequency.

In the synthesis stage of sinusoidal modeling, for every frame, it is represented as a sum of sinusoids:

$$s_j(t) = \sum_{k=1}^{N_j} A_{i,k}(t) \sin(\theta_{j,k}(t)) \tag{2.24}$$

where j is the frame number, N_j is the number of harmonic elements present in frame j, $A_{j,k}(t)$ is the strength of the k^{th} harmonic component, and $\theta_{j,k}(t)$ is the phase value of the k^{th} harmonic component.

The amplitude at the start of every frame are obtained from the vocal tract spectrum. The amplitude values were computed by a interpolation of the amplitude at the start and the end of the frame. The phase values $\theta_{j,k}(t)$ are estimated based on the F0 values and calculated as a interpolation of the F0 at the start and end of the frame. The initial phase is assumed to be zero, the phase at the start of the frame is represented as the phase at the end of the previous frame.

The unvoiced frame was formed by weighting white Gaussian noise with the vocal tract envelope interpolated from the vocal tract envelope amplitude in the frequency domain. Finally, the synthesis is done by mixing both voiced and the unvoiced signals according to the voicing strength. The quality of the synthesized speech from the HMM-adapted sinusoidal method shows that the application of the sinusoidal model apparently enhances the intelligibility of speech in contrast to the conventional MCEP computed from STFT spectrum.

2.6.2 Harmonic-noise models

The Harmonic plus Noise Model (HNM) was formerly introduced by Stylianou *et al.* [91, 92] for concatenative speech synthesis. This approach has been the source for several implementations used in SPSS [93–97]. The first adaptation of HNM into the framework of SPSS was introduced by Banos *et al.* in 2008 [94]. To make this approach suitable for SPSS, an approximation has been done in HNM. In HNM, the speech signal is represented as two elements: a harmonic element and a noise element. The harmonic element is determined from the summation of sinusoids. The noise element is computed by deducting the harmonic element from the initial signal. In the analysis stage of the HNM, the harmonic element is defined by considering the phase and amplitude of the harmonic elements in every frame. Subsequently, the harmonic and noise elements are parameterized for using it in the framework of HMMs.

The synthesis stage is relatively easier compared to analysis stage. The amplitude of the harmonic part of speech frame are computed by sampling the amplitude spectrum of the LPC envelope at integer multiples of the F0. The phase of speech frame is computed from the minimum phase response LP spectrum. An additional linear phase α is calculated in order to have consistency with those of the earlier frame. The amplitude and phase part forms the harmonic part of HNM model. The noise part of every frame is formed by filtering white Gaussian noise. The noise part of every frame is appended to the harmonic part to get the voiced excitation. For unvoiced frames, only the noise part is modeled.

It was reported that based on the synthesis quality, the HNM based method was an appealing

alternative to the well recognized STRAIGHT based analysis-synthesis framework with good intelligibility and naturalness. Moreover, it was employed in speech manipulation and voice transformation applications as well. However, incorporating phase knowledge in the HNM based method is not done extensively.

2.6.3 Evaluation Parameters

Speech synthesis is a perceptual phenomenon and to evaluate the synthesis quality similar to speech recognition, objective evaluation was used [98–101]. However, the perceptual phenomenon can be accurately judged by the subjective evaluations [102]. Different methods are there for the subjective evaluations. These methods basically try to capture the naturalness and intelligibility of the synthesized speech.

2.6.3.1 Naturalness

The naturalness of synthetic speech can be defined as voice quality closest to the human speech. Synthetic speech differs from a natural speech in characteristics of the glottal source shape, vocal tract features, and suprasegmental prosodic features [102]. Thus, each of these characteristics contributes in part to the perception of synthetic speech as unnatural. The glottal source characteristics refer to the individual shape of the glottal pulse for a particular speaker. The variation in the shape of the glottal pulse between natural and synthetic speech results in a change in the quality of speech. The acoustic features usually capture average source and system features for a fixed window size and results in a lack of dynamic variations in the synthesized speech. For example, synthesizers generate the speech from average acoustic features where the coarticulation effect will be also smoothed out and result in unnatural speech. The prosodic aspects are captured in duration, intonation, and amplitude variation of speech sounds. The difference in duration, intonation, and amplitude variations between natural and synthetic speech results in unnatural speech.

Naturalness can be assessed, for example, by playing synthetic speech samples to the listeners and asking them to rate the samples on a scale from 1 to 5. The mean opinion scores (MOS) are taken from subjects to get to know the quality of each system. However, minor differences in the synthesis quality between two systems are difficult to assess using the MOS. Also, it is difficult to determine which aspects listeners paid attention when assessing the samples. The MOS tests do not necessarily give absolute results that are comparable between evaluations performed at different times and in

different conditions. In order to compare two or more systems directly and to assess minor differences between systems, a preference test (PT) was used [103]. In such a test, subjects listen to two samples, one from each system, and choose the sample they prefer or rate the quality difference between the two samples. The subject is instructed to select the preferred sample, a preference score is obtained by calculating the percentage of how often a system was preferred over the other.

Apart from MOS and PT evaluations, MUSHRA (multiple stimuli with hidden reference and anchor) and ANOVA (analysis of variance) tests are also reported in the literature for evaluation of speech synthesis system [103, 104]. The main advantage of MUSHRA test over MOS test is that it requires fewer participants to obtain statistically significant results. This is because all wave files are given to subjects at the same time so that a paired t-test can be used for statistical analysis. Moreover, scores are given in the range of the 0-100 scale, hence, it is possible to rate very minute differences. The ANOVA evaluation is a statistical hypothesis test used in comparing the scores given for two systems from the different subjects. The statistically significant results for a particular null hypothesis will happen, when a probability (p-value) is less than a threshold [103].

2.6.3.2 Intelligibility

Intelligibility of speech refers to the understanding of the message information present in the speech signal. The message information of speech is presented in the vocal tract shape or formant locations. The variations in formant locations and changes in its dynamic range in comparison to original speech files result in degradation of intelligibility of speech. It can be measured either by evaluating overall speech message or by evaluating the recognition of single speech segments, phonemes or words in isolation or in a sentence. In a word recognition test, words are played either in isolation or in sentences to subjects who are requested to indicate what they have heard. The intelligibility is measured by the word error rate evaluated from the different subjects [100, 104]. The words in semantically sound sentences are rather easy to guess even if a word is not properly heard. Therefore, semantically unpredictable sentences are often used in order to prevent guessing [105]. These sentences are grammatically correct so that they form valid sentences, but the sentences may not give any meaningful message information. In order to know the errors present in the acoustic features, objective tests like Mel-cepstral distortion (MCD) and the likelihood of the training are used. These give an indication of how well the synthesis model represents when compared to natural speech [103]. Apart from this, perceptual evaluation of speech quality (PESQ) test is also conducted in literature [99, 106].

The PESQ measure can be interpreted as a MOS which gives the similarity of synthesized speech to the original waveform [101].

2.7 Summary and Discussion

In the previous sections, a detailed review of different acoustic features, analysis/synthesis framework, and modeling techniques for SPSS is discussed. In this section, advantages and disadvantages of all the approaches and future directions in SPSS is discussed. The advantage of SPSS lies in the compact representation using statistical models. It also provides flexibility in changing the voice characteristics, speaking style, and emotions. Further, the intelligibility of the SPSS is on par with the recorded speech. However, the naturalness of SPSS is degraded when compared with the USS method. Hence, in this section, different parameters and modeling techniques, which are contributing to the naturalness and intelligibility of SPSS are discussed.

Based on the studies done in the literature, to model suprasegmental features (F0 and voicing decision), the continuous model is better compared to MSD model [59]. The continuous model gives accurate voicing decision, particularly around voiced/unvoiced boundary regions. Further, it also gives the flexibility to model the voicing decision with various voicing parameters in separate HMM streams rather than modeling just binary decisions.

The vocal tract spectrum in SPSS usually will be modeled by the LSP/MCEP. The spectrum computed from LSP is smooth, however, it models only vowel sounds and it can not capture some of the sounds like nasals. The main drawback of the conventional MCEP computed from the STFT spectrum is that it consists of temporal and spectral fluctuations, which reduces the intelligibility of synthesized speech. Hence, to remove the temporal and spectral fluctuations caused by the fixed window analysis and harmonic effect, pitch adaptive windows are used in STRAIGHT method. The spectrum obtained from the STRAIGHT is smooth, hence, the synthesized speech is intelligible. However, one of the limitations of the STRAIGHT methods is its dependency on the accuracy of pitch estimation. Hence, some alternate vocal tract estimation method is required to get the smoothed spectrum without dependent on the pitch and also gives an intelligible speech.

In excitation model, a different mechanism is developed to mimic the glottal pulse source model and to get the natural speech in SPSS. Different factors influencing the naturalness of speech includes parameters representing glottal pulse shape, aperiodicity present in voiced sounds, phase information,

Table 2.1: Overall summary of different acoustic features for SPPS

Modeling technique	Advantages	Disadvantages
Different Voicing models		
Multi-space distribution model	• Single stream for modeling both F0 and voicing labels	• Errors in the voicing label
Continuous model	• Two separate streams for modeling F0 and voicing label • Accurate voicing decision	
Vocal tract spectral model		
LP spectrum	• Models vowel sound properly	• modeling anti-resonance sounds is difficult
STFT MCEP	• Generalized model for all sounds	• Source effect is not completely removed
STRAIGHT MCEP	• Smoothed envelope without any source effect • Gives intelligible speech	• Depends on pitch synchronous window • Complex procedure for parameter extraction
Source model		
Impulse/noise excitation	• Simple model	• Speech quality is not natural and buzziness present
Mixed band excitation	• Representation of Aperiodic component • Naturalness improved	• Procedure to extract features are complex
Residual model	• Learning residual from the models	• Separate memory • Complex algorithms involved in selection of codebook
Glottal source model	• Control over parameters of glottal pulse shape • Easy to change voice characteristics	• Complex algorithm for selecting glottal pulses
Other signal processing models		
Sinusoidal model	• No source-filter mechanism involved	• Number of parameters involved is almost double
Harmonic and Noise model	• Voiced speech comes from harmonic and noise component • Quality is comparable or better than STRAIGHT method	• Modeling phase is difficult
Acoustic modeling		
HMM	• Compact representation with intelligible speech • Adaptability and flexibility	• Speaker similarity is lost • Modeling decision tree is difficult in HMM
Hybrid Synthesis	• Natural speech compared to HMM • Smoothed join compared to USS	• Complex algorithm required for combining
DNN	• Natural and intelligible quality compared to HMM	• Database required for training is high
WaveNet	• Quality is almost near recorded speech	• Not much exploration is done

and prosody. The parameters used to model glottal pulse shape such as LF parameters in SPSS give quality better than impulse excitation. However, it is not similar to natural voice, primarily due to the parameters are estimated from approximated representation of glottal pulses. Further, usage of

actual glottal pulse or LP residual needs more memory and computation for selecting the codebook. Hence, there is a need for an alternate mechanism to get natural pulses from the modeling technique. Moreover, the phase components of the glottal pulse can be modeled by the complex cepstrum. The initial experiment was done by Maia *et al.* [26] shows promising results for modeling of the phase components, which resulted in a synthesized speech that is nearby to natural signal with improved perception quality.

The investigations revealed that the phase components of the glottal pulse are really necessary for high voice quality speech. The source signal has a particular phase design arising from the vocal-fold excitation signal and the asymmetry due to vocal-fold adduction and abduction process. The phase composition of the source signal is particularly important for earphone listening. Moreover, the high voice quality also depends on the aperiodic element present in the source signal. Hence, there is a need for modeling both phase and aperiodic components in source modeling to get high-quality speech.

2.7.1 WaveNet

Apart from HTS and DNN models, recently, WaveNet model is proposed by Google DeepMind team in 2016 [107]. WaveNet takes raw speech data and models using probabilistic and autoregressive approach by taking one sample at a time. It directly uses raw speech signals as input and learns the predictive speech samples conditioned to the previous one using the deep neural network and dilated causal convolution. The conventional generative models are based on a lot of assumptions like fixed-length analysis window by assuming the signal is a stationary process. The signal is processed with linear filter by ignoring the non-linear characteristics of successive samples. The signal is modeled with the assumption of Gaussian process. These assumptions are helpful in modeling signal in a convenient way. However, these assumptions in the modeling result in a synthetic signal and perceptually not natural compared to the original recorded speech. Hence, in WaveNet model, conventional block based processing, linear filtering, and assumption of Gaussian process in the model is avoided.

WaveNet model was used to overcome all the issues without using any prior assumptions about speech signal and try to model each sample at a time. The synthesis quality was reported as better than the USS, HTS, and DNN systems. Further, WaveNet can even capture the characteristics of non-speech sounds like breathing and mouth movements, which shows the greater flexibility of this model. The gain in the voice quality is because of modeling each sample and synthesizing speech sample by sample. However, since it is proposed recently, the robustness of this algorithm for speaker

adaptation, emotional analysis *etc.*, is yet to be tested. The overall summary of all the methods with its merits and demerits are summarized in Table 2.1.

2.8 Organization of the Present Work

The majority of the speech regions are constituted by the voiced sounds where the activity of glottis is present and these regions are perceptually very important. Hence, it is important to find out the glottal activity region. In Chapter 3, glottal activity region detection algorithm is proposed using the combination of source features. The features used are SoE, NAPS, and HOS, which represent the different aspects of the glottal activity. The proposed glottal activity regions detection is compared with others algorithms such as RAPT, STRAIGHT, summation of residual harmonics (SRH), and robust epoch and pitch estimator (REAPER).

The focus of this thesis is to get a speech from the statistical framework with high voice quality. Hence, in Chapter 4, the significance of different features present in the glottal activity region like epochs, strength of excitation, voicing decision, phase information is shown for speech synthesis. Further, voicing decision from the glottal activity features is enhanced using the classifiers. Finally, the HMM based speech synthesis systems are developed using proposed glottal activity features and then compared with other state-of-the-art algorithms to demonstrate the merits of using glottal activity features for synthesis task.

In Chapter 5, 2-D based framework is proposed for demodulation of speech spectrogram using Riesz transform. This approach attempts to obtain the coarticulation effect in a better way in contrasted to traditional 1-D analysis and produces smoothed vocal tract envelope. In addition, Riesz transform also provides voicing decision, F0, and coherence map. Lastly, usefulness all the features obtained from Riesz transform is shown in both analysis/synthesis framework and as well as in statistical framework.

The naturalness of speech is mainly due to the well-designed source signal. Hence, in Chapter 6, integrated linear prediction residual (ILPR) based source modeling is proposed for SPSS. Initially, a method is proposed for showing the importance of epochs and the instants around the epochs for synthesis. Further, representations for aperiodic component and phase information present in the glottal activity region is explored and its significance is showed for SPSS. The aperiodic component is obtained by separating the ILPR signal into harmonic and noise component. To capture the phase information present in the source signal an all-pass filter is designed with its output as cosine phase

of ILPR signal using the iterative procedure.

In Chapter 7, all the acoustic features derived in the glottal activity region are used together for SPSS framework to get a better voice quality. Hence, a combination framework using suprasegmental, system, and source feature for SPSS is shown for improvement in prosody, intelligibility, and naturalness of synthetic speech. Further, proposed framework is tested for two Indian languages and also in DNN framework.

Chapter 8, reviews the work presented in this thesis, highlights the main contributions of the work and gives some directions for future research.

3

Glottal Activity Region Detection

Contents

3.1 Motivation for detecting glottal activity region 46
3.2 Features for characterization of glottal activity 47
3.3 Robustness of features for characterizing glottal activity 51
3.4 Glottal activity region detection using different attributes 54
3.5 Summary 59

Objective Based on the review, it is evident that to get good quality speech, modeling of different components present in glottal activity region is necessary. In the speech production mechanism, glottal activity is present in the majority of speech sounds and these sounds are categorized as voiced sounds. The different components present in these regions are glottal closure instants, glottal opening instants, harmonicity, aperiodicity, phase information, and time-varying vocal tract response, which are perceptually relevant for speech synthesis. Hence, there needs a mechanism to detect glottal activity region. In a conventional way, these regions are identified using the strength of excitation (SoE). In this chapter, the normalized autocorrelation peak strength (NAPS) and higher-order statistics (HOS) is used as additional features for characterizing glottal activity region. The three features, namely, SoE, NAPS, and HOS, respectively, are the indicators of different attributes of glottal activity region, namely, energy, periodicity, and asymmetrical nature of the resulting source signal. The effectiveness of these features is analyzed using the different source representation like differential electroglottograph signal, zero-frequency filtered signal, and integrated linear prediction residual. The combination of glottal activity information from these three features outperforms any of them, demonstrating different information represented by each of these features.

3.1 Motivation for detecting glottal activity region

During speech production mechanism, the process of exciting the vocal tract system by the vibration of vocal folds during speech production is termed as a glottal activity. It can be characterized by the features present in the resulting excitation source signal. The earlier work demonstrated the significance of SoE in characterizing glottal activity region [108]. Even though SoE is effective in characterizing glottal activity regions, it represents only one attribute of glottal activity, namely energy, and may not always be high in the entire glottal activity region. This is due to the time-varying nature of the glottal vibration. Figure 3.1 shows a segment of electroglottography (EGG) and differentiated electroglottography (DEGG) signals. In the DEGG signal, the large negative peaks represent the SoE. Around 0.05 s, the transition from glottal to nonglottal activity region and hence a decrease in SoE starts. It can be observed from Figure 3.1(b), the DEGG is periodic (quasi-periodic), and has higher strength during glottal closing when compared with glottal opening resulting in asymmetrical nature. These attributes may also be exploited in addition to SoE for better characterization of the glottal activity regions.

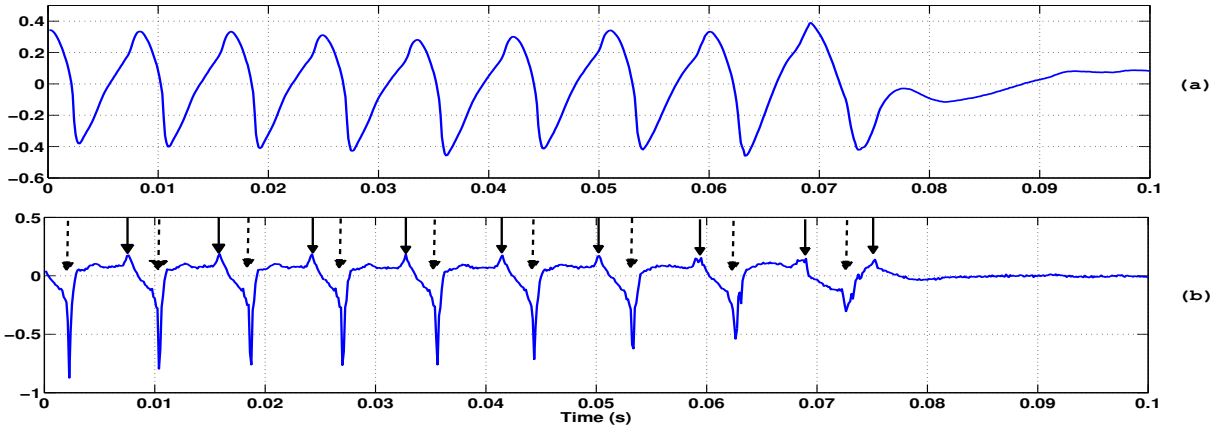

Figure 3.1: Nature of EGG (a) and DEGG (b) for glottal (0-0.05 s), glottal to nonglottal transition (0.05-0.08 s), and nonglottal (0.08-0.1 s) regions with glottal opening and closing marked in continuous and dashed arrow, respectively.

In this chapter, the effectiveness of periodicity, asymmetrical nature, and their combination with SoE are explored for characterizing the glottal activity region. The periodicity is computed using the normalized autocorrelation peak strength (NAPS) of source signal [64]. The asymmetrical nature is extracted using the higher-order statistics (HOS) of source signal [109]. The effectiveness of these features for characterizing glottal activity region is demonstrated initially using DEGG signal. However, in practice, the DEGG is seldom available with speech. Therefore, the two commonly used source signal representations derived from speech, namely, zero-frequency filtered signal (ZFFS) [110] and integrated linear prediction residual (ILPR) [111] are used for further exploration. Initially, all the three features, namely SoE, NAPS, and HOS, are evaluated independently and then combined. The rest of the chapter is organized as follows: The proposed features for characterizing the glottal activity regions using DEGG and speech are described in Section 3.2 and Section 3.3, respectively. The experimental evaluation of proposed features for the detection of GA regions is described in Section 3.4. The chapter is finally summarized in Section 3.5.

3.2 Features for characterization of glottal activity

In this work, initially, different features in the glottal activity region are computed from the DEGG signal and then the proposed features are applied on speech. The DEGG signal is a close approximation of excitation source signal, which captures glottal activity region in a noninvasive way.

This method was first developed by Fabre [112]. In DEGG, the glottal source signal is indirectly measured by capturing the changes in electrical impedance across the throat while speaking [112, 113]. It is characterized by energy, periodicity, and asymmetrical glottal pulse shape, and their variations [114, 115]. Therefore, DEGG is used for describing the SoE, NAPS, and HOS features, which are indicators of energy, periodicity, and asymmetrical nature of source signal, respectively.

3.2.1 Strength of Excitation (SoE)

In the glottal activity region, the significant excitation occurs during the closing of vocal folds and termed as instants of significant excitation or epochs [116]. The strength near an epoch can be obtained from DEGG by passing it through the zero-frequency filter and computing the slope of the filtered signal near the epoch [108]. The SoE ($s_e[k]$) is defined as the absolute slope of ZFFS ($z[n]$) given by

$$s_e[k] = |z[k+1] - z[k]| \tag{3.1}$$

where k is the epoch location. $s_e[k]$ gives the strength at the epoch location. Figure 3.2(a) shows the DEGG for a segment *Philip produced a coup* taken from BDL speaker of CMU ARCTIC database. The reference marks of glottal activity regions are shown by dotted lines. The SoE computed at epoch location is interpolated and Figure 3.2(b) shows the interpolated SoE evidence from DEGG. The strength is relatively high in glottal activity regions and low in non-glottal activity regions.

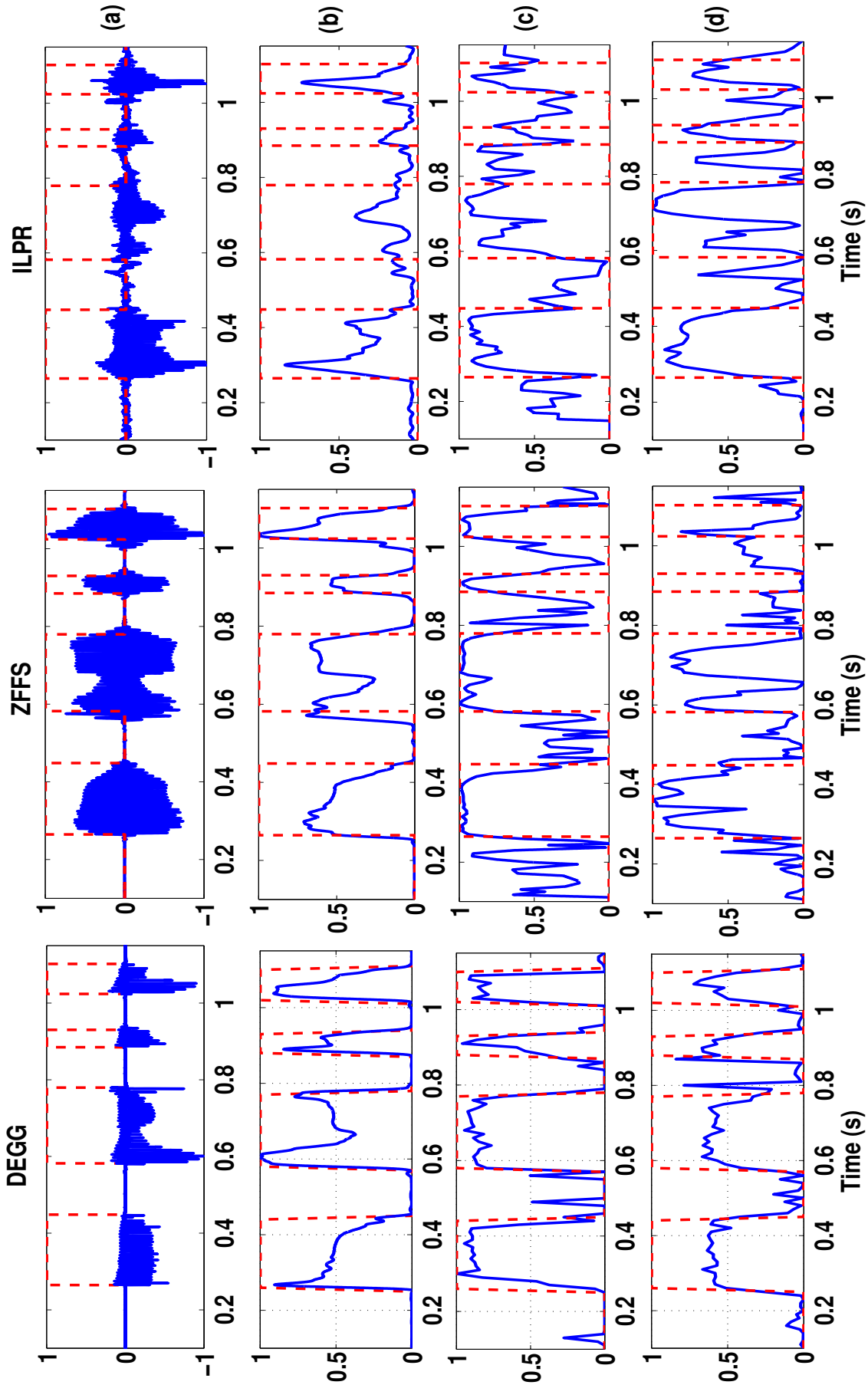

Figure 3.2: Characterization of glottal activity regions from source signal for a segment *Philip produced a coup:* (a) segment of source representations from DEGG, ZFFS, and ILPR, in three different columns of subplots; (b)-(d) three features SoE, NAPS, and HOS, respectively, obtained from each source representation (in three columns); reference marks are shown by dotted line in all the subplots.

3.2.2 Normalized autocorrelation peak strength (NAPS)

The DEGG is quasi-periodic in glottal activity regions. This nature can be extracted by computing NAPS of DEGG [64]. The NAPS ($n_p[k]$) for DEGG ($d[n]$) is given by

$$n_p[k] = \frac{\sum_{n=1}^{n=N} d[n]d[n-k]}{\sum_{n=1}^{n=N} d^2[n]} \tag{3.2}$$

where k is the delay and in the case of NAPS, it represents the location of the largest peak. The value of k varies from 2.5 to 15 ms and represents periodicity. It is computed for each frame of 20 ms with a shift of 10 ms and interpolated. Figure 3.2(c) shows the interpolated NAPS evidence obtained from DEGG. It is relatively high in glottal activity and low in non-glottal activity regions. This evidence seems to be equally effective like SoE in discriminating the glottal activity regions and the non-glottal activity regions.

3.2.3 Higher-order statistics (HOS) measure

During the glottal activity region, significant excitation occurs at the closing as well as the opening instants of vocal folds. However, the strength during closing instant is more when compared with that of opening instant. This causes the glottal pulses to be asymmetric in nature [114]. To capture this asymmetric nature, in [109], the appropriate power of the skewness-to-kurtosis ratio (SKR) is taken for making the evidence independent of signal energy and only a function of moments [109]. It is given by

$$\text{SKR} = \frac{\gamma^2}{\beta^{1.5}} \tag{3.3}$$

where γ and β are skewness and kurtosis of DEGG signal ($d[n]$), respectively, and are given by

$$\gamma = \frac{\frac{1}{N}\sum_{n=1}^{N}(d[n]-\bar{d})^3}{\left(\frac{1}{N}\sum_{n=1}^{N}(d[n]-\bar{d})^2\right)^{\frac{3}{2}}}, \beta = \frac{\frac{1}{N}\sum_{n=1}^{N}(d[n]-\bar{d})^4}{\left(\frac{1}{N}\sum_{n=1}^{N}(d[n]-\bar{d})^2\right)^2} - 3, \tag{3.4}$$

where $\bar{d}$ is the mean of $d[n]$. In this work, similarly asymmetrical nature of glottal pulse is computed from the source representation using SKR. This measure is computed for a frame of 20 ms with a shift of 10 ms and interpolated. The evidence is plotted in Figure 3.2(d) for the same DEGG segment shown in Figure 3.2(a). The SKR values are high in the glottal activity regions.

3.3 Robustness of features for characterizing glottal activity

The previous section described the features using DEGG. Since DEGG is the closest approximation to the excitation source signal, the derived features may represent the glottal activity region in the best possible manner. However, DEGG is seldom available along with speech. It is therefore required to analyze how well these features characterize glottal activity region using the source representations derived from speech. In this work, two recent methods of source representation, namely ZFFS and ILPR, are used for the characterization of glottal activity regions.

3.3.1 Zero-frequency filtered signal

ZFFS can be obtained by applying a zero-frequency filter on the speech signal. The zero-frequency filter is proposed by Murthy *et al.* [110] to compute the epoch location from the approximated source signal (ZFFS) from the speech signal. A brief procedure to find out the ZFFS from the speech signal ($s[n]$) is given below:

- Differentiate the input speech signal

$$x[n] = s[n] - s[n-1] \tag{3.5}$$

- Pass the differentiated speech signal $x[n]$ through the zero Hz resonator twice.

$$z_1[n] = \sum_{k=1}^{4} a_k z_1[n-k] + x[n] \tag{3.6}$$

  where a_1=4, a_2=-6, a_3=4, and a_4=-1. This is equivalent to four-time successive integration.

- The integrated signal ($z_1[n]$) is an exponentially growing signal. Hence, the trend in the $z_1[n]$ is removed by a moving average filter, which is given by:

$$z[n] = z_1[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} z_1[n+m] \tag{3.7}$$

  where $2N+1$ corresponds to an average pitch period.

- The trend removed signal $z[n]$ is termed as ZFFS.

The filtered signal shows high amplitude due to impulse-like excitations in the glottal activity regions when compared with the non-glottal activity regions. It is sinusoidal in nature due to periodic vibration

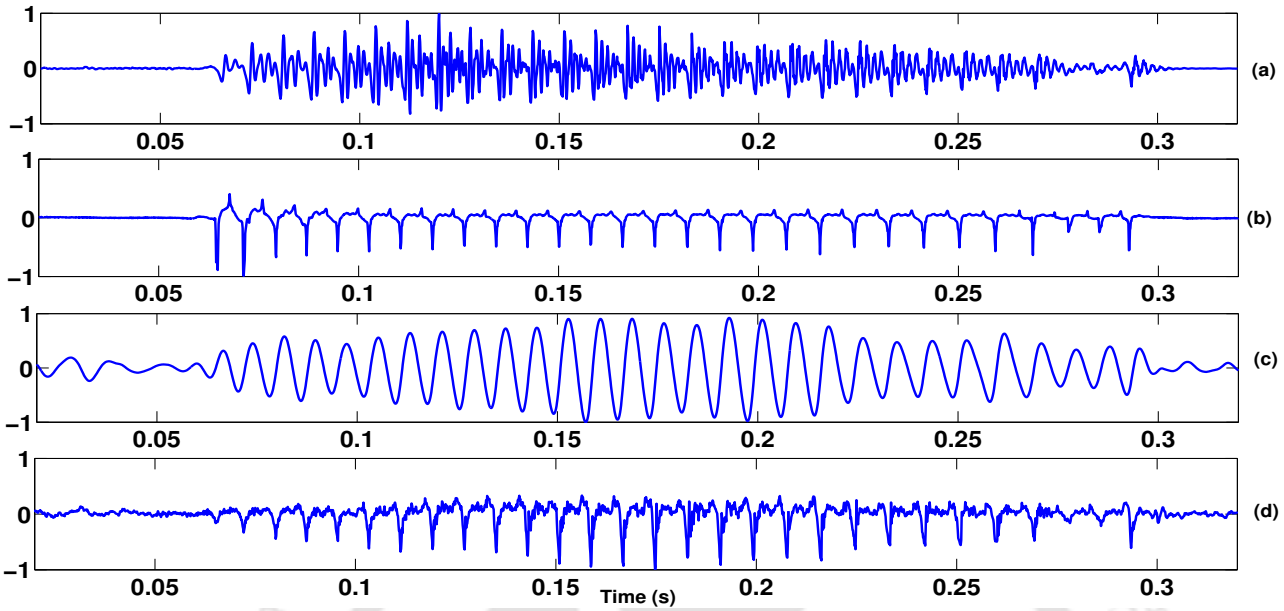

Figure 3.3: Source signal representation for a speech segment consisting of glottal activity regions: (a) and (b) speech segment and its DEGG; (c) and (d) ZFFS and ILPR source signal derived from speech.

of vocal folds. Also, the filtered signal is not purely sinusoidal in nature resulting in some asymmetry in each glottal cycle. Figure 3.3(a) and (b) show the speech segment and reference DEGG signal for one segment of glottal activity region, respectively. Figure 3.3(c) shows ZFFS having higher energy, periodicity, and less asymmetrical nature in glottal activity regions.

The glottal activity region evidence derived from ZFFS are shown with a solid line (in red color) in the second column of Figures 3.2(b)-(d). The three features are relatively high in glottal activity regions. In the case of non-glottal activity regions, these evidences are low. However, in some glottal activity regions, a slight degradation in these features is observed. This may be because the source signal used is the approximate representation. This can be observed in the second column of Figure 3.2(b) and (d), where, SoE and HOS values are low around 0.7 s region. The NAPS values are high in the same region in Figure 3.2(c). Also, in Figure 3.2(c) around 0.2 s region, NAPS values are high in the non-glottal activity regions. However, SoE and HOS have low values in the same region in Figure 3.2(b) and (d), respectively. Hence, the combination of all three features can be used to characterize glottal activity regions.

To study the effectiveness of the proposed features for speech, scatter plots of three features computed between ZFFS and DEGG are shown in Figures 3.4(a)-(c). The scatter plots of SoE and NAPS have most of the values along the diagonal indicating that the strength and periodicity obtained

from ZFFS are closely matching with that of DEGG. However, the scatter plot of HOS has a lot of values in the off-diagonal places indicating the less-asymmetrical nature of ZFFS. To quantify the linear orientation of these features, Pearson's correlation coefficient between DEGG and ZFFS is calculated [117]. It is evaluated for all the speakers of CMU ARCTIC and PTDB-TUG databases [22, 118]. The results given in Table 3.1 further confirm that the SoE and NAPS features are diagonally orientated when compared with the HOS feature.

3.3.2 Integrated linear prediction residual signal

The voiced speech signal is a result of exciting vocal tract system with a quasi-periodic train of impulses that forms the source signal. However, the source can be separated from the speech signal by inverse filtering operation using LP filter. Nevertheless, in conventional way, speech is pre-emphasized before applying LP filter to estimate the filter coefficients that increases the gain at higher frequencies relative to lower frequency regions in the computed LP spectrum. The pre-emphasized speech used during inverse filtering stage results in LP residual signal. This signal represents the approximated source signal, however, it contains high frequency component due to pre-emphasis operation. Alternatively, when the non-pre-emphasized speech is used during inverse filtering operation, the source signal obtained is called as ILPR signal, which is a close approximation of glottal flow derivative [111]. The ILPR signal $r_i[n]$ is given by

$$r_i[n] = s[n] + \sum_{k=1}^{p} a_k.s[n-k] \tag{3.8}$$

where $s[n]$ is non-pre-emphasized speech signal, p is the prediction order and it is set for $fs/1000 + 4$ (where fs is sampling frequency of speech signal) and a_k are the LP coefficients. The derived ILPR has more pronounced peaks during glottal closure and relatively smaller peaks during the glottal opening, causing the signal to have asymmetrical nature. The ILPR is similar to glottal flow derivative having periodicity property and high strength in glottal activity regions. Figure 3.3(c) shows ILPR and looks similar to DEGG signal.

The glottal activity region evidence derived from ILPR are plotted with a solid line (in red color) in the third column of Figures 3.2(b)-(d). This evidence has a large value in glottal activity region and low value in non-glottal activity regions. The HOS feature performs better in the case of ILPR when compared with ZFFS. This can be observed around 0.3 s region of ZFFS column and ILPR column in Figure 3.2(c). It may be due to higher asymmetrical nature of ILPR.

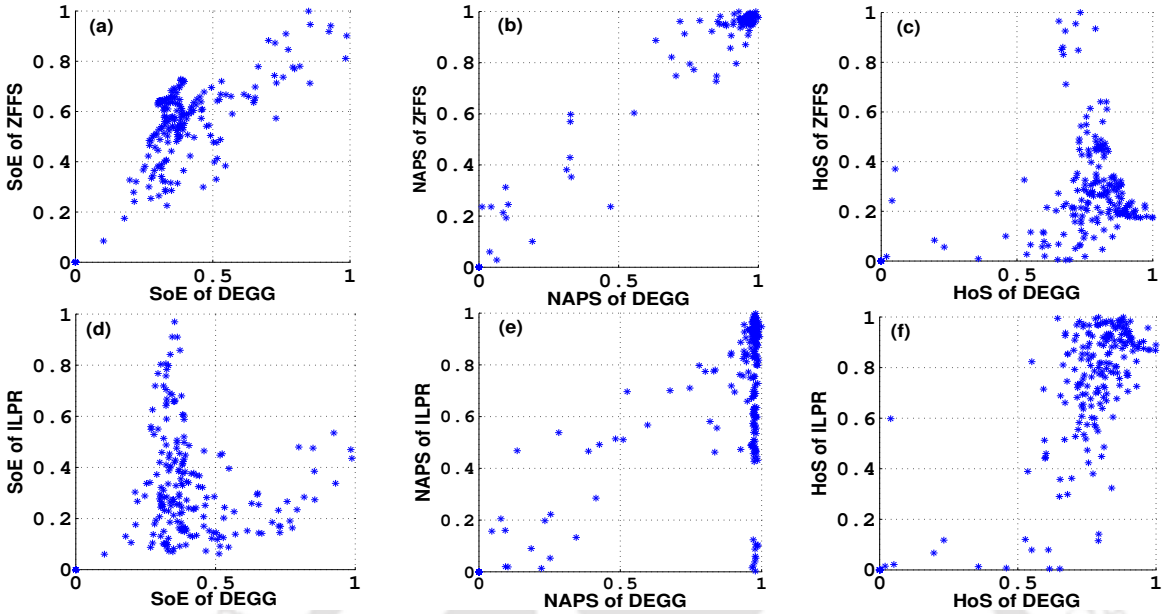

Figure 3.4: Scatter plot of DEGG vs. ZFFS, and ILPR source representation

Table 3.1: *Pearson's correlation coefficient between DEGG vs. ZFFS and ILPR*

	DEGG		
	SoE	**NAPS**	**HOS**
ZFFS	**0.9637**	**0.9838**	0.6010
ILPR	0.8382	0.8091	**0.8973**

To study the similarity of these features derived from ILPR, scatter plots between ILPR and DEGG are shown in Figures 3.4(d)-(f). The scatter plot of HOS has most of the values along the diagonal showing that the asymmetrical nature of ILPR matches closely with that of DEGG. However, scatter plots of SoE and NAPS have less values along the diagonal when compared with ZFFS. It is further signified by Pearson correlation coefficient of ILPR for HOS has higher values when compared with that of the other two features. This indicates that the combination of SoE and NAPS features derived from ZFFS, and HOS derived from ILPR may better characterize glottal activity regions.

3.4 Glottal activity region detection using different attributes

The features derived from the source signals are evaluated for glottal activity region detection task. In this work, two databases, namely, CMU ARCTIC and PTDB-TUG databases [22, 118] are used for evaluation. The CMU ARCTIC database consists of 5 speakers (4 Male and 1 Female) with

Table 3.2: Glottal activity region detection performance in clean conditions represented in terms of EER for both DEGG and speech signal

Different features		SoE	NAPS	HOS	Combined
DataBase	Source signal				
CMU	DEGG	4.08	2.74	2.05	**1.74**
ARCTIC	ZFFS	4.65	**4.12**	20.72	5.40
Male	ILPR	6.49	8.50	**4.48**	5.28
CMU	DEGG	3.21	2.11	1.90	**1.76**
ARCTIC	ZFFS	3.54	**2.32**	25.13	3.75
Female	ILPR	8.42	10.72	**6.79**	7.61
PTDB	DEGG	3.42	6.02	2.92	**2.72**
TUG	ZFFS	5.42	**3.48**	22.28	6.67
Male	ILPR	7.02	6.02	**4.25**	5.34
PTDB	DEGG	3.29	7.87	3.19	**2.47**
TUG	ZFFS	**4.67**	6.98	23.19	7.12
Female	ILPR	8.04	11.08	**5.27**	7.97
Average	DEGG	3.5	4.68	2.51	**2.17**
	ZFFS	4.57	**4.22**	22.82	5.73
	ILPR	7.49	9.08	**5.19**	6.55
	ZFFS+ILPR	4.57	4.22	5.19	**3.66**

simultaneously recorded speech and EGG signal. Similarly, PTDB-TUG database has reference EGG signal along with speech and consists of 20 speakers (10 Male and 10 Female). The glottal activity features are applied on DEGG, ZFFS, and ILPR. These features are processed using a frame of 20 ms with a shift of 10 ms and evaluated individually to get glottal activity region detection. As each of these evidence depends on the nature of the source signal, there are misclassifications and hence the three evidences are added and normalized using the min-max method. The min-max method will normalize the evidence into the range of [0.01 0.99] to obtain the combined evidence. The individual and the combined evidences are smoothed using local regression to avoid spurious evidences. The spurious evidences may be due to the fact that the features are derived from the source signal extracted from approximated representation of speech signal.

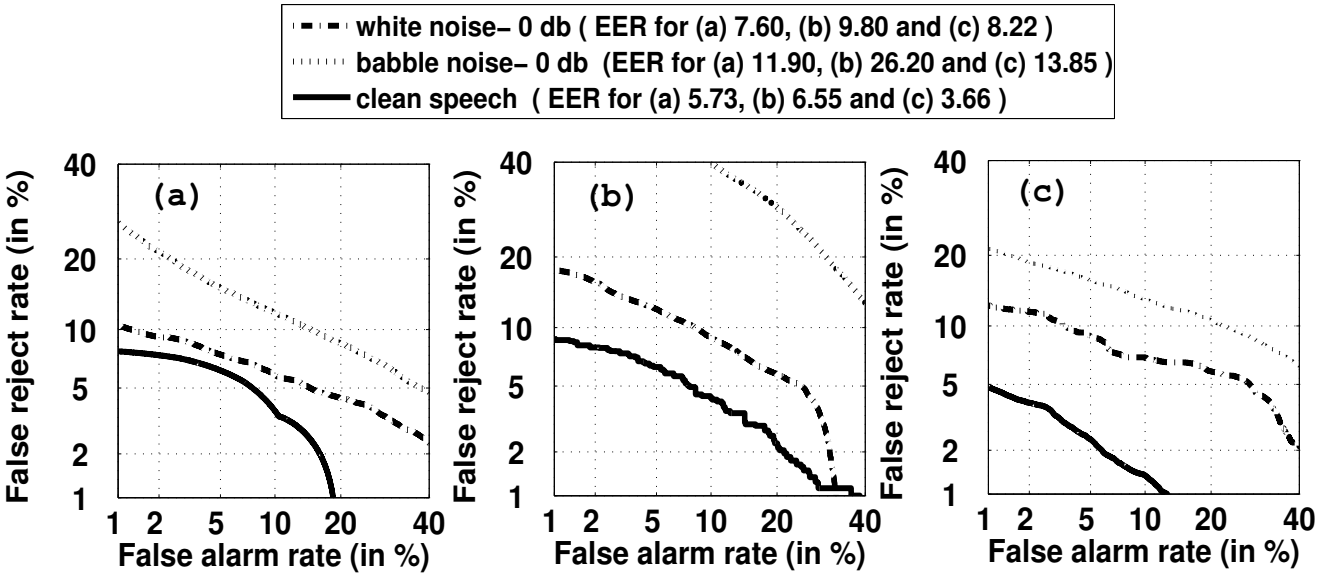

Figure 3.5: DET curve of glottal activity region detection task from different methods: The DET curve is shown for ZFFS, ILPR and combined source signal in each subplot of Figure (a), (b), and (c), respectively, for clean, white, and babble noise cases at 0 dB.

The proposed method is evaluated using detection error trade-off (DET) curve [108]. DET curve shows the trade-off between false alarm rate and false rejection rate. The equal error rate (EER) is a commonly accepted overall measure of system performance. The lower the EER value, the higher the accuracy of the glottal activity region detection method. The ground truth of glottal activity for EER calculation is derived from DEGG signal. The peaks in DEGG are calculated frame wise and reference marking of glottal activity region is detected by thresholding above the 1% mean value of peaks as mentioned in [119]. The positive class label for EER calculation above the threshold is taken as glottal activity region, whereas negative class label below the threshold is taken as non-glottal activity region. The overall performance of DEGG, ZFFS, and ILPR is tabulated in Table 3.2. In the case of DEGG, the combined method achieves an EER of 2.17, which is better than the individual evidences. The HOS evidence gives EER of 2.51 and it is lower for all the databases, which shows that DEGG is more asymmetrical in nature. The EER of speech has a value of 5.73 and 6.55 for ZFFS and ILPR, respectively. This value comes close to that of DEGG. There is a decrease in performance of ILPR. This is due to the higher EER of 7.49 and 9.08 for SoE and NAPS features, respectively. Similarly, HOS feature has an EER of 22.82 and results in the degradation of combined performance of ZFFS. Further, results indicate that asymmetrical nature of ILPR is better compared to ZFFS having a lower

Table 3.3: Comparison of the proposed method with other methods in clean conditions represented in terms of glottal activity region detection frame error

Database	Signal	wavesurfer	SRH	REAPER	Proposed
CMU-ARCTIC	DEGG	8.21	6.86	4.93	1.74
Male	speech	8.05	6.36	5.14	3.47
CMU-ARCTIC	DEGG	4.69	3.01	2.62	1.76
Female	speech	6.49	4.17	3.53	2.72
PTDB-TUG	DEGG	11.80	12.18	13.62	2.72
Male	speech	12.08	12.09	6.14	4.21
PTDB-TUG	DEGG	8.11	6.48	5.05	2.47
Female	speech	6.92	6.52	5.08	4.27
Average	DEGG	8.20	7.13	6.55	2.17
	speech	8.38	7.28	4.97	3.66

EER value of 5.19. Therefore, the best features from ZFFS and ILPR are considered and combined for the evaluation of glottal activity region detection. The SoE and NAPS features are taken from ZFFS. The HOS feature from ILPR gives an improved glottal activity region detection performance with EER of 3.66.

To study the effect of noise on glottal activity region detection, both the source representations are evaluated on synthetically generated noisy speech data at signal-to-noise ratio (SNR) of 0 dB. The noise waveforms are taken from Noisex-92 database [120] and tested for white and babble noise. Figure 3.5 shows the DET curves of the proposed features for ZFFS, ILPR, and combined source signal under clean and noise conditions at 0 dB. It shows that white noisy conditions performance of ZFFS is better than ILPR with EER of 7.60 and 9.80, respectively. For babble noise, the performance of ILPR is further reduced to EER of 26.2. The ZFFS gives EER of 11.90 for the same noise. It shows that glottal activity region detection of ZFFS is more robust to noisy conditions when compared with ILPR. The performance of combined source signal features is slightly less than ZFFS with EER of 8.22 and 13.85 for white and babble noise conditions, respectively. The slight decrease in performance of combined source signal may be due to HOS feature taken from ILPR, which is less robust to noise when compared with that of ZFFS. However, in this thesis work, the main focus is on the speech synthesis where clean speech is used for extracting the different acoustic features. Hence, this slight decrease in

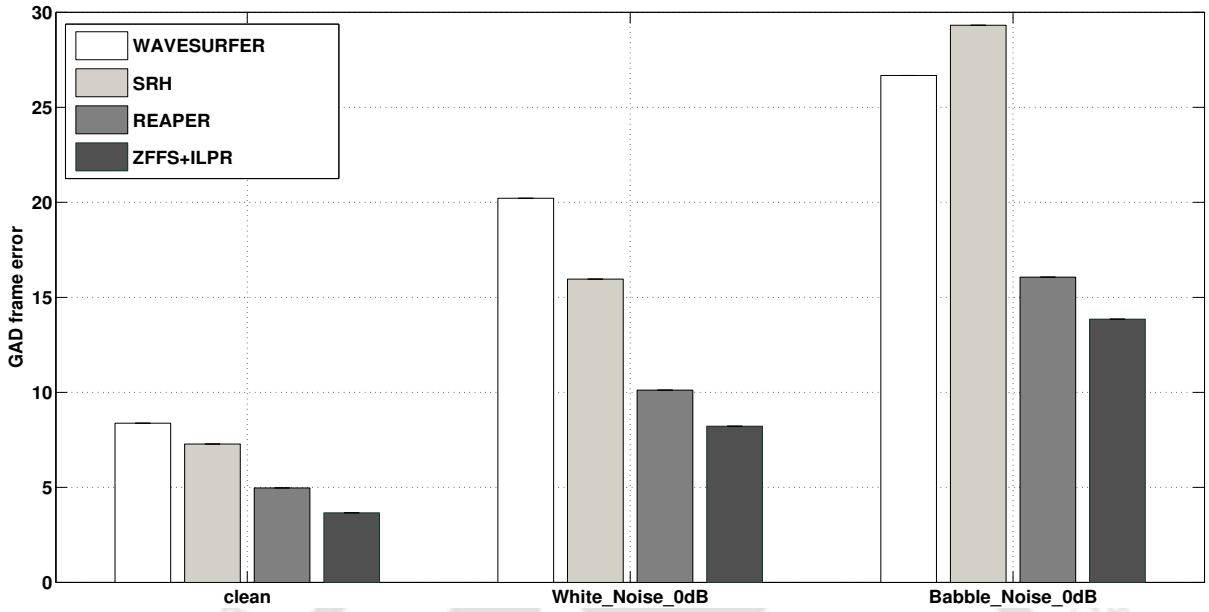

Figure 3.6: Glottal activity region detection frame error (%) for all the methods in clean speech and noisy speech at 0 dB of SNR with two types of noise.

the performance of combination method will not effect for the speech synthesis. The characteristics of babble noise are correlated to speech and hence the performance is poor when compared with white noise case.

The proposed combination method is compared with the three state-of-the-art algorithms, namely, wavesurfer, summation of residual harmonics (SRH), and robust epoch and pitch estimator (REAPER) available publicly [66, 67, 121]. Since wavesurfer , SRH , and REAPER algorithms are pitch-tracker algorithms and give binary voicing decision directly in terms of 0 or 1, the evaluation is done based on glottal activity region detection frame error instead of EER. The glottal activity region detection frame error is percentage error of glottal activity decision made with respect to ground truth. The ground truth for glottal activity region detection frame error is the same reference marking used in EER calculation. The comparison results are shown in Table 4.2. The proposed method is slightly better than the rest of the algorithms in clean speech case. The reason may be due to the fact that all the comparison algorithms depend only on the pitch of the source signal for glottal activity. However, the proposed algorithm depends on the epoch strength, periodicity, and asymmetrical nature of the source signal. Moreover, the proposed method performed significantly better in noisy conditions. This can be observed in glottal activity region detection error frame plot shown in Figure 3.6 for all the methods in clean and noisy condition at 0 dB. The REAPER and proposed methods are more robust

to noise when compared with other methods.

3.5 Summary

In this chapter, the significance of NAPS and HOS features, and their combination with the SoE feature to characterize glottal activity region are explored. The features capture different attributes of glottal activity region, namely periodicity, asymmetrical nature, and energy. Initially, the features are described using DEGG and their robustness was studied using ZFFS and ILPR obtained from speech. All these features are evaluated individually with CMU ARCTIC and PTDB-TUG databases for both clean and degraded conditions. In the case of ZFFS, SoE and NAPS features performed better. The HOS feature performed better in the case of ILPR. The performance is further improved, when the combination of best features is taken from ZFFS and ILPR. The proposed method performed better than state-of-the-art algorithms such as wavesurfer, SRH, and REAPER. In the next chapter, the significance of glottal activity region detection for speech synthesis task is discussed. Further, the role of different components present in the glottal activity region for improving the speech quality is showed in an SPSS framework.

4

Glottal Activity Features for Speech Synthesis

Contents

4.1 Motivation for processing glottal activity region for synthesis 62
4.2 Significance of Glottal activity features for Speech Synthesis 64
4.3 Glottal activity features for voicing decision 68
4.4 Improvement in the detection of glottal activity region using classifiers . 72
4.5 Glottal activity features for Synthesis 77
4.6 Experimental evaluation 79
4.7 Summary 84

Objective

In this chapter, the significance of features present in the glottal activity region is shown for improving the quality of speech synthesis. A method is proposed to improve voicing decision using glottal activity features for statistical parametric speech synthesis (SPSS). In existing methods, voicing decision relies mostly on fundamental frequency F0 and results in errors when the prediction is inaccurate. Even though F0 is a glottal activity feature, there are other features that characterize this activity, which may help in improving the voicing decision. The glottal activity features used here are the strength of excitation (SoE), normalized autocorrelation peak strength (NAPS), and higher-order statistics (HOS). These features are obtained from approximated source signals like zero-frequency filtered signal and integrated linear prediction residual. Further, to improve voicing decision and to avoid heuristic threshold for classification, glottal activity features are trained using different statistical learning methods such as the k-nearest neighbor (k-NN), support vector machine (SVM), and deep belief network (DBN). The voicing decision is best with SVM classifier and its effectiveness is studied using it in SPSS framework. The glottal activity features SoE, NAPS, and HOS are trained using hidden Markov model (HMM) along with F0 and Mel-cepstral coefficients (MCEP) to get the better voicing decision. The objective and subjective evaluations demonstrate that the proposed method improves the naturalness of synthetic speech.

4.1 Motivation for processing glottal activity region for synthesis

This chapter focuses on the better extraction of acoustic features present in the glottal activity region for analysis/synthesis framework. In speech production mechanism, speech is mainly categorized into voiced and unvoiced regions, based on whether the glottal activity or glottal vibration is present or not. The glottal activity regions can be characterized by variations in the locations of glottal closure instant or epoch due to quasi-periodic vibration of vocal folds, epoch strength due to variation in the strength of excitation signal, and an aperiodic component due to the turbulence noise generated in voiced speech. In existing literature [110, 116, 122], epochs correspond to glottal closure instants (GCI), glottal opening instants (GOI), and onset of bursts [116]. However, GCI are the major excitation present in speech production. Therefore, in this chapter, GCI are referred to as epochs. The glottal pulse is characterized by the duration of closing and opening phase of the glottal cycle, skewness or glottal pulse shape *etc.* [114, 115]. In this chapter, the significance of some these features

present in the glottal activity region is illustrated for speech synthesis. The glottal activity parameters are trained in HMM and used to improve the naturalness of SPSS.

In conventional SPSS using HMM, MCEP parameters are used to model the vocal tract transfer function. The fundamental frequency (F0) or pitch parameter is used to model excitation feature. F0 represents an only quasi-periodic characterization of glottal activity, whereas, other parameters present in the glottal activity regions are ignored. This is one of the reasons for the lack of naturalness in the synthesis quality of SPSS when compared to USS method [5]. In addition, F0 in HMM is trained along with voicing decision plays a critical role in the quality of synthetic speech. In HMM training, due to pitch estimation errors, the voiced frame may classify as a unvoiced frame. If it occurs within the middle of voiced region, degradation in synthesis quality will be higher. In addition, when the speech sounds are weakly periodic or strength of excitation is low [19, 20], there is a chance of missing the classification of some portions of voiced region. For instance, misclassification can happen around voiced-unvoiced (V-UV) transitions and UV-V transitions due to the low amplitude of speech signal [19, 20].

This chapter focuses on deriving robust voicing decision using different features representing glottal activity region and incorporating it in statistical modeling for improving the naturalness of synthetic speech. The focus is on voiced speech as it is perceptually very important and relatively easier to model. The present chapter is focused on showing the significance of different features present in the glottal activity regions for speech synthesis and training these features in HMM framework. Further, the glottal activity features are used as a voicing decision in speech synthesis to generate the excitation signal. To use features present in the glottal activity regions for voicing decision, some heuristic threshold is to be applied. This is avoided by using the classifiers. The advantage of using classifiers is transforming features into a higher dimensional space to get better separability for classification. The classifiers such as k-NN, SVM, and DBN are studied. The classifiers from simple to complex ones have been explored to examine their capability for the task of voicing classification and to find the classifier that performs best for using these features.

The main contributions presented in this chapter are:

- To show the significance of different features present in the glottal activity regions like epoch location, epoch strength, phase component, and voicing decision for speech synthesis.
- Improving the voicing classification from glottal activity features like SoE, NAPS, and HOS

using classifiers such as k-NN, DBN, and SVM.

- Training the glottal activity features SoE, NAPS, and HOS as a separate stream in HMM with the continuous distribution.
- Using the voicing decision for SPSS to improve the naturalness of synthetic speech.

The rest of the chapter is organized as follows: a brief description of different features present in the glottal activity region and their significance for speech synthesis is described in Section 4.2. The issues present in conventional voiced/unvoiced classification used in HTS and behavior of glottal activity features for a better voicing decision are described in Section 4.3. The refinement of voicing decision using different classifiers is described in Section 4.4. The integration of proposed glottal activity features for SPSS is explained in Section 4.5. The experimental evaluation and effectiveness of proposed voicing classification for synthesis is described in Section 4.6. The chapter is finally concluded in Section 4.7.

4.2 Significance of Glottal activity features for Speech Synthesis

In speech production mechanism, the majority of speech sounds are voiced sounds where the maximum glottal activity is present. The voiced sound unit refers to vowel, semivowel, nasal, voiced fricative, and voiced stop. Usually, the excitation signal for voiced sound is approximately modeled by impulse-like excitation. The remaining sounds such as unvoiced stop and unvoiced fricative are modeled by the white Gaussian noise. The quality of synthesis in the voiced regions in terms of excitation source signal mainly depends on different characteristics of glottal activity, like, strength of epochs, epoch interval, and so on. Further, to model the complex characteristics of voiced sound, skewness or shape of glottal pulse, duration of closing/opening phase of a glottal cycle, , changes in the vocal tract dynamics, and phase component present in the glottal cycle are to be considered. To model these characteristics of voiced sounds, proper representation of different aspects of glottal activity region is necessary. In this section, different parameters that represents the glottal activity and their importance to speech synthesis is discussed.

4.2.1 Cosine phase as excitation

In the generation of speech sounds, usually, magnitude only representation is used by ignoring the phase characteristics. However, for perceptual improvement of speech, phase component present in the

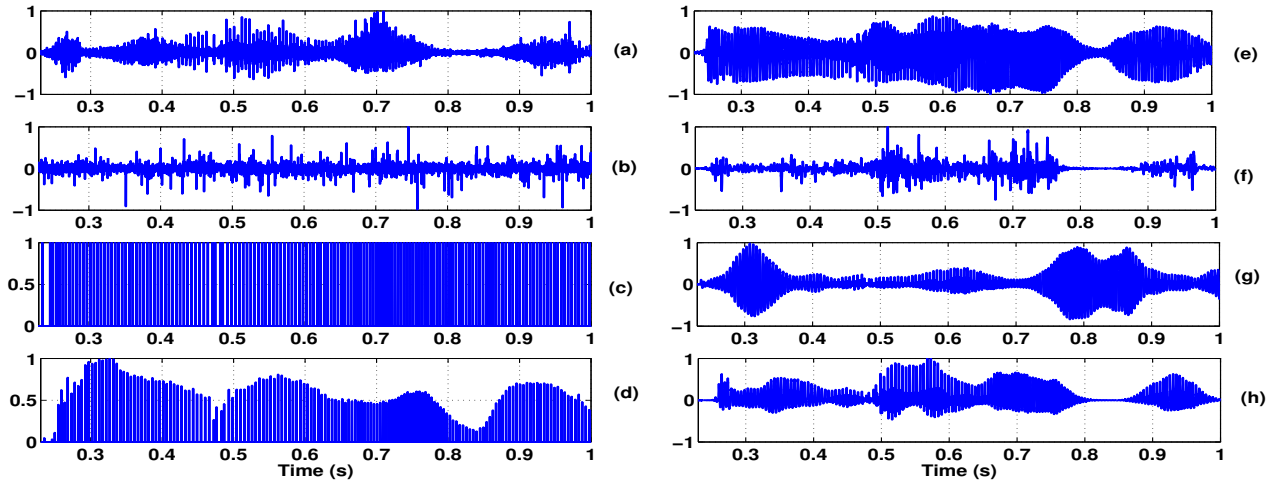

Figure 4.1: Excitation and synthesized speech for voiced segment: ((a)-(d)) A different types of excitation signal representing LP residual signal, cosine phase of LP residual, epoch based excitation signal, and strength weighted impulse signal, respectively; ((e)-(h)) corresponding synthesized speech from excitation signal obtained using LP residual, cosine phase of LP residual, impulse excitation signal, and strength weighted impulse signal

glottal activity region plays an important role [65,123]. In recent vocoders, the random phase is added to voiced excitation for synthesis using group delay to model the phase component [18]. However, the random phase may not be able to capture the actual phase component present in the glottal activity region and need some alternative representation. In this chapter, to know the significance of the phase component present in the glottal activity region, the cosine phase ($c[n]$) of linear prediction (LP) residual is used. Cosine phase is obtained from the residual signal by taking Hilbert transform of it to get the analytic signal ($r_a[n]$) or complex time function (CTF), which is given by,

$$r_a[n] = r[n] + jr_h[n] \tag{4.1}$$

where $r_h[n]$ is the Hilbert transform of $r[n]$. This transform does not change the magnitude of the signal. It just alters its phase, i.e shifts the phase of a positive frequency by -90 and that of negative frequency by +90. Thus, the signal and its Hilbert transform have the same amplitude, but have a phase difference of ± 90, and thus are orthogonal to each other. Further, to remove periodicity and strength of excitation (SoE) from the ILPR signal, cosine phase ($c[n]$) is computed as follows:

$$c[n] = r[n]/h[n] \tag{4.2}$$

where $h[n]$ be the Hilbert envelope (HE) and it is computed from CTF as follows:

$$h[n] = |r_a[n]| \tag{4.3}$$

$$h[n] = \sqrt{r^2[n] + r_h^2[n]} \tag{4.4}$$

Compared to random phase excitation generated from the white Gaussian noise, the cosine phase signal is useful as it preserves the phase component and some speaker-specific information as reported in [124]. To illustrate the significance of this phase representation of glottal activity, speech is synthesized from the cosine phase signal. For synthesis, the cosine phase signal is used as an excitation signal with synthesis filter, where the response of the filter is obtained from LP coefficients. Figure 4.1(b) shows the cosine phase excitation signal for the utterance taken from the CMU-ARCTIC database and synthesized speech is shown in Figure 4.1(f). Further, to understand the significance of cosine phase excitation for synthesis, speech quality is evaluated using the perceptual evaluation of speech quality (PESQ) measure. The average score for the cosine phase excitation signal is shown in Table 4.1. The above result is evaluated for 25 sentences of the ARCTIC database for one male speaker (BDL) and one female speaker (SLT). For comparison, the synthesis quality of the speech signal obtained from the cosine phase excitation is compared with random phase excitation. From the PESQ score shown in Table 4.1, it can be seen that the perceptual quality of the cosine phase excitation signal is better than the random phase excitation, which signifies the preservation of phase component in the cosine phase signal.

4.2.2 Epoch and its location

The periodic vibration of vocal folds can be approximately represented by the impulse-like excitation located around GCI or epoch [116]. The location of epochs varies with each glottal cycle resulting in aperiodicity. To represent this aperiodicity present in glottal activity region, impulses are introduced in the epochs and their location has persevered. In this chapter, the epoch locations are identified by filtering the speech signal using the zero-frequency filter (ZFF) and trend removing filter [110]. A brief procedure to find out the epoch location from the speech signal ($s[n]$) is given in Chapter 3.

To know the significance of epoch locations, impulses are inserted around the epoch locations, which is shown in Figure 4.1(c). In this chapter, this excitation is referred as epoch based excitation.

The speech synthesized from the LP filter by using the knowledge of epoch gives better PESQ score than cosine phase based excitation and also frame level fixed impulse based excitation. The results signify the importance of epoch locations, which is part of the glottal activity region. The main difference between the epoch based excitation and the impulse based excitation is in the former case F0 interval varies for each glottal cycle, whereas in impulse excitation F0 interval is fixed for each frame.

4.2.3 Epoch strength

The voiced sounds are produced due to the airflow pressure from the lungs and simultaneous adduction/abduction process of vocal folds. This results in high energy excitation during closing of glottis and gives higher strength to voiced speech at the epoch location. This strength around the epoch locations may be estimated by filtering speech using the zero-frequency filter and computing slope near the epoch locations of filtered signal [108]. The strength of excitation ($s_e[k]$) of the filtered signal ($z_2[n]$) is defined as

$$s_e[k] = |z_2[k+1] - z_2[k]|\,, \tag{4.5}$$

where k is the epoch location. $s_e[k]$ gives the strength of the impulse-like excitation at the epoch location. The epoch strength shows variations in strength for different sounds and it can also be seen in Figure 4.1(d). The variations in the amplitude of epoch are used as excitation and speech is synthesized using LP filter. The PESQ score also shows improvement in the naturalness when compared with a constant amplitude impulse used in the epoch location.

4.2.4 Voicing decision

To model different attributes of voiced sounds along with the characterization of glottal activity, accurate detection of the glottal activity region is required. The accurate detection of these regions will help in generating excitation i.e voiced or unvoiced excitation, which is commonly used in most of the vocoders. The misclassification in the glottal activity region leads to voiced region classified as a unvoiced region and gives noisy voiced speech. The unvoiced region classified as voiced region gives hoarseness to the synthesized speech. Hence, the accurate classification of glottal activity region is necessary for synthesis. Table 4.1 shows the PESQ score of 2.97 for the synthesized speech after using the impulse or white Gaussian noise excitation obtained with a voicing decision from glottal activity features. The procedure to compute the voicing decision from glottal activity features is explained

Table 4.1: PESQ for different types of excitation

Excitation signal	PESQ
LP residual	4.5
Cosine phase	1.61
Random phase	1.27
Impulse	1.81
Epoch based excitation	1.94
Epoch strength weighted excitation	2.13
Voicing decision	2.97
Voicing decision + epoch strength + cosine Phase	3.54

in Section 4.3. Scores show that excitation based on voicing decision gives better perceptual quality compared to other methods. Further, when the cosine phase is added to variable epoch strength along with voicing decision and used as excitation (voicing decision + epoch strength + cosine phase), there is an improvement in a PESQ score of 3.54, which is highest compared to all individual parameters. These results signify the importance of different features present in glottal activity region like epoch location, its strength, phase component, and overall detection of voicing decision for speech synthesis.

4.3 Glottal activity features for voicing decision

In the base version of HTS, the voicing decision is made independently for each state using the MSD model of F0. The pitch parameter is computed using a robust algorithm for pitch tracking (RAPT) [17]. The RAPT algorithm performs the frame by frame autocorrelation analysis to capture periodicity information. The errors in pitch detection algorithm may happen due to the low energy boundary regions of voiced sound. These errors may even procure to HMM training and during parameter generation, voiced frame may classify as unvoiced frame. In addition, if the misclassified unvoiced HMM state happens to occur in the middle of a voiced sound with voiced neighboring frames, quality will be poor.

Figure 4.2 is a synthesized sample of the English word “sleep” from SPSS, where F0 and voicing decision are obtained from RAPT algorithm [17]. In this word, due to the voicing decision error,

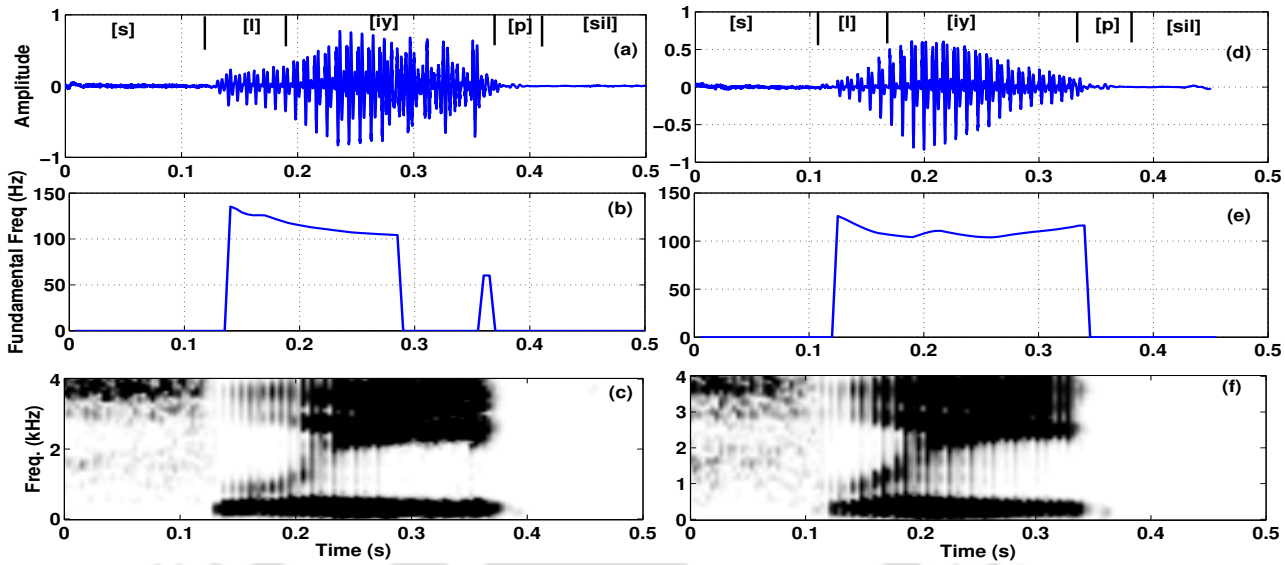

Figure 4.2: (a) SPSS synthesized speech for an English word "sleep"(/s/, /l/, /iy/, /p/,/sil/) using RAPT algorithm; (b) fundamental frequency with voicing decision; (c) spectrogram for the same word; ((d)-(f)) shows the SPSS synthesized speech for the same word using the proposed glottal activity features, fundamental frequency with voicing decision and spectrogram, respectively.

the unvoiced frames appeared within the voiced sound /iy/. This vowel sounds very dry and hoarse, which degrades the naturalness of synthetic speech. Thus, the features used for voicing decision are insufficient the capture the dynamics of voiced speech. There is a need for better features to represent the voicing information, along with conventional periodicity feature.

4.3.1 Analysis of F0 for different voiced sounds

In the basic version of HTS, voicing decision is made using glottal activity feature like F0. To know the significance of this feature, the relative frequency of F0 feature for all the sentences of SLT speaker taken from the ARCTIC database [22] is plotted in Figure 4.3. It shows the distribution of F0 values for different voiced sound categories like vowels, semivowels, nasals, voiced fricatives, and voiced stops. F0 is calculated from RAPT algorithm with a frame shift of 5 ms. Even though most of the F0 values for sound units such as vowels, semivowels, and nasals are around 110 Hz to 230 Hz, around 10 % values have F0 = 0 i.e these frames are classified as unvoiced sounds. For voiced stops and voiced fricatives, around 40 % of frames are classified as unvoiced sounds. There is a need to have better glottal activity features, which can classify these frames correctly.

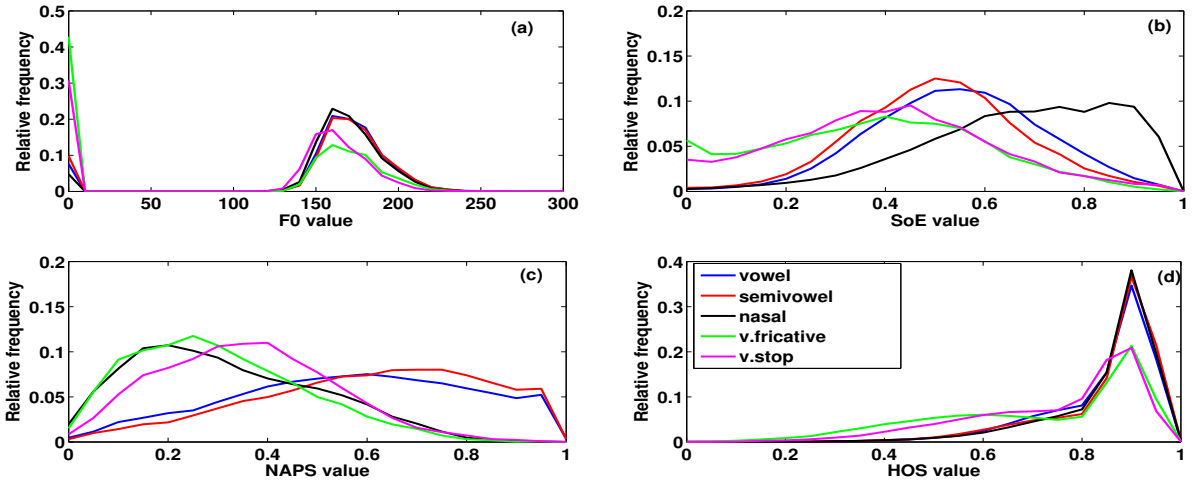

Figure 4.3: The distribution of different features present in glottal activity region for voicing decision: (a)-(d) relative frequency of feature values for F0, SoE, NAPS, and HOS, respectively

4.3.2 Analysis of SoE, NAPS, and HOS for voicing decision

In earlier chapter 3, three glottal activity features and their combination are proposed to identify the regions where the activity of glottis is present [125]. Three glottal activity features considered are SoE, NAPS, and HOS of the source signal. This section describes the voicing decision of phoneme units using these glottal activity features. To show different aspects of glottal activity present in voiced region, distribution of SoE, NAPS, and HOS values for different voiced sounds are plotted in Figure 4.3((b)-(d)). The SoE values for voiced fricatives and voiced stops are relatively low with around 5 % of frames present in the total database and it is better than F0 distribution shown in Figure 4.3(a). In addition, the relative frequency of the NAPS and the HOS features works better for voiced fricative and voiced stop sounds. Hence, the combination of features will be helpful for the correct classification of voiced sounds, particularly, NAPS and HOS features are helpful in low energy voiced sounds such as voiced consonants/semivowels.

Figure 4.4(a) shows a natural speech for a phrase “a big canvas” consisting of some weakly voiced sound units such as /b/, /g/, and /v/. The voicing decision obtained from the RAPT algorithm is shown in Figure 4.4(b) along with the normalized fundamental frequency plot showed in a continuous line. The RAPT algorithm works fairly well for a weakly voiced sound unit like /g/. For voiced consonant /b/ (around 0.4 s), pitch estimation fails due to the sudden dip and results in the sound unit being classified as the unvoiced region. A similar observation can be seen around 0.9 s region

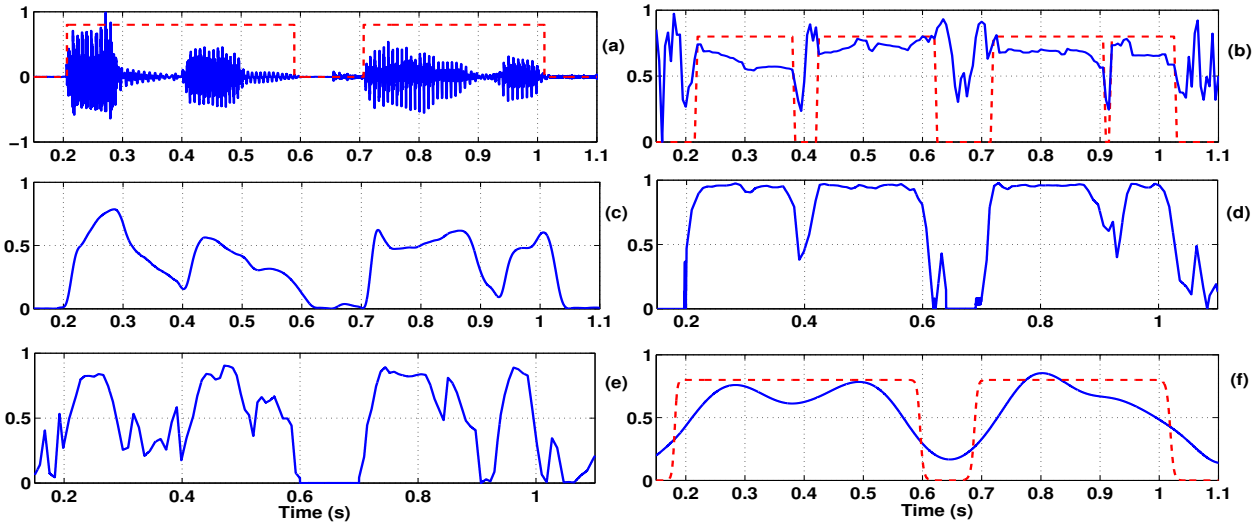

Figure 4.4: (a) Natural speech for a phrase "a big canvas"(/a/,/b/,/ih/,/g/,/k/,/ae/,/n/,/v/,/ah/,/s/) with reference voicing marking in dotted line; (b) voicing decision obtained from RAPT with amplitude 0.8, and normalized F0 evidence; (c)-(e) glottal activity evidence obtained from features, SoE, NAPS, and HOS, respectively; (f) proposed voicing decision (dotted line) using combined evidence of glottal activity features (continuous line)

for voiced sound unit /v/. Figure 4.4((c)-(e)) shows the SoE, NAPS, and HOS features, applied on the same segment, respectively. Even though the SoE evidence is low for /b/ and /v/ sounds, the NAPS (around 0.4 s and 0.9 s) and the HOS (around 0.4 s) values are high around these regions. Figure 4.4(f) shows the combined evidence and the final classification with the correct classification in the regions around 0.4 s and 0.9 s. The combined evidence is computed by normalizing all the three individual evidence to a maximum value and then averaging them to get the combined evidence in the range of 0.01 to 0.99. The reason for the combination of features is due to the fact that the NAPS evidence is computed on the ZFFS, an approximated band-pass filter and gives high gain for the components around fundamental frequency, resulting in enhancing the sinusoidal nature of low amplitude voiced sound units. Similarly, the HOS feature is computed from the ILPR signal captures the boundary/transition region very well since the residual errors around these regions are relatively high.

Figure 4.2(d) is a synthesized sample of the same English word "sleep" generated from SPSS using the proposed combination of glottal activity features. The voicing decision is shown in Figure 4.2(e) is computed from the proposed three glottal activity features trained in HMM. The detailed procedure of computing voicing decision is shown in Algorithm 1. In the synthesized word, there is no voicing decision error, which can be seen in Figure 4.2(e). This is mainly due to the better voicing decision

obtained from the combination of glottal activity features SoE, NAPS, and HOS. Hence, the additional features along with F0 are helpful in improving the voicing decision.

4.4 Improvement in the detection of glottal activity region using classifiers

An extensive study of the glottal activity features for classification of voiced sounds using different classifiers is explored in this section. There are two advantages of using classifier. First, classifiers will be helpful for enhancing the classification accuracy. Secondly, the heuristic threshold for classification can be avoided. The first classifier is the k-NN, a simple classifier stores all the features and their labels during the training process. During testing, the distance (Euclidean) of an unknown sample to all the training samples is computed. The labels of k nearest samples to the unknown sample are considered and the dominant label is assigned to the unknown sample [126].

Another classifier explored is the SVM, a binary classifier. This classifier finds optimal hyperplane by maximizing the distance of hyperplane from the support vectors. The details of SVM can be found in [127]. In this chapter, LIBSVM [128] toolkit is used for the SVM experiments. The radial basis function is used as kernel function here and it is given by:

$$K(x, y) = \exp(-Y||x - y||^2) \tag{4.6}$$

where x and y are training samples and labels, respectively, Y is the width parameter. There is also a cost parameter c for SVM and details can be found in [129]. The parameters Y and c are varied so as to achieve the best performance.

DBN have been used for voice activity detection task by Zhang *et al.* [76]. In DBN, serially concatenated multiple features is given as input and transferred through the multiple nonlinear hidden layers of DBN. The final layer of DBN consists of soft-max layer acts as a classifier, and used for classification of these glottal activity features. A stack of restricted Boltzmann machines (RBM) constitute a DBN and the training process takes place in two steps [42]. The first step consists of a pre-training stage, where each RBM is pre-trained in an unsupervised manner and the output of first RBM is given as input to the next RBM. This step is repeated until the final layer of DBN. The above process is performed to find the initial parameters of DBN, which are close to a good solution. The next step is the supervised back-propagation step and it is performed to fine tune the parameters.

For a 2-class classification problem like the voicing decision, a given observation o is assigned to a

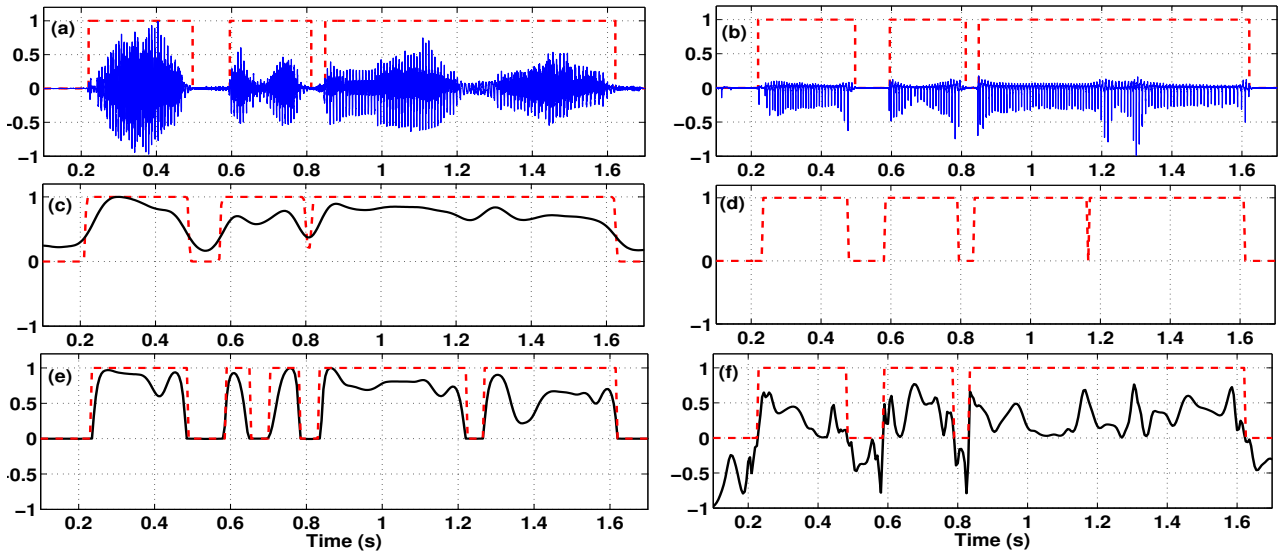

Figure 4.5: Voicing decision using classifiers: (a)-(b) speech and DEGG segment with reference marking in dotted line; (c) signal processing combined evidence in continuous line (black color) and final nonlinear mapping in dotted line (red color); (d)-(f) the voicing evidence obtained from k-NN, DBN, and SVM, respectively, with likelihood probability in continuous line (black color) and mapping in dotted line.

class y_k, whose analogous output unit is given a value of 1. The output unit $y_k, k = 1, 2$, is calculated similar to [76], by the following relations:

$$y_k = \begin{cases} 1, & \text{if } p_k > p_i, \forall_i = 1, 2, i \neq k \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

where p_k is the probabilistic soft output of the voicing class "$y_k = 1$", p_k is defined as $\exp(d_k)/\sum_i \exp(d_i)$ with d_k defined as

$$d_k = \sum_i w_{k,i}^{(L+1,L)} f_i^{(L)} \left(\sum_j w_{j,n}^{(L,L-1)} f_n^{(L-1)} \left(\ldots \sum_b w_{a,b}^{(2,1)} f_b^{(1)} \left(\sum_c w_{b,c}^{(1,0)} x_c \right) \right) \right) \tag{4.8}$$

with $f^{(l)}(.)$ representing the nonlinear activation function of the l^{th} hidden layer, $l = 1, ..., L$, $(w_{i,j}^{(l,l-1)})_{i,j}$ represents the weights between the neighboring layers with i as the i^{th} unit of l^{th} layer and j as the j^{th} unit of $(l-1)^{th}$ layer and $(x_c)_c$ represents the input feature vector [76].

The input to the classifiers consists of different glottal activity features and the labeled reference from differentiated electroglottography (DEGG) signal. The features are concatenated to form a vector for each frame. In addition, contextual information is incorporated wherein a five frame context was used. This means five frames to left and five frames to right of the current frame are used to get

the final input vector to the classifier. To compare glottal activity features without classifiers, these features are combined and nonlinearly mapped to get the binary classification. The nonlinear mapping function [130] is given by

$$E_p = \frac{1}{1 + \exp^{-(S_p - \Theta)/\tau}} \tag{4.9}$$

where E_p is nonlinearly mapped value and S_p is the combined glottal activity feature value obtained after adding SoE, NAPS, and HOS features and normalizing to the range of 0.01 to 0.99. The variables Θ and τ are the slope parameters, τ is chosen to be very small with a value of 0.001 and Θ is the main threshold chosen to be 0.7 in this chapter found empirically. The advantage of using nonlinear mapping compared to the direct threshold is to get the better voicing decision, which is also shown in the paper [131] for speech/music classification. The reason for getting better results may be coming from the fact that due to the sigmoid function used in the nonlinear mapping, the range of input feature obtained from the combined evidence of glottal activity features is increases. That is, higher feature values further increase and mapped to binary decision 1 and lower feature value further decrease and mapped to binary decision 0.

4.4.1 Voicing classification

The classifiers mentioned above are used for voicing decision using glottal activity features obtained from speech basically characterizes voiced sounds. The remaining sounds where the activity of glottis is minimal are classified as unvoiced sounds. The reference labels for training classifiers were obtained using DEGG. The features, SoE, NAPS, and HOS are applied on DEGG. The obtained three features are added and normalized. The combined evidence for each frame is classified as labeled as voiced frame when the evidence value is above 1% mean value [119], otherwise the frame is labeled as an unvoiced frame. The authors have further verified manually the reference labels of the voicing decision estimated from DEGG signal for setting the threshold to get reliable voicing labels as mentioned in [119].

The evidence obtained by the different classifiers are shown in Figure 4.5. The reference voicing label marking along with speech and DEGG are shown in Figure 4.5(a) and (b), respectively. The evidence obtained using the nonlinear mapping of combined glottal activity features show accurate detection of voicing decision, except for the region around 0.8 s as shown in Figure 4.5(c). The evidence obtained using k-NN is given in Figure 4.5 (d). Note that for k-NN classifier, only the mapped evidence is displayed, since the k-NN output gives only a binary value based on the dominant label. Even k-NN

gives a good detection accuracy for this speech sample. The evidence obtained from DBN slightly fails in the region around 0.7 s and 1.25 s as seen in Figure 4.5 (e). The SVM evidence output performs the best as seen in Figure 4.5 (f). It can be seen that the nonlinear mapped value of likelihood probability of SVM is very close to the reference value.

4.4.2 Evaluation parameters

The classifier results are evaluated by finding the percentage of voicing frame error (VU_E) between the reference and detected voicing decision using different classifiers. The voicing frame error is defined as follows

$$VU_E = V_E + U_E \tag{4.10}$$

where V_E and U_E represent the percentage of voiced and unvoiced errors, respectively, given by

$$V_E = V_N/T_{ref} \tag{4.11}$$

$$U_E = U_N/T_{ref} \tag{4.12}$$

where V_N and U_N are the total number of error frames for voicing and unvoicing decisions, respectively. T_{ref} is the total number frames present in the testing database. For evaluation of these parameters, two speakers SLT (female) and BDL (male) are taken from the CMU ARCTIC database available publicly [22] consist of 1132 sentences. A subset of the database around 800 sentences is used for training. The parameters of the classifiers are optimized using exhaustive grid search using fivefold cross-validation to get the best performances. The remaining sentences are used for validating voiced-unvoiced errors.

4.4.3 Results

The voicing decision results obtained from the classifiers k-NN, SVM, and DBN are tabulated in Table 4.2. The results are further correlated with state-of-the-art algorithms such as RAPT, speech transformations and representations using adaptive interpolation weighted spectrum (STRAIGHT), summation of residual harmonics (SRH), and robust epoch and pitch estimator (REAPER) [18, 66, 67, 121]. For all the methods, the glottal activity features are obtained for the frame shift of 5ms. The VU_E error rate for the signal processing method is evaluated using nonlinear mapping. Note that the contextual information is incorporated in the input of classifiers, where a five frame context has been used. The use of this information showed drastic improvement in the results. There may not be much

Table 4.2: Comparison of the different classifiers: represented in terms of the percentage (%) of voiced, unvoiced, and combined voicing error

Database	SLT			BDL			KEELE			CSTR		
	V_E	U_E	VU_E	V_E	U_E	VU_E	V_E	U_E	VU_E	V_E	U_E	VU_E
Nonlinear	3.34	1.85	5.19	4.0	1.92	5.92	4.51	1.92	6.43	4.03	1.98	6.01
k-NN (k=23)	2.84	1.48	4.32	3.41	1.6	5.01	4.06	1.78	5.84	3.67	1.84	5.51
SVM(c=8,Y=16)	**1.72**	**1.02**	**2.74**	**1.81**	**1.15**	**2.96**	**2.01**	**1.18**	**3.19**	**1.92**	**1.1**	**3.02**
DBN (40-25-15-5-2)	3.24	1.71	4.95	3.51	1.94	5.45	2.86	1.5	4.36	2.92	1.79	4.71
RAPT	4.04	2.45	6.49	4.85	3.2	8.05	5.36	3.92	9.28	4.82	2.76	7.58
STRAIGHT	3.6	2.33	5.93	4.02	2.83	6.85	4.65	3.55	8.20	5.12	3.49	8.61
SRH	3.86	2.31	6.17	4.79	2.57	7.36	4.81	4.11	8.92	4.55	2.78	7.33
REAPER	2.02	1.51	3.53	3.35	2.06	5.41	3.97	2.38	6.35	3.65	2.16	5.81

change in the glottal activity. However, when processing on a frame by frame basis, a finer level of processing may result in plenty of mis-classifications. The use of left-context and right-context of the current frame may provide additional information to the classification task and may help to reduce the sudden changes, which inturn helps in reducing the classification error. The results of varying the number of frame level context to the left and right of the current frame are shown in Table 4.3 for SVM and DBN classifiers. From the table, it can be seen that in the frame level context 5, the ARCTIC database (for both SLT and BDL speaker) gives the best results and further increasing the context is not helping in improving the voicing decision.

The SVM classifier performed best among all the classifiers and even superior to state-of-the-art voicing algorithms such as RAPT, STRAIGHT, SRH, and REAPER classifiers. In classifier, SVM performed even better than DBN due to the fact that classification task here is binary classification. Moreover, in this chapter the database used is only around 1 hr (800 sentences), which may not be sufficient for DBN to give good accuracy for this classification task. In addition to the ARCTIC database, the proposed voicing decision is evaluated using KEELE and CSTR databases [132, 133], and compared with other state-of-the-art methods. The trend for these databases also remains similar, with a proposed method having voicing decision error as 3.19 and 3.02, respectively, for KEELE and CSTR databases. The slight decrease in the voicing decision in both the databases due the slightly degraded nature of wave files present in both the databases compared to the ARCTIC database. The same trend is observed in the other methods also.

Table 4.3: Comparison of the frame level context information used in different classifiers: represented in terms of the percentage (%) of voicing error

No. of context	SVM			DBN		
	SLT	BDL	Average	SLT	BDL	Average
0	4.68	6.19	5.43	7.92	9.34	8.63
1	4.55	6.02	5.28	7.68	8.64	8.17
2	3.53	4.27	3.90	7.41	8.25	7.83
3	3.01	3.95	3.48	6.61	7.35	6.98
4	2.78	3.36	3.07	5.86	6.57	6.21
5	**2.74**	**2.96**	**2.85**	**4.95**	**5.45**	**5.20**
6	3.15	4.07	3.61	7.36	8.54	7.95

4.5 Glottal activity features for Synthesis

In this section, integration of proposed voicing decision using glottal activity features for SPSS is explained. The proposed framework of SPSS is provided in Figure 4.6. It provides a unified framework to train vocal tract, excitation, and duration parameters simultaneously in HMM [5]. This framework of SPSS is mainly classified into training and synthesis. In training, excitation, vocal tract, and duration parameters are derived from each phoneme for the whole database. All the phonemes are modeled with 5 states and in each state 3 streams are used to model the different parameters extracted for each phoneme [59]. The first stream consists of vocal tract parameters with 35 MCEP parameter including the zero*th* coefficient, additionally, their delta and delta-delta coefficients are also used. It is trained by continuous density HMMs. The source parameters F0 and its delta and delta-delta coefficients are modeled in a single stream using continuous distribution instead of three independent streams used in the MSD modeling of F0 [59]. In the third stream, glottal activity features SoE, NAPS, and HOS are modeled with continuous distribution along with delta and delta-delta coefficients. The details of each stream and its distribution are as follows:

- The first stream for MCEP and its derivatives with the continuous probability distribution.
- The second stream for fundamental frequency, its delta, and delta-delta with the continuous distribution.

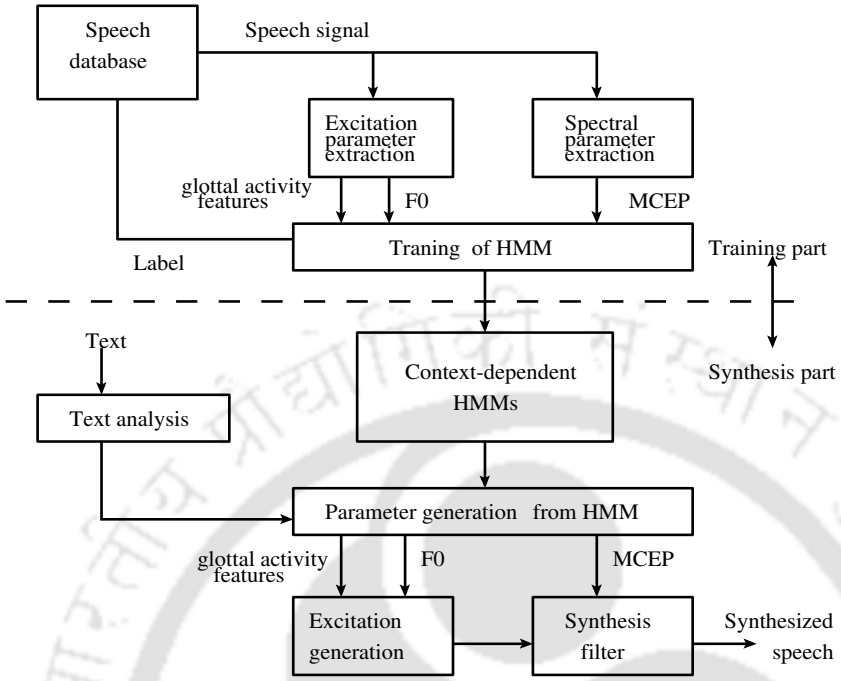

Figure 4.6: Block diagram of SPSS

- The third stream for SoE, NAPS, and HOS features and its derivatives with the continuous probability distribution.

For each phoneme, parameters mentioned above are extracted along with their corresponding labels. In training part, the maximum likelihood estimation of each parameter is computed using Baum-Welch re-estimation algorithm [5]. During synthesis, using the maximum likelihood parameter generation algorithm, frame wise MCEP, glottal activity features, and F0 parameters are generated for a given input text [5]. The generated glottal activity features are used for voicing decision to generate the excitation source signal.

4.5.1 Integration of voicing decision in SPSS

The block diagram of integrating glottal activity features for the voicing decision of SPSS is shown in Figure 4.7. In this chapter, voicing decision for SPSS is tested using HTS framework. The excitation signal is generated according to the voicing decision modeled from the proposed glottal activity features. The voicing decision is obtained from the generated SoE, NAPS, and HOS parameters by passing them through the SVM trained model. The F0 is computed from ZFF. It is based on the sub-segmental analysis by calculating the interval between the successive

epochs [110]. The ZFF method gives an accurate estimation of F0 [71]. The obtained F0 over an epoch interval is averaged over a frame segment to get frame wise F0. During synthesis, the voicing decision is obtained from the trained SVM model, where the input to SVM is the glottal activity parameters generated from HMM. The detailed procedure for integration of glottal activity features to the HMM is explained in the Algorithm 1. For the voiced frame, impulse train is generated. In the case of a unvoiced frame, the white Gaussian noise is generated. The generated excitation is passed through the synthesis filter. In the present chapter, the Mel-log spectral approximation (MLSA) filter is used as synthesis filter. The entire process is done frame wise and the convolved source and system response is overlapped and added to obtain the synthetic speech.

Algorithm 1. Algorithm for computing voicing decision from the proposed glottal activity features in HMM training and synthesis

Training:
Step 1: Compute **F0** from ZFF algorithm, which gives instantaneous **F0** values. To get frame wise **F0** value, average the **F0** values in each frame.
Step 2: Compute glottal activity features SoE and NAPS from ZFFS, and HOS feature from ILPR.
Step 3: Apply glottal activity features with five frame context information to SVM classifier. Optimize cost parameter (c) and width parameter (Y). Save the SVM model.
Step 4: Model **F0** and glottal activity features (SoE, NAPS, and HOS) with two streams in HMM as continuous Gaussian distribution.

Synthesis:
Step 5: Generate the glottal activity parameters and **F0** from the maximum likelihood generation algorithm of HMM for a given text.
Step 6: Obtain the voicing decision from the SVM trained model by feeding glottal activity features (SoE, NAPS, and HOS obtained from HMM) with five frame context.
Step 7: Generate the excitation signal from the voicing decision as shown in Figure 4.7.

end

4.6 Experimental evaluation

To evaluate the proposed glottal activity features for the vocoder framework, HTS system is built for two speakers: SLT (US female) and BDL (US male) [22]. The SLT and BDL speaker consist of 1132 sentences. For training, randomly selected 1000 sentences were used and remaining sentences were used for testing. The parameters are analyzed for a frame size of 25 ms with a frame rate of 5

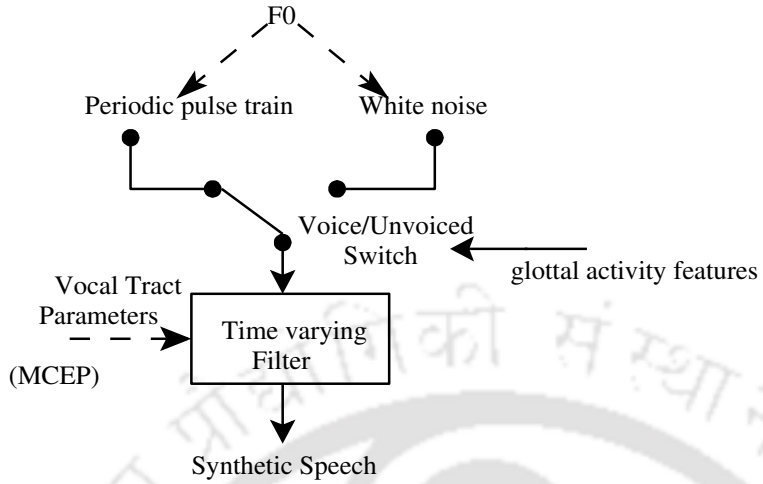

Figure 4.7: Integration of glottal activity features to the vocoder of SPSS

ms. The glottal activity features and F0 parameter are extracted frame wise along with the MCEP representing the vocal tract information. These parameters are trained in the HMM framework. The proposed method is called as glottal activity (GA) continuous method, as the glottal activity features are modeled with a continuous distribution. For the comparison purpose, along with the proposed method, three more systems based on the voicing decision computed from RAPT, STRAIGHT, and REAPER methods are developed in the HMM framework [18, 66, 121]. In all the three methods, F0 is modeled with MSD model, whereas, in the proposed method, F0 is modeled as continuous distribution [59]. The proposed method is also compared with the original continuous model proposed by Yu *et al.* [59], where the voicing decision is modeled in a separate stream using voicing strength.

In RAPT and REAPER methods, the voicing decision is based on the autocorrelation analysis of speech and then by using dynamic programming for decision making. The STRAIGHT method is based on wavelet transform, where based on voicing strength in different bands, the voicing decision is made. In this chapter, version 40 of STRAIGHT is used. The SRH based voicing decision is not used here since its accuracy is less compared to other algorithms. For a fair comparison of the voicing decision, in all the methods F0 is computed from ZFF [71]. Since ZFF method will give a very good estimate of F0 [71]. The speech waveforms are synthesized using MLSA filter. Some of the synthesized samples for all the methods can be accessed from the following link[1]

[1] http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/gadhts.php

4.6.1 Subjective evaluation

In this evaluation, two tests are conducted, namely, mean opinion score (MOS) and preference test (PT) for all the SPSS system. In MOS test, 25 sentences which are not used in training are given to subjects along with the original waveform. The subjects were asked to give the ratings with the scale of 1 to 5 (1: poor and 5: excellent) to the wave files by comparing with the original wave file. Two groups of subjects were used in the subjective evaluation. They are 10 speech experts and 15 naive listeners. The evaluators are not from native English, however, they are fluent in speaking English. For evaluations, listeners were asked to examine naturalness present in each file and give their overall scores accordingly. The average scores obtained from expert listeners are given in Table 4.4. From the table, it can be seen that the proposed voicing decision method outperform RAPT, STRAIGHT, and REAPER algorithms. Moreover, the proposed method is marginally superior to continuous model with MOS score of 3.46 signifies the importance of accurate estimation of voicing decision to improve the naturalness of the synthetic speech.

Table 4.4: Subjective evaluation results of MOS and PT with 95% confidence intervals from expert subject

Experimental	**SPSS system using different types of voicing decision**						**p value**
Evaluation	**GA continuous**	**RAPT**	**STRAIGHT**	**REAPER**	**Continuous**	**Same**	
MOS	3.36	2.62	2.96	3.06	3.12	-	
PT	72%	20%	-	-	-	8%	4.38×10^{-9}
	53%	-	31%	-	-	16%	2.12×10^{-5}
	47%	-	-	35%	-	18%	1.56×10^{-2}
	45%	-	-	-	34%	21%	1.92×10^{-1}

Table 4.5: Subjective evaluation results of MOS and PT with 95% confidence intervals from naive subjects

Experimental	**SPSS system using different types of voicing decision**						**p value**
Evaluation	**GA Continuous**	**RAPT**	**STRAIGHT**	**REAPER**	**continuous**	**Same**	
MOS	3.21	2.65	2.82	2.96	3.04	-	
PT	58%	12%	-	-	-	28%	3.19×10^{-7}
	44%	-	24%	-	-	32%	2.07×10^{-4}
	39%	-	-	28%	-	33%	1.78×10^{-3}
	33%	-	-	-	31%	36%	4.14×10^{-1}

In addition, to know the distribution of scores for male and female speakers, the bar chart is plotted in Figure 4.8. It shows MOS score with a standard deviation of all five systems. From the bar plot,

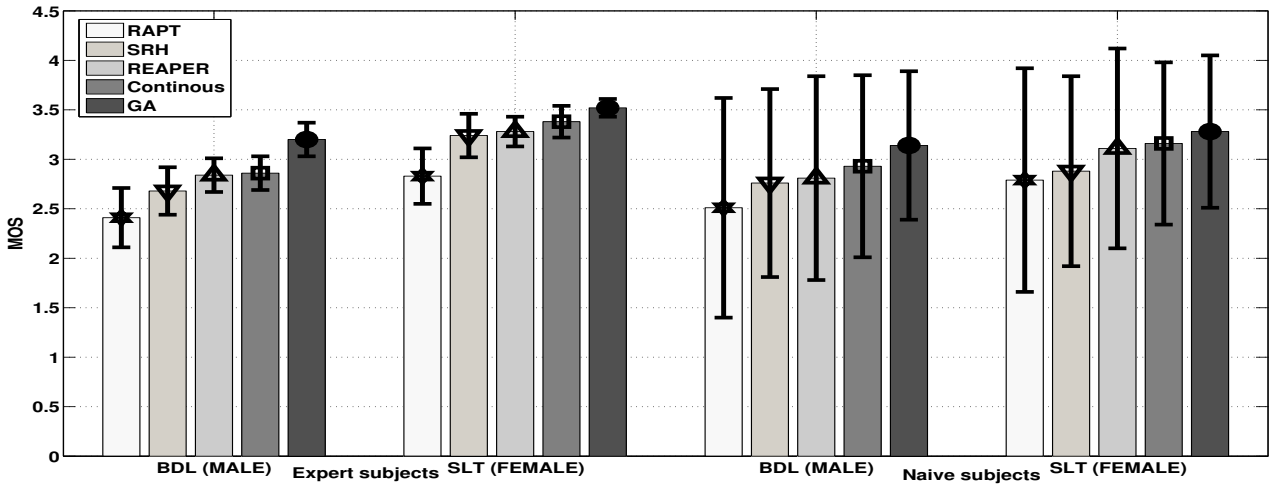

Figure 4.8: Average MOS of five different SPSS systems with RAPT, STRAIGHT, REAPER, Continuous, and GA model based voicing decision, respectively, for SLT (female speaker) and BDL (male speaker)

it can be seen that the proposed method for both male and female speaker is slightly better than the continuous method. Results for naive listeners are presented in Table 4.5. The conclusions that can be drawn from these results are similar to that of expert listeners, except that MOS scores for naive listeners are relatively lower. This can be explained by the fact that compared to expert listeners, naive listeners pay less attention to the minute variations in the naturalness. Further, contrary to expert listeners, naive listeners having higher variance compared to expert listeners. Nonetheless, it is appreciable to note that even the naive listeners are able to observe differences between proposed GA based SPSS system with the other methods with a considerable difference in MOS score. This is consistent with both male and female speakers as shown in Figure 4.8.

In the preference test, for each sentence, subject were requested to listen to two versions of the system shuffled randomly from five systems at a time and asked to choose any one system or prefer the same as their preference. The percentage of preference scores with p value from expert listeners can be viewed in Table 4.4. A clear preference for the proposed method over RAPT, STRAIGHT, REAPER, and continuous voicing methods can be observed according to the preference test and the p values given by hypothesis tests. In the case of naive listeners, similar to MOS test, preference to GA system is higher when compared with RAPT system with a preference score of 72%. The preference between STRAIGHT/REAPER/continuous algorithms with the proposed GA method is slightly reduced. The reason for this may be that naive listeners not able to distinguish small errors occurring due to the voicing decision. The proposed glottal activity features based voicing decision

using continuous modeling outperforms base version of the continuous voicing model where only voicing strength is used. This signifies the contributions of other glottal activity features like NAPS and HOS for the improvement in speech quality.

4.6.2 Objective evaluation

In this chapter, three objective measures are used, namely, PESQ [106], log spectral distance (LSD), and voicing frame error (VU_E). The duration of synthesized speech and original speech may not be of the same length. Hence, alignment of original and synthesized speech is done using dynamic time warping algorithm. The PESQ measure should be interpreted as an MOS regarding the similarity to the original waveform. The PESQ scores obtained for all the types of voicing decision are tabulated in Table 4.6. It can be observed from the table that the proposed voicing decision is having a PESQ score of 1.59 with the standard deviation of ± 0.02, which is better than all the other voicing decision methods. This signifies an improvement in the naturalness of synthetic speech by the proposed method.

Second objective evaluation is the LSD measure gives distortion error in the spectral domain. The reason for choosing this measure is due to voicing error, voiced frame may classify as a unvoiced frame, and vice versa. This mismatch results in a change in excitation and reflected in the synthesized speech spectrum. The lower LSD value indicates smaller distortion and better the synthesis quality. This measure is evaluated between reference original speech and synthesized speech for the same text. The average LSD for all the five voicing methods given in Table 4.6 along with standard deviation. The LSD for the proposed method is lesser with distortion of 2.05, indicating the better spectral modeling of proposed method compared to the continuous F0 model and as well as the MSD methods such as RAPT, STRAIGHT, and REAPER algorithms. Similarly, as mentioned in Section 4.4 (c), the VU_E error rate is also evaluated for the synthesized waveforms. From Table 4.6, it can be seen that VU_E values are relatively higher in the case of synthesized files when compared to the evaluation done for the original waveforms on the whole database. This may be due to the fact that HMM training errors are also included in the synthesized files. However, the VU_E error rate in the synthesized waveforms is lesser for the proposed method. This signifies the importance of modeling features present in the glottal activity regions for improving the naturalness of SPSS framework.

Table 4.6: Objective evaluation results of PESQ, LSD, and VU_E with standard deviation

Experimental Evaluation	SPSS using different types of voicing decision				
	GA Continuous	**RAPT**	**STRAIGHT**	**REAPER**	**Continuous**
PESQ	1.59±0.02	1.21±0.03	1.20±0.06	1.41±0.03	1.46±0.03
LSD	2.05±0.26	2.64±0.27	2.49±0.32	2.43±0.26	2.27±0.29
VU_E(%)	4.61	11.02	9.15	7.12	5.93

4.7 Summary

This chapter demonstrated the significance of different features present in the glottal activity region for speech synthesis. The feature studied in this chapter includes epoch location, epoch strength, phase component, and voicing decision. A robust voicing decision using glottal activity features is proposed for SPSS. The voicing decision is performed based on the SoE, NAPS, and HOS features, which represent strength, periodicity, and asymmetrical nature of glottal activity regions, respectively. The features used for voicing decision are further refined using different classifiers like k-NN, SVM, and DBN. The SVM performed best for voicing decision. The performance of the proposed approach is compared with the current pitch estimation based voicing decision algorithms. The results demonstrated that the classification error of proposed approach is notably less compared with other methods. The glottal activity features SoE, NAPS, and HOS are modeled along with F0 and MCEP using continuous distribution in SPSS. The quality of synthesized speech using the proposed approach is compared with three SPSS systems developed using RAPT, STRAIGHT, and REAPER voicing decision methods. The proposed method is also compared with the continuous F0 model. The objective and subjective assessment results show that the proposed voicing decision using glottal activity features for SPSS significantly reduces voicing decision errors. It also helps in improving the naturalness of synthesized waveforms when compared with other methods.

In particular, in this chapter, importance of accurate estimation of voicing decision using features present in the glottal activity region is shown. In the subsequent chapters, other features present in the glottal activity region like vocal tract features, aperiodic components, and phase components are explored for improving the quality of speech.

5

Riesz Transform for Speech Synthesis

Contents

5.1 Motivation behind the 2-D processing 86
5.2 The Riesz transform based demodulation 88
5.3 Riesz transform based demodulation for speech spectrum 91
5.4 Carrier spectrogram analysis 94
5.5 Synthesis methodology 98
5.6 Experimental validation 102
5.7 Summary 109

Objective

In this chapter, 2-D spectro-temporal analysis/synthesis method is proposed using Riesz transform. In the traditional amplitude-modulation, frequency-modulation (AM-FM) method, demodulation is based on the fixed short time frame analysis, which results in errors for both source and filter parameter estimation. The 2-D spectro-temporal analysis is motivated by the fact that the human auditory cortex is tuned to localized spectro-temporal modulations. The spectro-temporal receptive fields of these cortical cells look like 2-D spectro-temporal Gabor filters. The demodulation of 2-D spectro-temporal patches using Riesz transform yields smoothed spectral envelope, carrier spectrogram, and coherence map, representing vocal tract spectrum, source signal, and periodicity, respectively, in a single framework. The analysis/synthesis representation gives better synthesis quality than the state-of-the-art STRAIGHT vocoder, which is based on the pitch-synchronous analysis. Further, smoothed spectral envelope and coherence map are compactly represented using Mel-cepstral coefficients (MCEP) and trained in the HMM framework. The compactly represented parameters are used as input to the synthesis module of vocoder framework. The synthesized files are measured using objective and subjective evaluation. The results show that decomposition of the spectrum into an envelope, carrier, coherence map is equally effective like STRAIGHT method.

5.1 Motivation behind the 2-D processing

Recently, sophisticated analysis/synthesis algorithms have been developed to amend this problem, and their ultimate goal is to provide natural-sounding intelligible speech. Most of these algorithms focused on the analysis/synthesis framework. The speech signal is a non-stationary signal and contains variations in amplitude level, frequency level, and phase level. To analyze these characteristics, the signal can be processed either using the short-time analysis, pitch-synchronous analysis, or spectro-temporal analysis. Short time analysis involves processing of the speech signal with a fixed 20 to 30 ms. It works well for various applications in speech coding and recognition, where the pitch is slowly varying. Some of the examples of these types of processing are short-time Fourier transform analysis, linear prediction analysis, and sinusoidal analysis [77,80,90]. For the applications like emotional speech analysis, children speech recognition, *etc.*, the pitch-synchronous analysis is useful, however, it needs prior pitch estimation. Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum (STRAIGHT) is quite popular among the pitch synchronous analysis algorithms,

which uses a combination of adaptive Gaussian pitch synchronous windows for analysis [18]. The STRAIGHT method gives smoothed spectral envelope which is free from excitation interferences such as harmonicity. It is successfully used in speech synthesis application as it gives accurate vocal tract spectrum even in presence of pitch varying conditions and also gives the aperiodic decomposition of speech for each segment.

The spectro-temporal analysis method tries to focus on exploiting the joint temporal and spectral characteristics to capture the slowly varying vocal tract shape in the 2-D domain. The significance of the spectro-temporal analysis comes from auditory neuro-physiological studies where it has been shown that there are specialized neurons in the primary auditory cortex that respond to specific spectro-temporal patterns [134]. These patterns of cortical cells which look like 2-D spectro-temporal Gabor filters in the auditory cortex have motivated similar signal processing strategies for speech analysis. In [135], Quatieri *et al.* modeled the speech spectrum patch as multicomponent amplitude and frequency (AM-FM) modulated 2-D sinusoidal signal where the demodulation is achieved by estimating the carrier parameters from *Grating Compression Transform* domain [136]. The method proposed by Quatieri *et al.* needs prior pitch estimation. Hence, the accuracy of the demodulation critically depends on pitch estimation.

In this chapter, we proposed the unified framework in the 2-D domain to demodulate the speech signal into the AM and FM components using Riesz transform. It is an extension of Hilbert transform in a higher dimension and this method does not require any prior pitch estimation like other demodulation algorithms [137,138]. The spectro-temporal joint processing helps in removing both temporal and spectral variations presence due to the harmonic and windowing effect. The obtained AM component is smoothed vocal tract envelope without any source interference. Hence, this smoothed envelope is compactly represented using MCEP features. In addition, proposed demodulation also preserves the carrier spectrogram. The Carrier spectrogram is used to compute the fundamental frequency. Further, from the carrier spectrogram, coherence map is computed, which gives the amount of harmonic and non-harmonic components present in a particular band of the speech spectrum. This coherence map is also compactly represented using MCEP features. Next, we develop analysis/synthesis module by modulating the smoothed envelope and carrier spectrogram obtained from parametrized coefficients with different types of phase representation. The different phase representation includes F0 based zero phase excitation, random phase excitation, and aperiodic excitation. Further, as an application, the

parametrized coefficients are tested in the HMM based statistical speech synthesis framework.

5.2 The Riesz transform based demodulation

In this section, we briefly introduce the demodulation of AM-FM 1-D signal using Hilbert transform. Complex Riesz transform can be considered as a 2-D extension of Hilbert transform. We discuss few properties of the complex Riesz transform operator. In order to show the key idea for 2-D demodulation using Riesz transform, we take a stylized example of amplitude modulated mono-component 2-D sinusoid and develop the demodulation technique using quadrature property of Riesz transform.

5.2.1 Hilbert Transform

Hilbert transform was first proposed by Gabor [139], showing its importance for the representation of analytic signal. The frequency response of Hilbert transform is characterized by $H(\omega) = -j\frac{\omega}{|\omega|}$, where $\omega \in \mathbb{R}$ and $j = \sqrt{-1}$. Denoting Hilbert transform of a scalar function $f(t) : \mathbb{R} \rightarrow \mathbb{R}$ by $\mathcal{H}(.)$, consider an amplitude modulated cosine $f(t) = v(t)\cos(2\pi\omega_0 t + \theta)$ where $v(t)$ is a non-negative and slowly varying function referred as envelope, and $\cos(2\pi\omega_0 t + \theta)$ is referred as carrier signal. Under certain conditions on the spectrum of envelope and carrier (Bedrosian theorem [140, 141]), it is straightforward to show the following:

$$\mathcal{H}\{v(t)\cos(2\pi\omega_0 t + \theta)\} = v(t)\sin(2\pi\omega_0 t + \theta) \tag{5.1}$$

where $\sin(2\pi\omega_0 t + \theta)$ is referred as quadrature component of $\cos(2\pi\omega_0 t + \theta)$ and the property in Eq. 5.1 is referred as quadrature property of the Hilbert transform. Envelope and carrier estimates can be obtained by constructing a complex analytic signal $f_a(t)$ as follows,

$$\begin{aligned} f_a(t) &= v(t)\cos(2\pi\omega_0 t + \theta) + j\ \mathcal{H}\{v(t)\cos(2\pi\omega_0 t + \theta)\} \\ &= v(t)\cos(2\pi\omega_0 t + \theta) + j\ v(t)\sin(2\pi\omega_0 t + \theta) \\ &= v(t)e^{j(2\pi\omega_0 t + \theta)} \end{aligned}$$

From the analytic signal representation, envelope and carrier signal estimates are given by $v(t) \propto |f_a(t)|$ and $\cos(\angle(f_a(t)))$, respectively. In the next section, we extend the similar concepts in 2-D using Riesz transform.

5.2.2 Riesz Transform

Riesz transform is scalar function to vector function mapping $f(\boldsymbol{\omega}) \rightarrow f_{\mathcal{R}}(\boldsymbol{\omega})$ defined as follows [142],

$$f_{\mathcal{R}}(\boldsymbol{\omega}) \triangleq \begin{pmatrix} f_1(\boldsymbol{\omega}) \\ f_2(\boldsymbol{\omega}) \end{pmatrix} = \begin{pmatrix} (h_t * f)(\boldsymbol{\omega}) \\ (h_w * f)(\boldsymbol{\omega}) \end{pmatrix}$$

where $f(\boldsymbol{\omega})$ is a scalar function and $f_{\mathcal{R}}(\boldsymbol{\omega})$ is its Riesz transform. $h_t(\boldsymbol{\omega})$ and $h_w(\boldsymbol{\omega})$ are impulse responses of filters associated to Riesz transform along time and frequency axis, respectively, and given by [142]:

$$h_t(\boldsymbol{\omega}) = \frac{t}{2\pi||\boldsymbol{\omega}||^3}, \quad \text{and} \quad h_w(\boldsymbol{\omega}) = \frac{\omega}{2\pi||\boldsymbol{\omega}||^3},$$

The respective frequency responses of the filters are given by,

$$\hat{h}_t(\boldsymbol{\Omega}) = -j\frac{\Omega_t}{||\boldsymbol{\Omega}||}, \quad \text{and} \quad \hat{h}_w(\boldsymbol{\Omega}) = -j\frac{\Omega_w}{||\boldsymbol{\Omega}||},$$

where $||\boldsymbol{\Omega}|| = \sqrt{\Omega_t^2 + \Omega_w^2}$.

where Ω_t and Ω_w denoting frequencies in time-axis and frequency-axis of the spectrogram, respectively. The frequency response of complex Riesz transform (CRT) is written as follows,

$$\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega}) = \hat{h}_t(\boldsymbol{\Omega}) + j\hat{h}_w(\boldsymbol{\Omega}) = \frac{-j\Omega_t + \Omega_w}{\sqrt{\Omega_t^2 + \Omega_w^2}} \tag{5.2}$$

From Eq. 5.2 it is straightforward to show that $\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})$ has unit magnitude and odd symmetric function of $\boldsymbol{\Omega}$, $\hat{h}_{\mathcal{R}}(-\boldsymbol{\Omega}) = -\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})$. CRT is 2-D all-pass filter and has phase response given by,

$$\angle\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega}) = \tan^{-1}\left(-\frac{\Omega_t}{\Omega_w}\right)$$

The phase response of CRT kernel is spiral in nature [138].

5.2.3 Demodulation in 2-D using Riesz transform

Consider an amplitude modulated 2-D cosine with spatial frequencies $\Omega_0 \cos\theta_0$ and $\Omega_0 \sin\theta_0$ along t-axis and w-axis, respectively, with orientation θ_0, and let $\boldsymbol{\Omega} = (\Omega_0 \cos\theta_0, \Omega_0 \sin\theta_0)$, then, 2-D cosine signal is written as follows,

$$\begin{aligned} f(\boldsymbol{\omega}) &= V(\boldsymbol{\omega})\ \cos(\Omega_0(t\ \cos\theta_0 + w\ \sin\theta_0)) \\ &= \frac{1}{2}V(\boldsymbol{\omega})(e^{j\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle} + e^{-j\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle}) \end{aligned} \tag{5.3}$$

where $V(\boldsymbol{\omega})$ is a non-negative, smooth and slowly varying 2-D function referred as envelope. The goal is to obtain the quadrature component $\sin(\Omega_0(t\ \cos\theta_0 + w\ \sin\theta_0))$ of the amplitude modulated 2-D cosine. Denoting CRT operator by $\mathcal{R}(.)$ and using eigenfunction function property of linear shift invariant systems [143], the Riesz transform of $f(\boldsymbol{\omega})$ in Eq. 5.3 can be written as follows,

$$\begin{aligned} \mathcal{R}f(\boldsymbol{\omega}) &= \frac{1}{2}V(\boldsymbol{\omega})\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega}_0)e^{j\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle} + \hat{h}_{\mathcal{R}}(-\boldsymbol{\Omega}_0)e^{-j\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle}) \\ &= \frac{1}{2}V(\boldsymbol{\omega})\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega}_0)(e^{j\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle} - e^{-j\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle}) \\ &= e^{j\theta_0}\ V(\boldsymbol{\omega})\sin(\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle) \end{aligned} \tag{5.4}$$

$$\implies \underbrace{e^{-j\theta_0}\mathcal{R}}_{\text{vortex operator}}\ V(\boldsymbol{\omega})\cos(\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle) = V(\boldsymbol{\omega})\sin(\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle)$$

where we have used Eq. 5.2 to evaluate $\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})$ at $\boldsymbol{\Omega}_0$ and odd symmetry property of CRT. The CRT operator pre-multiplied by the factor $e^{-j\theta_0}$ is referred as vortex operator [137] and denoted by $\mathcal{V}$, thus

$$\mathcal{V}\{V(\boldsymbol{\omega})\cos(\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle)\} = V(\boldsymbol{\omega})\sin(\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle) \tag{5.5}$$

The orientation θ_0 can be estimated using structure tensor method [144]. Next, we construct a complex signal by combining 2-D cosine and its quadrature component as follows,

$$\begin{aligned} f_a(\boldsymbol{\omega}) &= V(\boldsymbol{\omega})\cos(\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle) + j\ V(\boldsymbol{\omega})\sin(\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle) \\ &= V(\boldsymbol{\omega})e^{j\langle\boldsymbol{\Omega}_0,\boldsymbol{\omega}\rangle} \end{aligned} \tag{5.6}$$

This construction is similar to the construction of analytic signals using Hilbert transform in 1-D. The estimates of envelope and carrier signal are given by $V(\boldsymbol{\omega}) \propto |f_a(\boldsymbol{\omega})|$ and $cos(\angle f_a(\boldsymbol{\omega}))$, respectively.

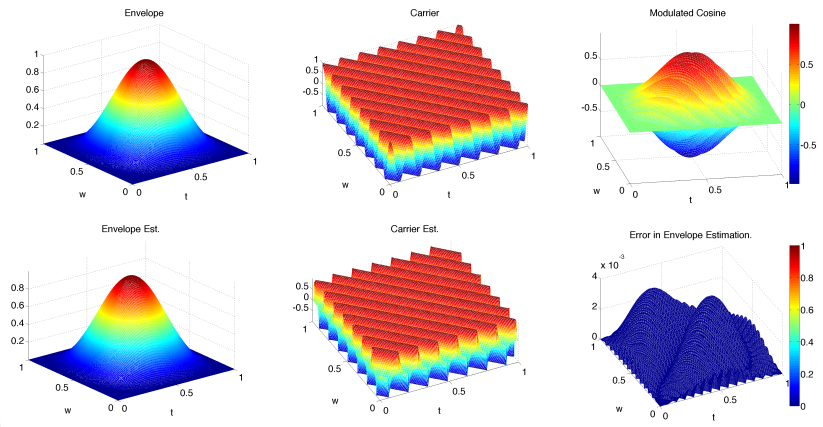

Figure 5.1: (a) Modulating signal: 2-D hamming window $V(\omega)$, (b) Carrier signal: 2-D cosine $\cos(\langle \mathbf{\Omega}_0, \boldsymbol{\omega} \rangle)$ at $\Omega_0 = 20\pi$ and $\theta_0 = \pi/4$, (c) Amplitude Modulated signal using hamming window, (d) Estimated envelope using CRT, (e) Estimated carrier using CRT, and (f) Error is envelope estimation.

Figure 5.1 illustrates the Riesz transform based demodulation of a 2-D cosine.

5.3 Riesz transform based demodulation for speech spectrum

The spiral based demodulation provides a single framework to extract different components of speech such as smooth AM envelope, carrier spectrogram, and the coherence map representing the vocal tract spectrum, dynamics of harmonics, and harmonic/inharmonic components of the speech signal, respectively. In the next section, we discuss the details of these components.

5.3.1 AM envelope

Aragonda *et al.* [145], proposed the demodulation of all voiced speech using Riesz transform. In this work, the demodulation is extended for continuous speech sounds. The Riesz transform based demodulation successfully captures the spectro-temporal dynamics assuming that vocal tract parameters are slowly varying relative to pitch and such an assumption is valid for speech spectrograms. Figure 5.2(a) shows smoothed spectral envelope obtained from Riesz transform in comparison to the envelope obtained from STRAIGHT shown in Figure 5.2(b). The plot is shown for a speech segment taken from Hindi database [146]. The significance of the smoothed envelope by the 2-D approach

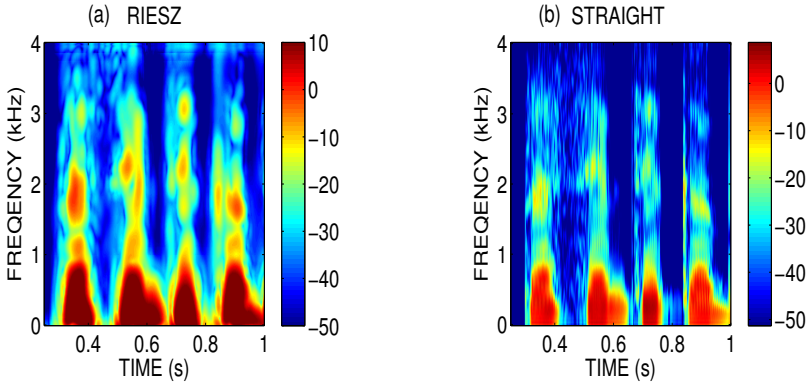

Figure 5.2: Smoothed spectral envelope obtained from (a) Riesz transform; (b) STRAIGHT method

is analyzed by estimating formant frequencies from the demodulated smoothed envelope. For this experiment, we used VTR database [147]. This database consists of manually corrected first four formant tracks. The evaluation was done on the test set containing 8 sentences (a subset of the TIMIT database [148]) spoken by 16 male and 8 female speakers resulting in a total of 192 sentences. The wave files were down-sampled from a sampling rate of 16 kHz to 8 kHz before further processing. For demodulation, the spectrogram was computed using a 512-point discrete Fourier transform. From the spectrogram, spectro-temporal patches were obtained with a size of 100 ms along the time axis and 600 Hz/1000 Hz along the frequency axis covering 3-4 pitch harmonics (600 Hz for male speaker and 1000 Hz for female speaker).

The accuracies of the first four estimated formants were measured in terms of the detection rate (DR). The detection of formant values was counted when formant values were found either within 20% deviation or 300 Hz absolute deviation of the reference value, whichever was smaller. The average DR computed for VTR database is (in %) given in Table 5.1. As a comparison, formants derived from the LP method and the STRAIGHT method is also given in table [18, 77]. We observe that the Riesz transform based approach performs significantly better for all formant locations when compared with LP method. In comparison with the STRAIGHT, proposed method is significantly better for F_3 and F_4 estimation whereas it shows comparable performance for F_1 and F_2 estimation.

It should be noted that the STRAIGHT spectrum is estimated using the version of the code STRAIGHTv407f available publicly [149].

Table 5.1: Performance of formant extraction in terms of detection rate using different method

Formants	Formant estimation using		
	LPC	**STRAIGHT**	**Riesz**
F1	97.43	**99.75**	99.21
F2	91.26	**95.72**	**95.86**
F3	69.01	72.45	**79.63**
F4	29.32	25.31	**35.31**

5.3.2 Carrier spectrogram

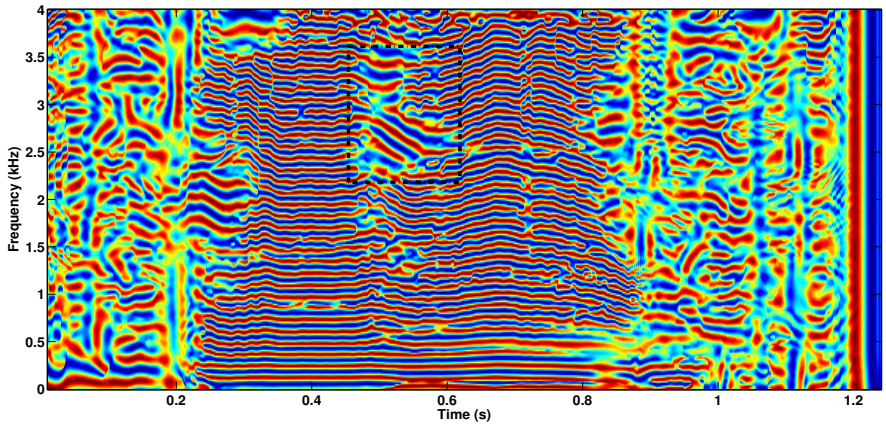

Figure 5.3: Carrier spectrogram computed for diphthong sound unit /ai/, where sound unit is present from 0.2 s to 0.85 s and nonharmonic component shown in rectangle dotted line around 0.5 s region.

The demodulation of NB spectrogram yields carrier spectrogram of the speech signal. Figure 5.3 shows the carrier spectrogram obtained for /ai/ diphthong sound unit. The carrier spectrogram consists of two types of spectro-temporal regions, harmonic and inharmonic. The smooth spectro-temporal variations of the fundamental frequency (F0) are present in harmonic regions corresponding to voiced sounds, whereas, inharmonic regions show different patterns in higher frequency bands which can be used as an indicator of aperiodic components present in the carrier spectrogram as shown in Figure 5.3. Carrier spectrogram also shows the spectro-temporal transitions from voiced to unvoiced sounds. In order to detect transitions between voiced to unvoiced sounds, we compute coherence map from carrier spectrogram. Hence, the carrier spectrogram can be used for voiced/unvoiced decisions of the speech signal and fundamental frequency estimation.

5.4 Carrier spectrogram analysis

5.4.1 Coherence map

Coherence map is a 2-D time-frequency map computed from carrier spectrogram. Coherence is computed by dividing carrier spectrogram into overlapping time-frequency patches (as in the case of NB spectrogram demodulation), at a later stage, the patches are overlap-added to get the full coherence map (Section-V). For each entry of carrier spectrogram matrix a 2×2 structure tensor matrix [138] is computed which is defined as follows,

$$J(\boldsymbol{\omega}) \triangleq \begin{bmatrix} (\psi * f_1^2)(\boldsymbol{\omega}) & (\psi * f_1 f_2)(\boldsymbol{\omega}) \\ (\psi * f_1 f_2)(\boldsymbol{\omega}) & (\psi * f_2^2)(\boldsymbol{\omega}) \end{bmatrix}$$

where $f_1(\boldsymbol{\omega})$ and $f_2(\boldsymbol{\omega})$ are the Riesz transforms along time and frequency axis, respectively. ψ is a Gaussian function with standard deviation σ. The eigenvalues and corresponding eigenvectors of structure tensor matrix give the distribution of gradient of underlying input image [138]. The relative discrepancy between the two eigenvalues of a carrier patch is an indicator of the degree of uniformity or periodic component present in that patch. This attribute is quantified by the coherence $C(\boldsymbol{\omega})$ defined as follows,

$$C(\boldsymbol{\omega}) \triangleq \begin{cases} \left(\frac{\lambda_1(\boldsymbol{\omega})-\lambda_2(\boldsymbol{\omega})}{\lambda_1(\boldsymbol{\omega})+\lambda_2(\boldsymbol{\omega})}\right)^2, & \lambda_1(\boldsymbol{\omega}) + \lambda_2(\boldsymbol{\omega}) \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\lambda_1(\boldsymbol{\omega})$ and $\lambda_2(\boldsymbol{\omega})$ are the eigen values of structure tensor matrix $J(\boldsymbol{\omega})$. Please note that explicit computation of coherence is possible by finding trace and determinant of structure tensor matrix $J(\boldsymbol{\omega})$. Coherence map takes on continuous values between 0 and 1, for harmonic regions coherence values are close to 1 and for inharmonic regions, the values are close to 0.5.

In general, for a perfectly periodic signal, the carrier spectrogram will have no spectro-temporal frequency modulations (flat harmonics' frequency bands), in that case, coherence map will have values equal to 1 (Figure 5.5(b)). Speech sounds are assumed to be periodic signal over a short segment. However, there will always be some perturbations in the fundamental frequency of excitation source signal in each glottal cycle. These factors introduce deviations from the precise repetition of the waveform in each glottal cycle and cause spectro-temporal variations of the 2-D carrier. If a carrier patch has slowly varying spectro-temporal structure such as the case of voiced speech, coherence values will be close to 1 whereas if there is an inconsistency in the spectro-temporal structure (unvoiced

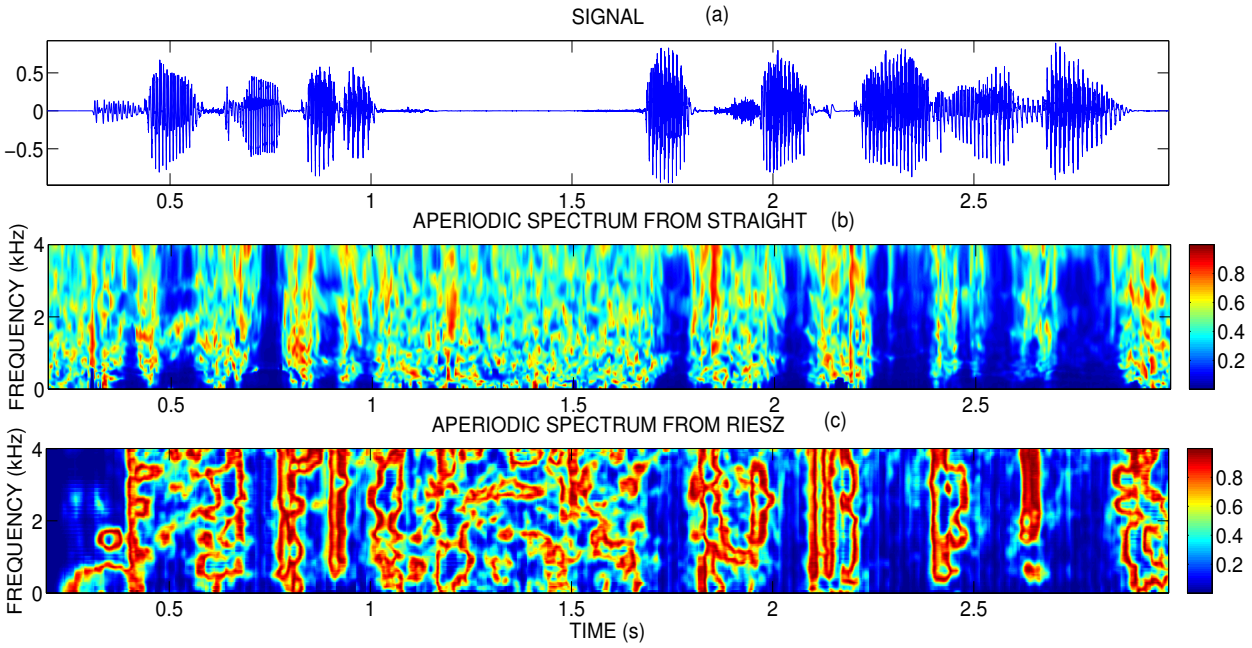

Figure 5.4: Aperiodic spectrum computed for speech utterance taken from TIMIT database: (a) speech signal; (b) and (c) shows the aperiodic spectrum computed from STRAIGHT and Riesz transform, respectively.

sounds, voiced to unvoiced or unvoiced to voiced transitions), coherence values will be close to 0. Hence, coherence map provides the time-frequency mapping of the harmonic (periodic) and inharmonic (aperiodic) component present in each time-frequency patch. In addition, the coherence map also shows clear boundaries between voiced and unvoiced regions of the speech signal.

Thus, coherence map can be used for the decomposition of different frequency bands of speech into periodic or aperiodic components. The aperiodic component ($A(\boldsymbol{\omega})$) for a patch is related to coherence by the following relation:

$$A(\boldsymbol{\omega}) = 1 - C(\boldsymbol{\omega}). \tag{5.7}$$

The computed $A(\boldsymbol{\omega})$ for each patch is overlap-added to get aperiodic map and is shown in Figure 5.4(c) for a speech utterance taken from TIMIT database [148]. For comparison, aperiodicity map computed from the STRAIGHT is also shown in the Figure 5.4(b). For voiced portion aperiodic component computed from Riesz transform looks more distinct compared to STRAIGHT method.

5.4.2 Voicing decision

Consider a frequency band 0 to 1000 Hz in coherence map shown in Figure 5.5(b), moving along the time axis within this band, one can observe that the coherence values are close to 1 for voiced speech

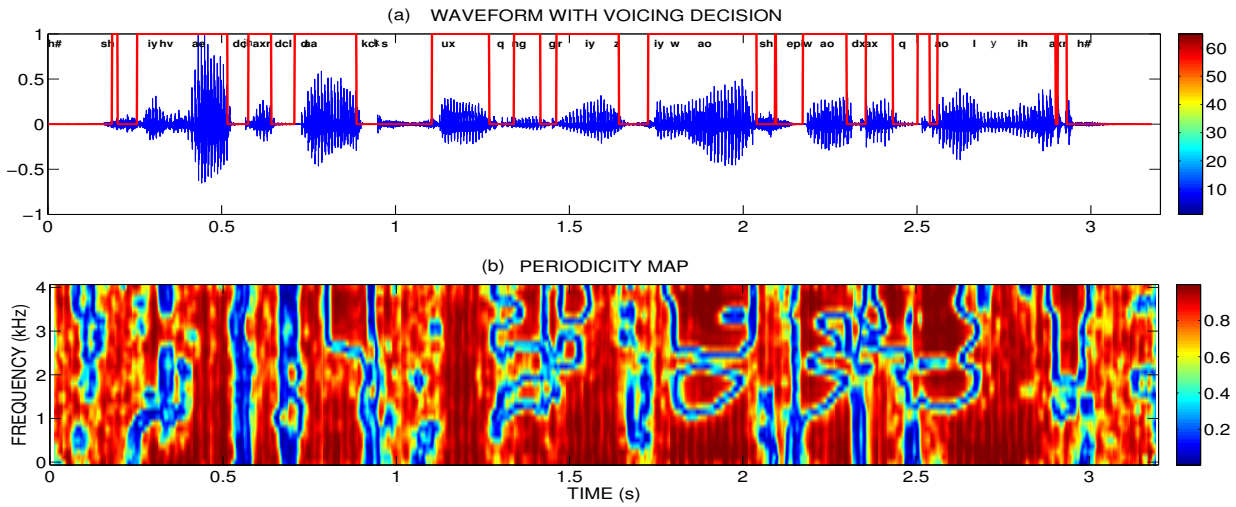

Figure 5.5: Voicing decision with corrections for silence frames using Energy of frame: (a) speech signal taken from TIMIT database with voicing decision and transcription ; (b) coherence map showing the evidence of voicing region

frames and close to 0 for unvoiced/silence frames. The voiced and unvoiced separation is achieved by simple thresholding of coherence map values. The threshold for each speech frame was chosen as the mean of coherence values within the specified frequency band. Figure 5.5(b), shows the coherence map computed for TIMIT sentence shown in Figure 5.5(a). It can be seen that even in the silence frames, coherence values are high at random instants. Hence, in order to avoid the misclassification, speech and silence regions are separated using the short time average energy calculated frame-wise. The computed voicing decisions are shown in Figure 5.5(a) with a red line along with speech signal.

5.4.3 Pitch estimation

The fundamental frequency for each voiced frame is estimated using carrier spectrogram and coherence map. Coherence map is used to separate voiced/unvoiced speech frames. We consider 50 to 1000 Hz frequency band which has a rich harmonic structure as is evident from the carrier spectrogram (Figure 5.3). For reliable pitch estimates, the specified frequency band is further divided into sub-bands each of width 350 Hz with a shift of 110 Hz. Since carrier is almost sinusoidal at a given time location, the pitch can be estimated. For each sub-band candidate pitch values are estimated by taking Fourier transform of carrier slice and by finding the location of the dominant peak in quefrency domain as shown in Figure 5.6. The final pitch value for a voiced frame is obtained by averaging all candidate pitch values for that particular frame. Figure 5.7(c) shows the pitch map calculated for one

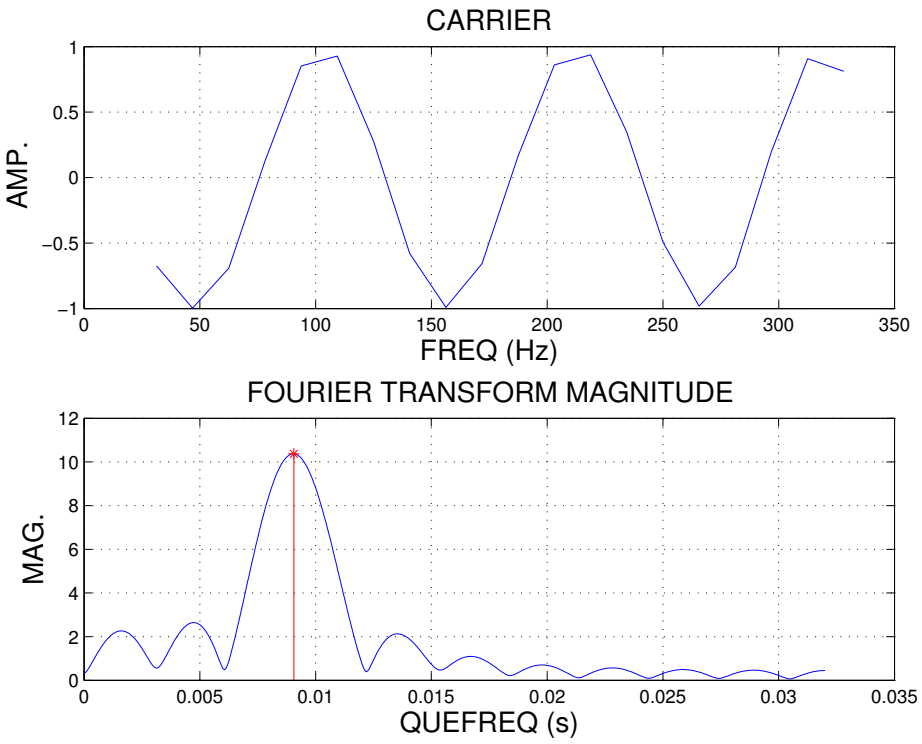

Figure 5.6: (a) Carrier slice for a voiced frame and for a frequency sub-band from 0 to 350 Hz, (b) Fourier transform magnitude.

TIMIT sentence. The pitch map is computed by finding the pitch values in all the sub-bands and overlap-added.

Table 5.2: Pitch estimation from KEELE and CSTR database

	KEELE		**CSTR**	
	F0 estimation using		**F0 estimation using**	
Parameter	**Riesz carrier**	**ZFF**	**Riesz carrier**	**ZFF**
GE (%)	9.08	10.27	7.24	6.17
ME (Hz)	4.41	5.31	4.88	3.93

The evaluation of the proposed method is performed on the pitch databases, namely, KEELE and CSTR database [132, 133]. The KEELE database consists of simultaneously recorded speech and reference pitch computed from Laryngograph. This database consists of one utterance recorded by 5 male and 5 female speakers sampled at 20 kHz. The CSTR database consists of 50 speech sentences uttered by one male and one female speaker recorded along with Laryngograph waveform sampled at

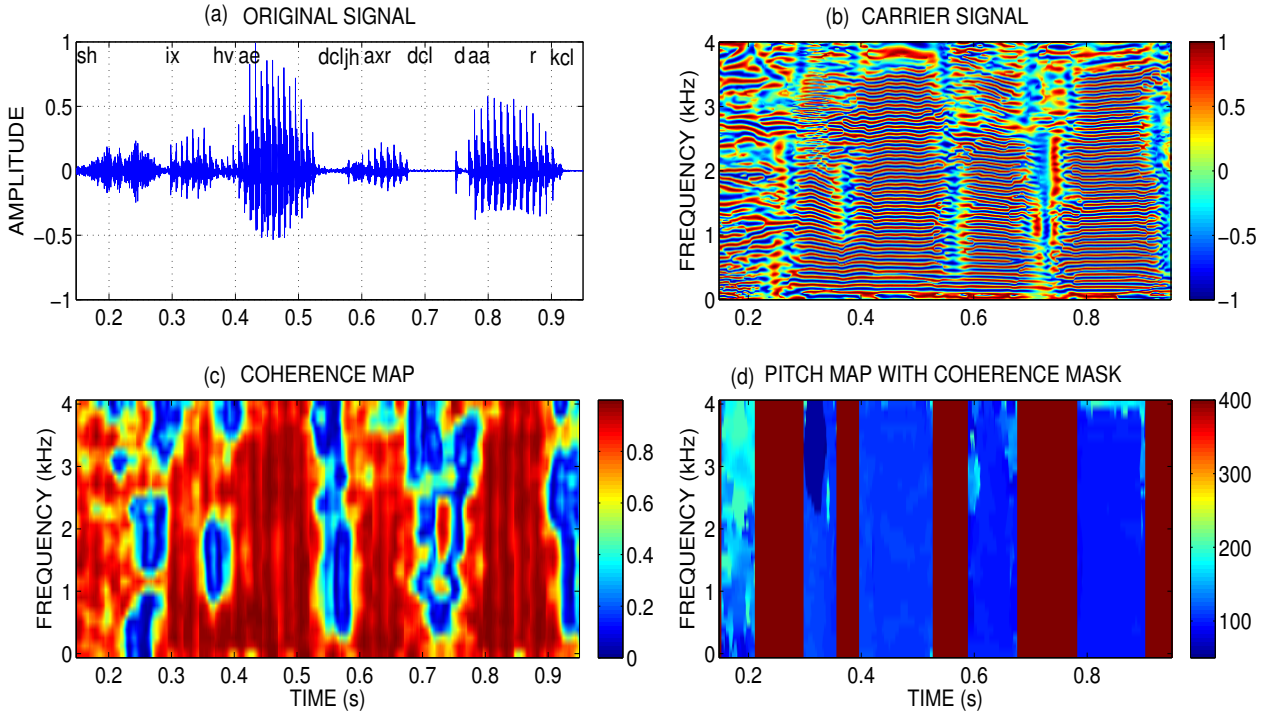

Figure 5.7: Pitch estimation using the coherence map: (a) speech signal along with the transcription; (b) carrier spectrogram obtained after demodulation of speech signal; (c) coherence map computed from carrier spectrogram using structure tensor method; (d) pitch map computed using carrier spectrogram and coherence map

20 kHz. For evaluations, the wave files were down-sampled to 8 kHz. The results obtained for the proposed method is shown in Table 5.2. For evaluation, two objective measures were used namely gross error (GE) rate and the mean error (ME). GE is defined as the percentage of voiced frames with an estimated pitch value which deviates from the reference value by more than 20%. ME is defined as the absolute difference between the mean of reference pitch and the estimated pitch. The estimated pitch is compared with the zero-frequency filter (ZFF) based pitch estimation method. Details of the pitch estimation from the ZFF is given in [71], same procedure is followed in this chapter. The final results obtained are shown in Table 5.2. Figure 5.8 shows pitch trajectory across frames for ZFF and using carrier spectrogram.

5.5 Synthesis methodology

The proposed framework provides a unique decomposition (AM and FM) of NB spectrogram. AM and carrier spectrogram can be combined with STFT phase to synthesize high-quality speech having

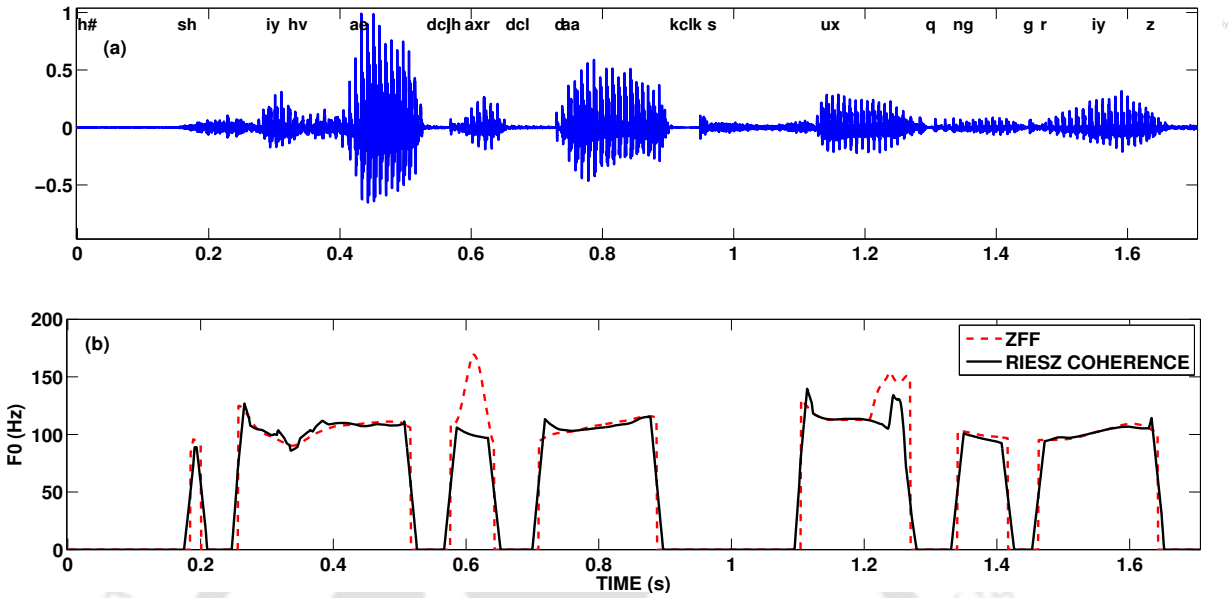

Figure 5.8: F0 estimation from Riesz transform: (a) Speech utterance from TIMIT with transcription; (b) comparisons of F0 estimation from ZFF and Riesz coherence methods shown in red and black color, respectively.

no perceptual difference with the original speech. Since it is difficult to model and parameterize STFT phase, a different approach is taken for generating an excitation signal. Based on the source-filter theory of speech production mechanism it is possible to design an appropriate excitation signal, which can be used to excite vocal tract filter for speech synthesis. Assuming source-filter model for speech synthesis, the proposed framework provides smooth AM envelope (Vocal tract response), pitch information for voiced speech frames and aperiodicity for noise modeling. Along with these informations, we test the suitability of the proposed framework by designing three types of excitation signals 1) impulse excitation utilizing only f_0, 2) impulse excitation with addition of random phase in high-frequency regions, and 3) excitation signal same as in case (2) with the incorporation of aperiodicity information for noise modeling. We describe the details in the following sections.

5.5.1 Synthesis using STFT Phase

Let $V(m)$ and $\cos\Phi(\omega)$ represents the smoothed AM envelope and the FM carrier spectrogram, respectively. The reconstructed spectrogram ($\tilde{S}(\ell, m)$) generated from the modulation AM and FM patches ($\tilde{S}_W^{i.j}(\ell, m)$) using overlap-add method in the least-squares sense [150] and is given by

$$\tilde{S}(\ell, m) = \frac{\sum\limits_{i,j} \tilde{S}_W^{i.j}(\ell, m) W(Tj - \ell, Fi - m)}{\sum\limits_{i,j} W^2(Tj - \ell, Fi - m)} \tag{5.8}$$

where T and F denote the hop size of the 2-D window along the time and frequency axes, respectively, $\boldsymbol{m} = (\ell, m)$ denotes the discrete time and frequency variables, respectively. i, j denotes patch location in the spectrogram. $S(\boldsymbol{m})$ is combined with the phase of the original STFT and inverted using overlap-add reconstruction with least-squares cost function to obtain an estimate of the speech signal ($\tilde{s}(n)$) given by

$$\tilde{s}(n) = \frac{\sum_{\ell} \tilde{s}_w(n, \ell) w(T\ell - n)}{\sum_{\ell} w^2(T\ell - n)} \tag{5.9}$$

where $\tilde{s}_w(n, \ell)$ is the inverse Fourier transform of the ℓth frame of STFT with magnitude $\tilde{S}(\ell, m)$ and the phase corresponding to the STFT of the original signal. The synthesis quality of this speech is natural and sounding similar to original speech.

5.5.2 Synthesis using F0

Synthesis using STFT phase with Riesz AM envelope gives a speech signal with near to natural quality. However, parametrization of phase is difficult due to the randomness involved in the phase signal. Hence, different approaches have been tried in the literature to synthesize speech. One of the basic methods is using a simple impulse/noise based excitation. This method requires only one F0 parameter and it is easier to model. Hence, in this work, we used F0 computed from the carrier spectrogram for generating an excitation signal. Based on the F0 values, impulse with unit amplitude is inserted frame wise for every voiced frame with a interval of F0. Whenever an unvoiced frame is present white Gaussian random noise is generated. The voicing decision for each frame is computed from the coherence map. To evaluate the synthesized file from the simple excitation method convolving with Riesz AM envelope, perceptual evaluation of speech quality (PESQ) score is computed for around 4 languages (Hindi, Tamil, Assamese, and Manipuri). In each language, 10 sentences are used for evaluation. The average PESQ score for synthesis from Riesz AM envelope with impulse/noise excitation can be seen in Table 5.3 with a score of 2.237, which is better than the STRAIGHT method when the same excitation is used with STRAIGHT envelope. The synthesis quality from the F0 excitation is intelligible. Perceptually, the quality is degraded due to the fixed F0 interval in both Riesz and STRAIGHT method, which introduces buzziness to synthetic speech.

5.5.3 Synthesis using Random phase

Based on the results shown in the previous section, it is clear that phase response plays an important role in perception. Further, Kawahara *et al.* showed in the STRAIGHT method that for the voiced speech, adding group delay of fixed-delay in the higher frequency improves the perceptual quality [18]. Hence, in this work also, we added group-delay based random phase in the higher frequency band. A random-phase signal can be generated by sampling the phase from the uniform distribution as follows

$$\Phi_{rand}(\omega) = \mathcal{U}(-\pi, +\pi) \tag{5.10}$$

and then multiplied sigmoid function to add the random phase-only in the higher frequencies and is given by

$$\Phi_{rand_w}(\omega) = \Phi_{rand}(\omega) * \sigma(\omega) \tag{5.11}$$

$$\sigma(\omega) = \frac{1}{1 + e^{-x}} \tag{5.12}$$

However, since the phase spectrum must be odd-symmetric around the Nyquist frequency, the sampling is performed for the positive frequency part and then the odd mirror image is appended to the phase spectrum. After the random phase addition to Riesz envelope along with impulse/noise excitation, the time-domain signal is constructed by the inverse discrete Fourier transform. The random phase addition helped in improving the PESQ score to 2.372. This signifies the importance of phase response in speech perception.

5.5.4 Synthesis using Aperiodicity

The voiced speech is usually assumed as produced from periodic excitation over a short segment of speech. However, even within a short segment, variations in the strength of excitation (shimmer) and variations in periodicity (jitter) will be present. Due to this jitter and shimmer, source signal will not be perfectly periodic. Hence, we need to incorporate this aperiodicity in generating the excitation signal. The simple F0 based excitation is not able to accommodate this aperiodic component present in the voiced speech. Hence, two types of excitation is generated within voiced speech segment to represent the periodic and aperiodic component of voiced speech, respectively. The periodic and aperiodic decomposition, which is computed from the coherence map give an indication of the amount of harmonic component present in the different frequency bands. The overall synthesis block diagram

Table 5.3: PESQ score for Riesz and STRAIGHT spectral envelope with the different types of excitation

Excitation signal	Riesz	STRAIGHT
STFT Phase	3.49	3.364
Impulse excitation	2.237	2.199
Impulse + Random phase	2.372	2.361
Impulse+Random phase+ aperiodic component	2.465	2.563

is shown in Figure 5.9, where to impulse/noise excitation aperiodic component and random phase component is also added. The synthesis equation using aperiodicity spectrum ($A(\omega)$) is given as follows:

$$S(\omega) = E(\omega)V(\omega) \tag{5.13}$$

where $V(\omega)$ smoothed vocal tract envelope computed from Riesz transform and $E(\omega)$ is the excitation spectrum given by:

$$E(\omega) = [1 - A(\omega)]I(\omega)\Phi_{rand_w}(\omega) + A(\omega)W(\omega) \tag{5.14}$$

Where $A(\omega)$ is the aperiodic spectra computed from Riesz transform, $I(\omega)$ is the spectra computed from impulse excitation, $\Phi_{rand_w}(\omega)$ is the random phase spectra computed from the white Gaussian noise weighted by sigmoid function, and $W(\omega)$ is the white Gaussian noise spectra.

The overall quality from the periodic/aperiodic excitation with random phase is perceptually better than impulse/noise + random based excitation. However, STRAIGHT is giving slightly better perceptual quality with PESQ score of 2.563, which has to be further investigated.

5.6 Experimental validation

In order evaluate proposed analysis/synthesis framework using Riesz transform, speech synthesis systems are developed in different languages and tested in statistical framework. Further, we compare the performance of the proposed approach with STRAIGHT method. First, the comparison in the analysis/synthesis framework is done. Then the usefulness of this framework for SPSS using HMM is shown.

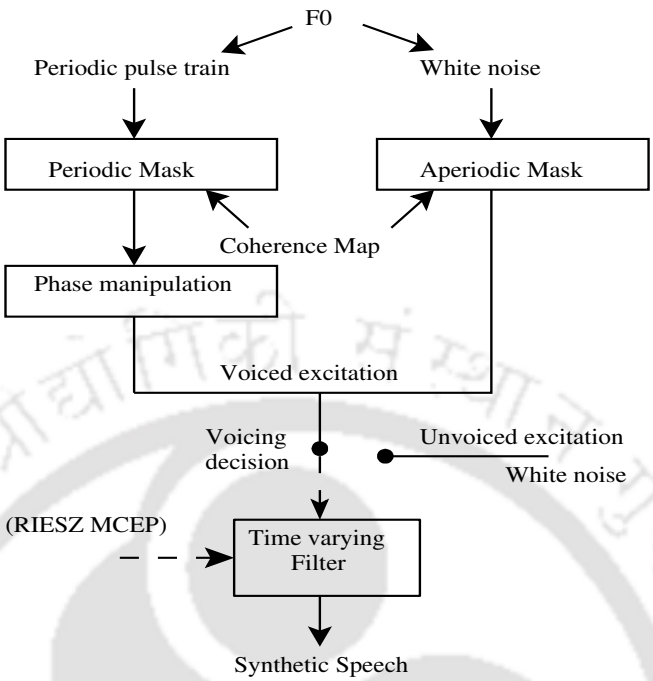

Figure 5.9: Synthesis framework of proposed method

5.6.1 Analysis and Synthesis framework

The features like smoothed envelope, F0, coherence map, and voicing decisions are computed from Riesz transform. The smoothed envelope and coherence map are compactly represented using MCEP feature. All these features are evaluated in analysis/synthesis framework. The overall synthesis framework is shown in Figure 5.9. The phase manipulation mentioned in the block diagram is done by addition of random phase with a fixed group delay similar to the procedure used in the STRAIGHT method [18]. For aperiodicity model, MCEP parameter of the order 25 computed from the coherence map is used. During synthesis, these MCEP features are converted back coherence spectrum. F0 decision is based on proposed carrier spectrogram with silence correction using short-time energy. The vocal tract spectrum is computed from parametrized MCEP parameters with an order of 35. For evaluation, four languages, namely, Indian English, Hindi, Assamese, and Manipuri are taken from Indian TTS database [146]. Each database has one male and one female speaker spoken in native accent. From each speaker 10 sentences were taken for evaluation.

Table 5.4: PESQ and SNR scores of Riesz and STRAIGHT methods for Indian TTS database

Objective Measure	Riesz		STRAIGHT	
	Male	Female	Male	Female
PESQ	3.78	3.40	3.47	3.57
SNR	16.23	14.76	13.23	14.41

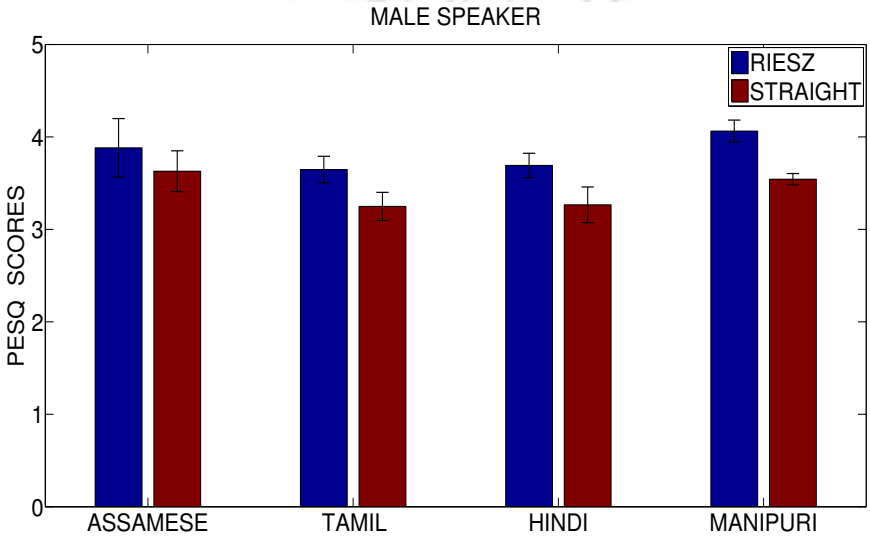

Figure 5.10: PESQ score of the male speakers for the proposed method in analysis/synthesis framework evaluated for 4 Indian languages

5.6.1.1 Objective evaluation

In this chapter, two objective measures are used, namely, PESQ and signal-to-noise ratio (SNR) [106]. The PESQ measure should be interpreted as an MOS regarding the similarity to the original waveform. The PESQ scores obtained for the proposed method and its comparison with STRAIGHT framework is shown in Table 5.4. It can be observed from the table that the proposed method is having a relatively higher PESQ score of 3.78 when compared with STRAIGHT method which is having a score of 3.47. This signifies an improvement in the naturalness of synthetic speech by the proposed method. Further, for female speaker, both methods are having a comparable score with the slightly higher score for STRAIGHT method.

In addition, to know the distribution of PESQ scores for male and female speakers, the bar chart is plotted in the Figure 5.10 and 5.11, respectively. It shows PESQ score with a 95% confidence interval. From the bar plot, it can be seen that the proposed Riesz method for male speakers is

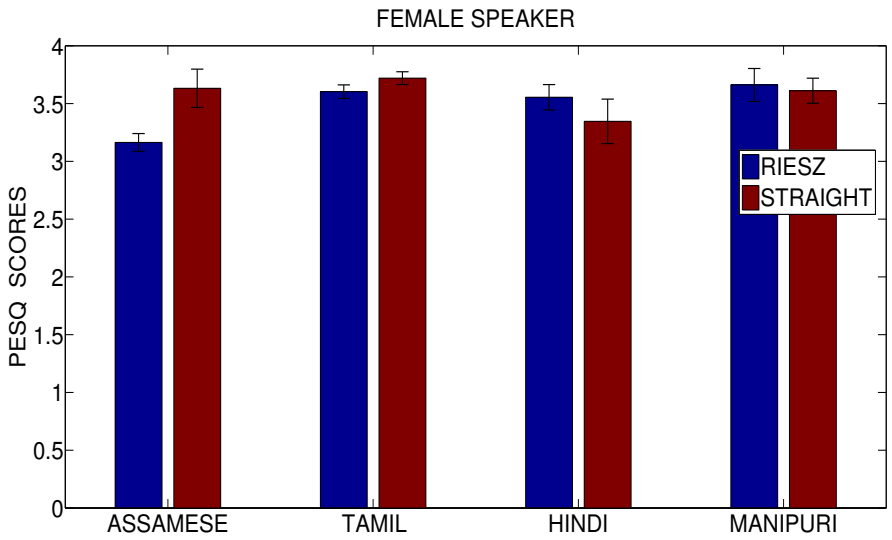

Figure 5.11: PESQ score of the female speakers for the proposed method in analysis/synthesis framework evaluated for 4 Indian languages

slightly better than the STRAIGHT method. For the female speaker, the proposed and STRAIGHT method performs almost similar.

Second objective evaluation is the SNR measure, which gives distortion error in the temporal domain. This measure is evaluated between reference original speech and synthesized speech for the same text. The average SNR for proposed method and STRAIGHT method given in Table 5.4. The SNR of proposed method is higher indicating the better spectral modeling of proposed method compared to the STRAIGHT algorithm.

5.6.1.2 Subjective evaluation

In this evaluation, mean opinion score (MOS) is conducted to know the perceptual quality of analysis/synthesis framework [151]. In MOS test, 10 sentences from each language are used and given to subjects along with the original waveform. The subjects were asked to give the score with the scale of 1 to 5 by comparing with the original wave file. For evaluations, listeners were asked to examine naturalness and perceptual distortions present in each file and to give their overall scores accordingly. The average scores obtained from subjects were given in Table 5.5. From the table, it can be seen that the proposed Riesz method outperform STRAIGHT algorithm in the analysis/synthesis framework.

In the preference test (PT), subjects have to select between the two versions of the system at a time, by listening to the different sentences from each system. Further, they can select both the

Table 5.5: MOS and PT results for analysis/synthesis framework with 95% confidence interval

Experimental Evaluation	Analysis synthesis framework			p-value
	STRAIGHT	Riesz	Same	
MOS	4.15	4.28	-	
PT	34%	48%	18%	< 0.01

system as same as their choice. The percentage of preference scores with p-value from listeners can be viewed in Table 5.5. It can be seen that subjects preferred the proposed technique over straight method with statistically significant p-values given by hypothesis tests. This indicates the importance of proposed analysis/synthesis framework for improving the naturalness of synthetic speech.

5.6.2 Statistical parametric speech synthesis

The integration of proposed vocoder framework for SPSS using HMM is provided in Figure 5.12. It provides a unified framework to model AM envelope, FM carrier, and coherence map simultaneously in different streams using HMM [5]. The framework for speech synthesis is mainly classified into training and synthesis. In training, AM envelope representing vocal tract envelope, FM carrier representing fundamental frequency, and coherence map representing periodicity are derived from each phoneme for the whole database. For evaluation of the systems, HMM based speech synthesis is developed for two speakers (one male and one female) taken from CMU ARCTIC database [22]. Each speaker consist of 1132 sentences, for training 1000 sentences are tested and remaining 132 sentences used for testing. For comparison, HMM based synthesis system using STRAIGHT features are also developed and procedure used is similar to the steps mentioned in [15].

All the phonemes are modeled with 5 states and in each state 5 streams are used to model the different parameters extracted for each phoneme [59]. The first stream consists of AM envelope parameters i.e 35 MCEP features including the zero*th* coefficient and their delta and delta-delta coefficients. It is trained by continuous density HMMs. The carrier parameter F0 and its delta and delta-delta coefficients are trained in a three separate streams with Mult-space distribution (MSD) model [58]. In the fifth stream, periodicity parameters obtained from coherence map is modeled with 25 MCEP parameters to represent periodic component present in each band. The details of each stream and its distribution are as follows:

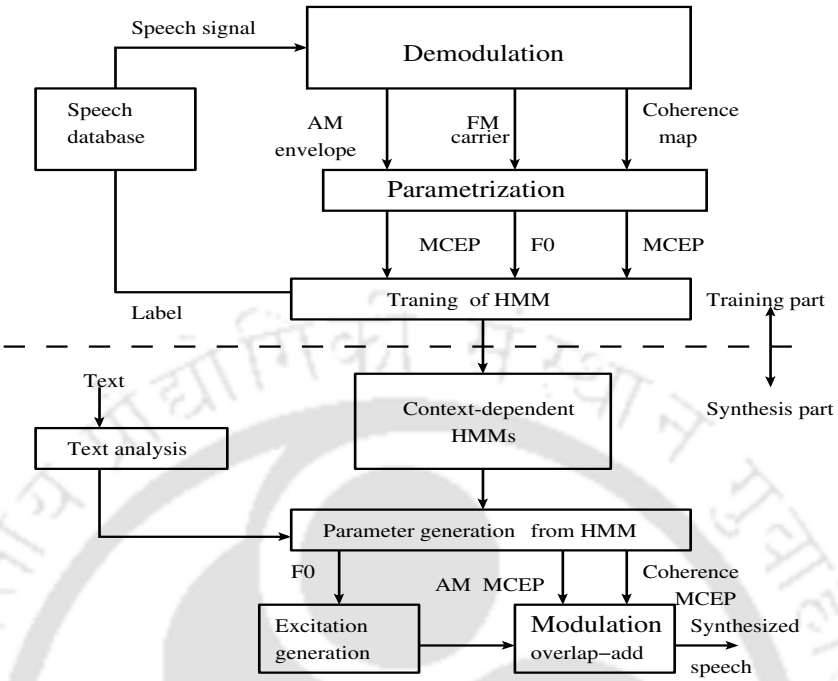

Figure 5.12: Block diagram of HMM based speech synthesis

- The first stream for MCEP (order: 35) and its derivatives with a continuous probability distribution to represent the smoothed AM envelope.
- In the next three streams the fundamental frequency, its delta, and delta-delta are modeled with MSD.
- The fifth stream MCEP (order: 25) parameters are computed from the coherence map and its derivatives, which are modeled with a continuous probability distribution.

For each phoneme, parameters mentioned above are extracted along with their corresponding labels. In training part, each parameter is computed using Baum-Welch re-estimation algorithm [5]. During synthesis, as per input text, frame wise AM MCEP, coherence MCEP, and F0 parameters are computed by maximizing output probability using the maximum likelihood parameter generation algorithm [5]. The generated coherence map MCEP parameters are used for obtaining the amount of aperiodic component present in each band of STFT spectra. Some of the synthesized samples from the Riesz method can be accessed from the following link[1]

[1] http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/rieszhts.php

Table 5.6: Comparison of Objective results using STRAIGHT and proposed analysis/synthesis framework for HMM based speech synthesis

Objective Measure	**MCD (dB)**	**BAP (dB)**	F0 **RMSE Hz**	**V/UV %**
STRAIGHT	4.36	1.40	9.31	11.66
Riesz	4.02	2.01	8.68	11.18

5.6.2.1 Objective evaluation

The objective results of the two systems using the Riesz and STRAIGHT framework are presented in Table 5.6. Here, 132 sentences are used for testing. The different parameters used for evaluation are the Mel-cepstral distance (MCD), error in band aperiodicity (BAP), F0 root mean square error (RMSE), and voiced/unvoiced (V/UV) error. It is observed that Riesz method performs better than STRAIGHT method in terms of lesser MCD, F0 RMSE, and voicing error. In the case of aperiodicity parameter, STRAIGHT method gives lesser BAP spectrum error. In general, objective results confirms that Riesz method can achieve a better result than STRAIGHT method.

5.6.2.2 Subjective evaluation

To evaluate the perceptual quality of the proposed technique synthesized from SPSS, MOS and PT tests are conducted. In MOS test, 20 sentences were used and asked the subjects to give their opinion score on the scale of 1 to 5 (1: poor and 5: excellent) by comparing with the original file. A total of 10 subjects were used in the listening test. For evaluations, listeners were asked to examine the naturalness of each file compared to the original file and give their scores. The scores obtained from the subjects are given in Table 5.7. From this table, it can be seen that the proposed Riesz transform based analysis/synthesis framework performs slightly better than the STRAIGHT method in the SPSS framework. In the preference test, 37% waveforms felt as same indicating the similarity in both Riesz method and STRAIGHT method. However, around 38% files are given as preference to Riesz method, which is better than STRAIGHT method.

Table 5.7: MOS and PT results for SPSS with 95% confidence interval

Experimental Evaluation	Analysis synthesis framework			p-value
	STRAIGHT	Riesz	Same	
MOS	3.21	3.3	-	
PT	28%	35%	37%	< 0.01

5.7 Summary

In this chapter, 2-D spectro-temporal analysis/synthesis method is proposed using Riesz transform. The demodulation of 2-D spectro-temporal patches using Riesz transform yields smoothed spectral envelope, carrier spectrogram, and coherence map, representing vocal tract spectrum, source signal, and periodicity, respectively, using a single framework. The analysis/synthesis representation gives a better synthesis quality than the state-of-the-art STRAIGHT vocoder. Further, smoothed spectral envelope and periodic component computed from the carrier spectrogram compactly represented using MCEP features and used in the analysis/synthesis framework. The synthesized speech from this framework is compared with state-of-the-art STRAIGHT system, which is based on the pitch-synchronous analysis. The quality of synthesized files are measured using objective and subjective evaluation. The results show that proposed method performed better than STRAIGHT system. The effectiveness of Riesz transform is further studied in statistical framework using hidden Markov model based speech synthesis and results show that decomposition of the spectrum into an envelope, carrier, a periodic map is equally effective like STRAIGHT method.

In the next chapter, source modeling for the statistical framework is proposed using integrated linear prediction residual. In particular, next chapter explores the different components present in glottal activity region like aperiodic and phase information to improve the naturalness of synthetic speech. Initially, these components are studied in analysis/synthesis framework and then they are tested in the statistical framework using HMM.

6

Integrated Linear Prediction Residual for Source Modeling

Contents

6.1 Different components of source modeling 112
6.2 Different components present in Glottal activity region 114
6.3 Periodic and Aperiodic modeling 120
6.4 Phase modeling 128
6.5 Summary 136

Objective

The objective of this chapter is to demonstrate the significance of different components present in the glottal activity region for source modeling. The different components include a sequence of epoch with varying excitation strength, a small segment of linear prediction (LP) residual around each epoch, Hilbert envelope of LP residual, and the phase components present in the glottal activity region. The strength weighted epoch sequence generates speech which is intelligible, but synthetic in nature. By considering a small region of samples of LP residual around the epochs, the naturalness of synthesized speech increases significantly. However, modeling samples around the epochs in the statistical framework is difficult. Hence, the signal is divided into periodic and aperiodic representation from the integrated LP residual (ILPR). The ILPR signal is modeled in the frequency domain by dividing the spectrum into two bands to characterize periodic and aperiodic components of the glottal activity region. The periodic components of ILPR signal below the maximum voiced frequency (f_m) are modeled using Mel-cepstral coefficients called as RMCEP. The aperiodic component above f_m is modeled using white Gaussian noise shaped by pitch adaptive triangular envelope and weighted by the strength of excitation (SoE). The RMCEP and SoE are trained in HMM framework along with MCEP and F0 parameters. The synthesized speech by the proposed source modeling reduces the buzziness and improves the speaker similarity. To improve the naturalness, phase component is also modeled from ILPR signal. The phase characteristics of excitation signal are estimated from the cosine phase of ILPR using an all-pass filter. The all-pass filter coefficients (APC) derived from the all-pass filter are used as features for modeling phase in SPSS. During synthesis stage, to generate the excitation signal, frame wise generated APC are used to add the group delay to the impulse excitation. The experimental results show that phase modeling results in a better perceptual synthesis quality.

6.1 Different components of source modeling

The primary motivation of this chapter is to model the glottal flow derivative signal to enhance the naturalness and improve the speaker similarity of the synthesized speech. However, in practice, it is challenging to model the source signal. This may be because the source signal in case of voiced speech *i.e.* glottal activity region consists of a harmonic structure in a low-frequency band and noise component in the high-frequency band [2, 25, 91]. Further, phase component present in the glottal activity region plays an important role improving the perceptual quality of speech.

In the HMM based speech synthesis, there are several attempts have been done to model the residual signal, which is an approximated source signal. The residual signal consist of both aperiodic and phase components. In [23,84,152], the residual signal is modeled as harmonic component and noise component, representing the deterministic and stochastic (DSM) part of the source signal, respectively. The spectrum of the residual signal is divided into two bands separated by the maximum voiced frequency (f_m). The lower band below f_m is modeled by processing pitch-synchronous residual frames and keeping it as codebooks. The stochastic component above f_m is modeled by random noise with its shape is weighted by pitch adaptive triangular window. In GlottHMM, glottal pulses are extracted from speech via iterative adaptive inverse filtering and stored as a library of pulses, resulting in improved synthesis quality [32]. However, storing codebook or glottal pulses need separate memory and a complex optimized algorithm is required to select the codebook or glottal pulses for creating excitation [86].

In recent years, some works are done in modeling the phase component for speech processing applications [153]. These works investigated the usefulness of phase component for speech recognition and speaker verification task [124, 154]. Particularly, in speech coding area, to get the sophisticated excitation signal, phase component is used along with minimum-phase synthesis filter to get the mixed phase characteristics of the speech signal. However, in speech synthesis, more specifically in SPSS, not much exploration is done in modeling the phase component. This is mainly due to the random nature of phase signal and it is not suitable for training directly in HMM. Yet, there are some works showed that phase spectrum also has a significant role in the improvement of naturalness of synthetic speech. Kawahara *et al.* in [65] showed that adding fixed group delay around the high-frequency region leads to improvement in the naturalness of synthesized speech. In [84,88,93,155], even for SPSS, it is shown that phase can be modeled and improvements in the quality of synthetic speech are reported. Further, in recent advances, phase component is used in deep-learning based speech synthesis [50]. In summary, to generate the excitation signal similar to glottal flow or residual signal, we have to move beyond the impulse excitation and minimum-phase assumption and need to model both phase component and aperiodic component.

In this chapter, the parametric representation of the residual signal is done to get better synthesizer for HMM based speech synthesis. Initially, in the analysis/synthesis framework, significance of deriving the different source component using LP residual signal with epoch as anchoring point is analyzed.

The different source components include strength weighted epochs, aperiodic representation, and phase information. Next, methods to parametrize these different components in statistical framework are shown. The rest of the chapter is organized as follows: different components present in the glottal activity region for source modeling and their significance to speech synthesis are shown in Section 6.2. The aperiodic and phase component extraction from ILPR signal and modeling these components in statistical framework are described in Section 6.3 and 6.4, respectively. The chapter is finally concluded in Section 6.5.

6.2 Different components present in Glottal activity region

In this section, different components present in the glottal activity region, which represent the source signal is described using epoch knowledge. Epoch can be extracted by different methods like group delay, DYPSA, ILPR, and zero-frequency filter [110, 111, 116, 122]. Even though all these methods are popular for epoch extraction, the zero-frequency filter method is shown to give the best performance compared to other methods [110]. This chapter, therefore, uses zero-frequency filter method for epoch extraction.

6.2.1 Epoch based excitation model

In this section, different methods to derive the excitation using the epoch knowledge are shown. The detailed procedures to extract epoch from speech signal are mentioned in the earlier chapter 4.

Epoch with strength of excitation:

In the earlier chapter 4, it is shown that epoch based excitation in the glottal activity region gives better synthesis quality than the impulse based excitation. In this method, the epochs with their strength calculated by zero-frequency filter method is used as excitation. Epochs are located for the given speech signal and then their strength of excitation around epoch locations is also computed. Epochs with the variable strength of excitation represent the shimmer of the excitation signal. In addition, epoch locations are preserved in the analysis framework, which represent the jitter of the excitation signal. These jitter and shimmer characteristics of the excitation signal gives the aperiodic component present in speech signal. Figure 6.1(a) shows a segment of speech containing portions of voiced and unvoiced region. Figure 6.1(b) shows the corresponding residual and Figure 6.1(c) shows the epochs extracted by the zero-frequency filter method. The epochs with its strength form the

excitation signal is shown in Figure 6.1(d). This signal is non-zero only at the epochs location.

Residual samples around epoch:

In this method, the epochs are used as anchor points, and a small range of residual samples are extracted from them to form instants of significant excitation. In the present chapter, 1 ms of residual samples on either side of the detected epochs are used to form the excitation [156, 157]. Figure 6.1(e) shows the residual extracted by considering 1 ms range on either side of the epochs. The number of non-zero values is more and this excitation signal can be represented in few parameter. Further, residual samples around the epoch locations represent the aperiodic and phase components of the source signal. As a result, the excitation signal, in this case, may provide significantly better naturalness and preserve the prosodic information in the synthesized speech.

Hilbert envelope of LP residual:

To make the residual samples suitable for modeling, the residual is further divided into two parts using the definition of analytic signal. The magnitude of the analytic signal derived from residual is termed as Hilbert envelope and the real part of the time domain phase of the analytic signal is termed as cosine phase. Both these components is independently modeled for deriving the excitation signal.

The details of deriving the Hilbert envelope and cosine phase from the analytic signal is mentioned in Chapter 4. In this method, the epochs are used as anchor points to select samples from Hilbert envelope of LP residual. A small range of Hilbert envelope samples of 1 ms around epochs are extracted from them to form excitation signal. Figure 6.1(f) shows the Hilbert envelope extracted by considering 1 ms range on either side of epochs. This excitation signal represents the magnitude representation of the residual signal and unipolar in nature.

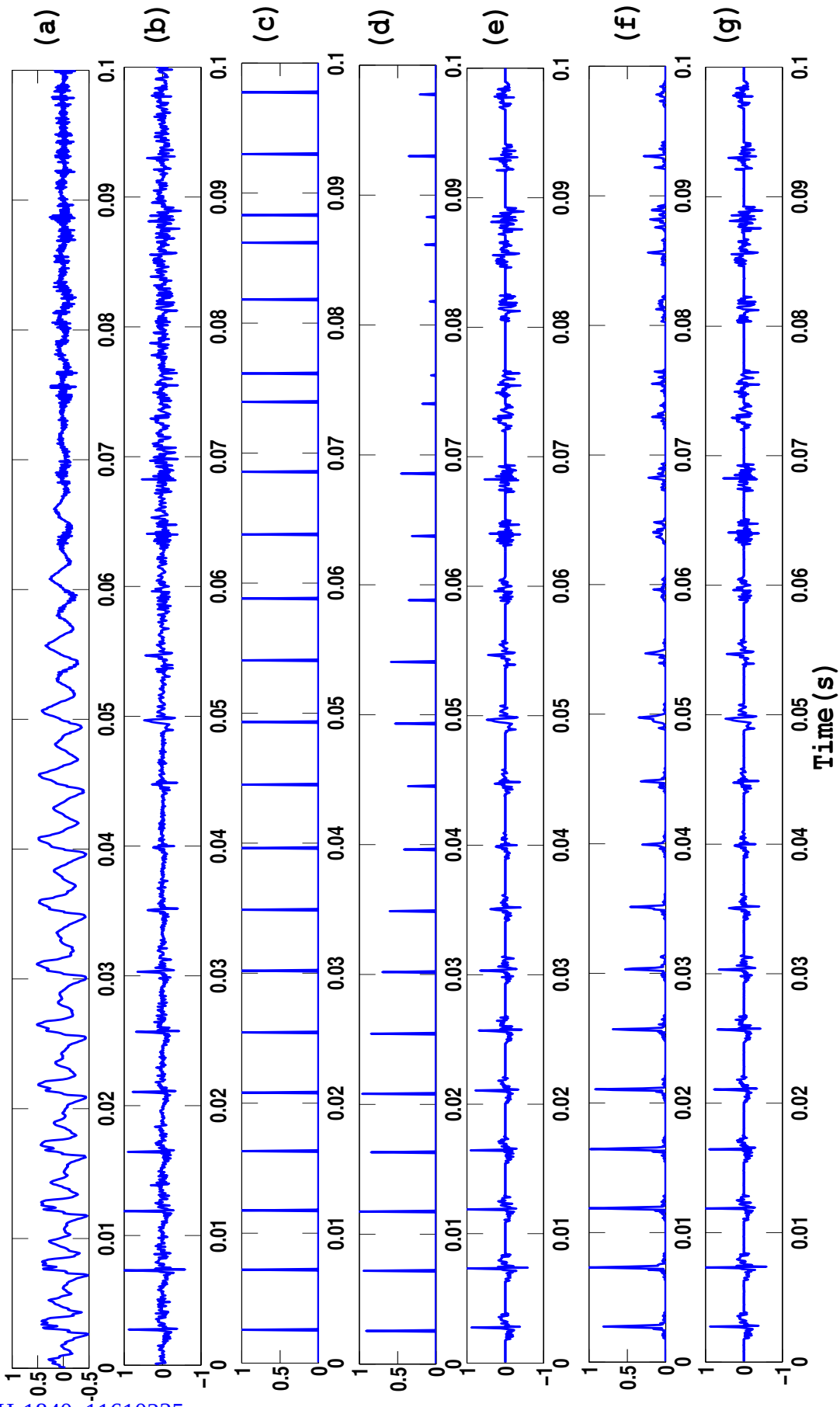

Figure 6.1: Epoch based excitation signal derived from different methods: (a) a speech segment consist of voiced and unvoiced speech portions, (b) LP residual, (c) epochs location calculated from zero-frequency filtered signal, (d) epochs with their strength, ((e)-(g)) excitation signals derived around epochs by considering small portion of LP residual, Hilbert envelope, and Hilbert envelope + cosine phase, respectively.

6.2.2 Experimental studies and discussion

To demonstrate the significance of different epoch based excitation signal, the speech signal sampled at 16 kHz is processed using 20^{th} order LP analysis, 25 ms frame size with a shift of 5 ms. The linear prediction coefficients (LPC) representing the vocal tract information and LP residual representing the excitation source information are obtained. The same speech signal is processed by zero-frequency filter to extract the epochs.

In the first study, epoch sequence with their strength is taken as an excitation signal to represent the jitter and shimmer of the source signal. This signal is given to the LP filter for obtaining the synthesized speech by overlap and add method [158,159]. But in this case, voiced/unvoiced separation is not happening and synthesized speech is intelligible which conveys message information, but less intelligible. In the second study, in order to incorporate naturalness and prosodic information, LP residual is considered. The excitation source signal now contains the sequence of residual samples anchored around epochs as discussed earlier. The informal listening of synthesized speech gave a feel that it is both natural and intelligible, near to that of the original speech. This indicates that the excitation signal should include a small range of samples around epochs for increasing naturalness and preserving prosodic information.

The motivation for the next two studies is to further understand whether strength or sequence associated with the LP residual is more important for preserving naturalness and intelligibility. The residual is decomposed into Hilbert envelope and cosine phase representing the magnitude and phase components, respectively. The Hilbert envelope values around the epochs are preserved as in the case of residual and used as excitation source signal. The synthesized speech is natural and preserves prosodic information. However, the quality seems to be slightly inferior compared to the residual case. This infers that apart from the strength, phase information is also important. The synthesis of speech using only cosine phase resulted in a lot of perceptual noise due to the large amplitudes associated with phase sequence values. For this, the excitation signal using Hilbert envelope are multiplied with respective cosine phase values and used as excitation source signal. The synthesized speech quality improves significantly compared to that of using only Hilbert envelope. This study indicates that we need to preserve both aperiodic and phase component in the excitation signal to get natural speech.

Figure 6.2 and Figure 6.3 shows the speech waveforms and spectrograms of synthesized speech for the word *I was* by using different excitation signals. It can be seen that there are no discontinuities

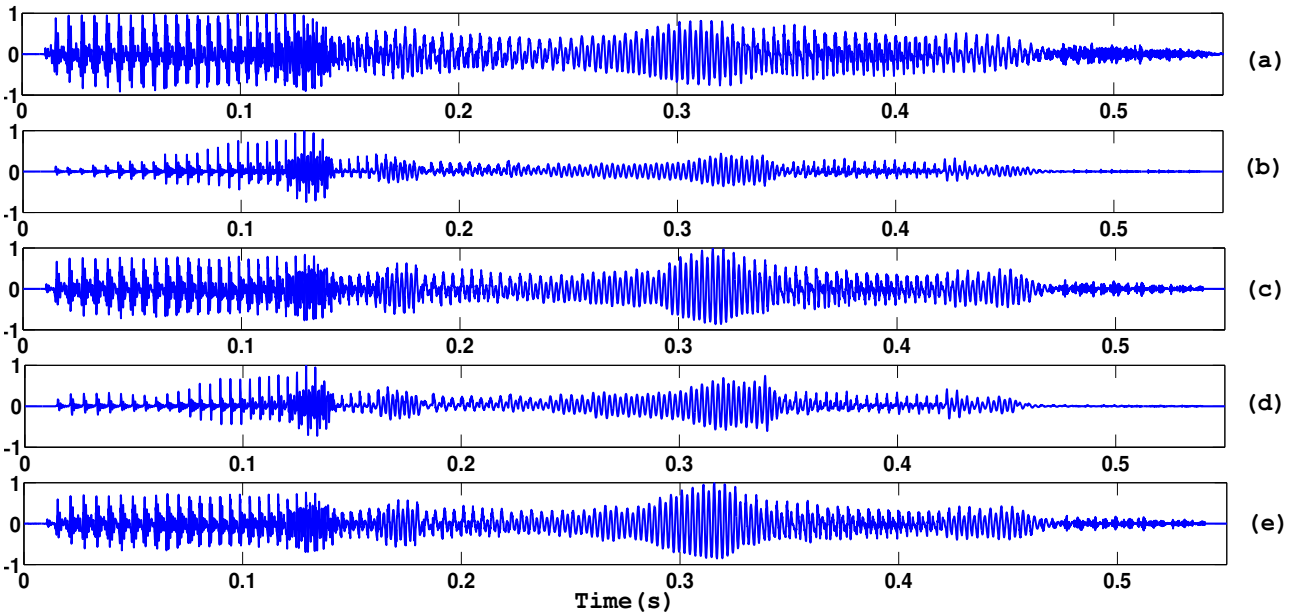

Figure 6.2: Time domain waveforms of synthesized speech for word *I was*: (a) original speech signal, synthesized speech based on ((b)-(e)) strength weighted epochs, LP residual, Hilbert envelope, and Hilbert envelope + cosine phase, respectively.

in the synthesized speech in the case of LP residual and Hilbert envelope+cosine phase based system. Most of the features such as pitch changes and formant transitions seem to be preserved well. Hence, the synthesized speech quality is comparable with the original speech. But in the synthesized speech using Hilbert envelope and strength weighted epoch method, some discontinuities are seen both in the spectrogram and time domain waveforms. Therefore, synthesized speech is not natural, which infers that source of excitation consist of some additional excitation information along with the epoch information.

Subjective Study

One sentence *I was about to do this when cooler judgment prevailed* from Arctic database is selected and recorded from 5 speakers (2 male and 3 female) for the study. The recording initially sampled at 48 kHz is down sampled to 16 kHz and used for synthesizing in four different epoch based excitation schemes as explained above. 15 research scholars of our lab participated in the subjective evaluation. The synthesized speech files using all four methods along with the original speech files are presented to the subjects for the evaluations. The speech files were randomized and file names were coded before

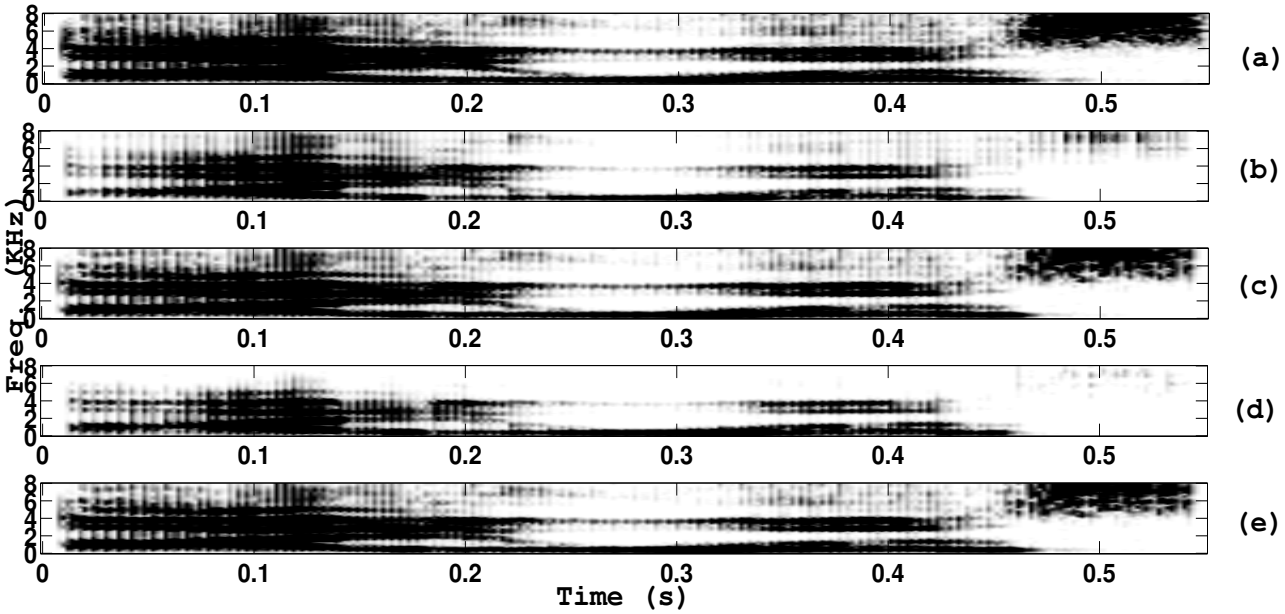

Figure 6.3: Spectrogram view of synthesized speech for word *I was*: (a) original speech signal, synthesized speech based on ((b)-(e)) strength weighted epochs, LP residual, Hilbert envelope, and Hilbert envelope + cosine phase, respectively.

presenting to the subjects for evaluation. A pilot test was given to each subject before the evaluation. The subjects were asked to observe the naturalness, intelligibility and perceptual distortions present in each file and give there opinion scores accordingly on a standard mean opinion score (MOS) test [160]. There are 25 (5*4 + 5 original files) files used for the evaluation. The mean of the scores obtained for all the files for a given excitation signal technique is calculated as the MOS. The MOS obtained for all the 4 techniques are given in Table 6.1. We can observe from Table 6.1 that there is a significant improvement in MOS scores for the LP residual and Hilbert envelope+cosine phase based source models as compared to other two methods. The synthesized files can be accessed from the following link: [1]

From the basic studies, it can be concluded that to get better source representation in the glottal activity region, aperiodic and phase component modeling is necessary. The samples around the residual samples can be preserved, but it is not suitable for statistical modeling. Hence, we need some parametric representation of these components for statistical modeling.

[1] http://www.iitg.ernet.in/cseweb/tts/Assamese/sourcemodeling.php

Table 6.1: MOS for different source modeling using epoch based excitation

Modeling technique	MOS
Strength weighted epochs	2.42
LP residual	4.05
Hilbert envelope	2.87
Hilbert envelope+ cosine phase	4.01

6.3 Periodic and Aperiodic modeling

This section describes modeling of the periodic and aperiodic component using the ILPR signal. As mentioned earlier, in this work, ILPR signal is used as it is a smoothed version of residual signal due to integration operation when compared to LP residual and also signal looks similar to glottal flow derivative signal. In the voiced speech, the frequency below the f_m contains harmonic components, whereas frequency above f_m contains the random noise spectrum [84]. Motivated by this, ILPR signal is divided into two bands, harmonic and noise components in the frequency domain based on voicing frequency (f_m). In the rest of the section, the nature of the ILPR signal and modeling of ILPR signal in two band excitation scheme is described.

6.3.1 ILPR signal

The ILPR signal is similar to glottal flow derivative, having both quasi-periodic nature and harmonic structure. Further, it is a smoothed waveform compared to LP residual signal. Figure 6.4(b) shows the ILPR signal obtained after passing a non pre-emphasized speech through the inverse LP filter. The signal looks similar to the differentiated electroglottography (DEGG) signal, which is shown in Figure 6.4(a).

The ILPR signal contains both periodic and aperiodic components in the voiced speech. Figure 6.5(a) shows the ILPR signal for a voiced speech segment having a quasi-periodic waveform with noise component embedded in it. Further, to know these two components, the ILPR signal is passed through a low-pass and a high-pass filter with a cut-off frequency of voicing frequency (f_m= 4 kHz) [84]. The low-pass and high-pass filtered ILPR signal are shown in Figure 6.5(b) and (c), respectively. The low-pass filtered signal retains the periodic nature of voiced signal and turbulence noise is de-emphasized to some extent. The turbulence noise is preserved in the high-pass filtered signal shown in Figure 6.5(c).

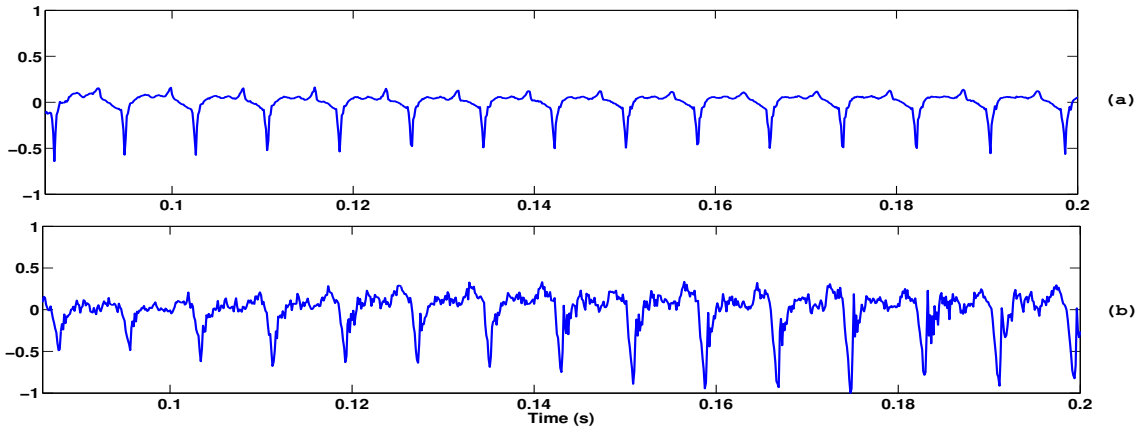

Figure 6.4: Source signal representation of a speech segment of voiced regions: (a) and (b) reference DEGG source signal and ILPR signal for a same speech segment, respectively.

It can be observed that the high-pass filtered ILPR signal consist of noise component synchronized with pitch period of speech and having a variable amplitude due to excitation around the epoch region. The excitation samples present around the epochs are perceptually very important for synthesis quality [156]. These events present in the production of voiced speech is due to the turbulence noise are partially produced at the time instants of opening and closing of the vocal folds resulting in the aperiodic component present in the glottal activity region [97, 156].

Hence, in this chapter to make ILPR signal suitable for parametrization and then train in HMM, it is decomposed into two components: harmonic and noise. The ILPR signal $r_i[n]$ is given by

$$r_i[n] = r_h[n] + r_{no}[n] \tag{6.1}$$

where the harmonic component $r_h[n]$ represents the periodicity in voiced speech and noise component $r_{no}[n]$ tries to capture the aperiodicity present in the voiced speech.

6.3.2 Residual MCEP representing harmonic component

To represent the harmonic structure of ILPR signal, the residual signal is divided into two bands in the frequency domain based on voiced frequency f_m. The lower band of the residual signal below f_m is parametrized using MCEP in the frequency domain. The MCEP approximates spectrum of the residual signal in the frequency domain with very small error and it is called as RMCEP in this chapter [161]. It captures the harmonic structure of the source signal. The value of voicing frequency (f_m) is fixed in this work to 4 kHz as mentioned in [96]. Moreover, in this chapter, more focus is given

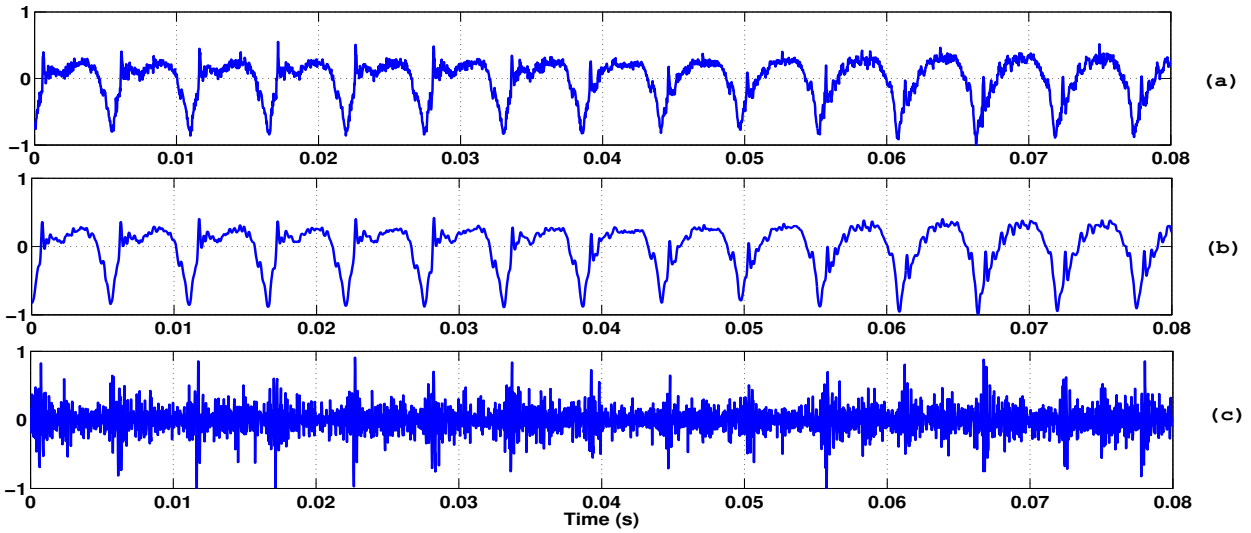

Figure 6.5: Periodic and noise component of the ILPR source signal for voiced regions: (a) the ILPR source signal for a segment voiced speech; (b) and (c) periodic and noise component of the ILPR signal obtained by low-pass filtering and high-pass filtering the ILPR signal, respectively.

on the modeling of harmonic components of ILPR using MCEP rather than showing the effectiveness of variable voiced frequency. The advantage of modeling harmonic components with MCEP instead of conventional MFCC are: MCEP are derived from the magnitude spectrum of the speech and Mel-warping function is applied in the cepstral domain using all-pass filter system instead of conventional filter-bank analysis used in the computation of MFCC. Further, warping of the spectrum using all-pass filter resulting in an invertible set of features between the cepstral and spectral domain. This property makes them particularly attractive in generative models like the SPSS [161].

6.3.3 Noise component

The noise modeling ($r_{no}[n]$) using the ILPR signal is followed similar to the procedure done in the harmonic noise model (HNM) [96]. In the HNM, it is assumed that white Gaussian noise $b[n]$ is convolved with an auto-regressive model $q[n]$. The time domain envelope is modulated by weighting function $w[n]$:

$$r_{no}[n] = w[n](q[n] * b[n]) \tag{6.2}$$

where $w[n]$ is the noise envelope, which is a pitch dependent triangular function, trying to fit the noise component present in the ILPR signal. Since f_m is fixed to 4 kHz in the proposed method and

the spectrum of the ILPR signal is flat over the entire frequency band, the auto-regressive model has assumed to be having the same effect for all the frames. Hence, $q[n]$ is considered as a high-pass filter (beyond f_m) slightly attenuated in the very high frequencies. In this chapter, instead of using the constant envelope amplitude for triangular function, variable amplitude obtained from the strength of excitation of the ILPR signal around the epoch is used as envelope amplitude.

Strength of excitation (SoE)

In the voiced region, due to the rapid movement of vocal fold, significant excitation occurs during the closing of the vocal fold. This results in high strength in the source signal near the epoch location. This can be observed in Figure 6.5(c), showing the high amplitude around the epoch location for the noise component in the high-pass filtered ILPR signal. Moreover, the amplitude of the noise component around the epoch location is variable, so estimating this amplitude may help in representing noise component in a better way. The strength near an epoch can be obtained from the ILPR signal by passing it through the zero-frequency filter and taking the slope of the filtered signal near the epoch location [108]. The strength of excitation ($s_e[k]$) for ILPR signal around epoch region is defined as the slope of the filtered signal derived from the ILPR signal($z_i[n]$) given by:

$$s_e[k] = |(z_i[k+1] - z_i[k])|, \tag{6.3}$$

where k is the epoch location. $s_e[k]$ gives the strength of the impulse-like excitation at the epoch location. The SoE parameter gives a variable amplitude to pitch adaptive triangular noise envelope.

6.3.4 ILPR source modeling for HMM based speech synthesis

The proposed source modeling is tested with publicly available open source toolkit HTS [34]. In the base version of the HTS, each phoneme is modeled with 5 states. In each state 4 streams are used to model the different features of phonemes. The first stream is used for MCEP and its derivatives, representing vocal tract transfer function. The next three streams are used for the fundamental frequency (F0), its delta, and delta-delta, respectively, to represent the source information. Here, F0 is trained as multi-space distribution (MSD), which models, both voiced and unvoiced regions in single model [58]. In this chapter, along with these 4 streams, RMCEP and its derivatives are modeled in the fifth stream to represent the harmonic component of the source signal and in the last stream, SoE and its derivatives are used, which gives the varying amplitude to noise model.

The voice/unvoiced decision to generate excitation is modeled by the weight of MSD, whereas du-

Table 6.2: Speech parameters used per frame for training the HTS

Feature	Number of parameters
Fundamental frequency (F0)	1
Strength of excitation (SoE)	1
Residual mel-cepstral coefficients (RMCEP)	20
Mel-cepstral coefficients (MCEP)	35

ration is modeled by a single Gaussian distribution for each state. The number of speech parameters used in the training of HMM per frame is summarized in Table 6.2. To represent the harmonic component, 20 RMCEP parameters are used in the proposed source model. In addition, one SoE parameter is used to represent the varying amplitude of noise component. During the synthesis, parameters are generated by the maximum likelihood algorithm, as described in [9].

Proposed source modeling using ILPR signal

A block diagram of the proposed source modeling using the ILPR signal is shown in Figure 6.9. The impulse train is generated according to F0. The resulting signal is passed through the low-pass filter and weighted with the residual spectrum generated from RMCEP to represent the harmonic part of excitation. The noise component $r_{no}[n]$ is generated by high-pass filtering of the white Gaussian noise. The resulting signal is modulated by a pitch adaptive triangular envelope weighted by the SoE. Both harmonic and noise components are added in the spectral domain. The added spectrum of the excitation signal is passed through the MLSA filter and then overlap-added to obtain the synthesized speech. In the case of unvoiced regions, only white Gaussian noise is used as the excitation. The voice/unvoiced decision is made based on the MSD weight of fundamental frequency.

6.3.5 Experimental evaluation

To evaluate the proposed vocoder, HTS system is built for two speakers: SLT (US female) and BDL (US male). The speakers SLT and BDL are taken from the CMU ARCTIC database available publicly [22], which consist of 1132 sentences. The first 1000 sentences are used for training and remaining sentences are used for testing. The parameters proposed in the previous sections are extracted for a frame size of 25 ms with a frame rate of 5 ms and trained in the HMM framework. For the com-

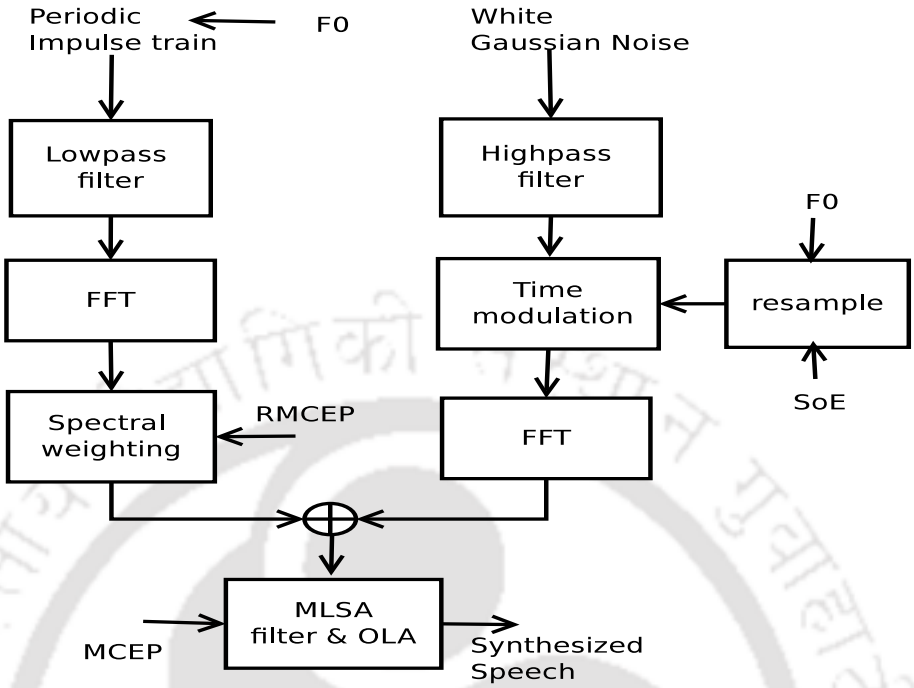

Figure 6.6: The work flow of the source modeling using ILPR signal

parison purpose, along with the proposed method, 3 more systems, based on impulse/noise, mixed, and STRAIGHT excitation source model are developed in the HMM framework. In the impulse/noise source model, impulse and white Gaussian noise are used as excitation, for voiced and unvoiced frame, respectively. The mixed excitation is based on a simple two band excitation for voiced speech with low-pass filtering the impulse train below the voicing frequency (f_m=4 kHz) for the lower band, whereas in the higher band, white Gaussian noise is high-pass filtered above the voicing frequency. In the STRAIGHT based excitation, impulse excitation is convolved with band-pass filter weighted by aperiodic components and a random phase is added generated from fixed group delay [149]. In all the systems, vocal tract system is modeled by MCEP computed on the STFT spectrum of speech. The synthesized files for all the 4 methods can be accessed from the following link [2].

Subjective evaluation

In this evaluation, two tests are conducted, namely, MOS test and preference test (PT). The 25 sentences which are not used in training are given to subjects along with the original waveform and asked to give the mean opinion score in the scale of 1 to 5. For evaluations, 10 people participated and asked them to observe naturalness, speaker similarity, and perceptual distortions present in each file and give

[2] :http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/ilprhts.php

Table 6.3: Subjective evaluation results of MOS and PT

Experimental Evaluation	Source model using different types of excitation				
	Impulse/noise	Mixed	STRAIGHT	ILPR	none
MOS	2.31±0.28	2.96±0.23	3.15±0.15	3.11±0.17	-
PT	9%	-	-	85%	6%
	-	32%	-	61%	7%
	-	-	45%	42%	13%

their scores accordingly. The average scores obtained from all the subjects are given in Table 6.3 along with standard deviation, which is computed for scores present within a 95% confidence interval of the mean. From the table, it can see that ILPR based source modeling outperform the impulse/noise based source model. Moreover, the proposed source model is slightly better than the mixed excitation with MOS of 3.11, which signifies the addition of the harmonic and the noise component helped in improving the naturalness and speaker similarity. Further, proposed method performs almost similar to STRAIGHT based excitation. The slight degradation may be due to the fact that random phase is not used in the proposed method for excitation, whereas random phase component is added to the excitation with the help of group delay in STRAIGHT method.

In addition, to know the distribution of score for male and female speaker bar chart is plotted in Figure 6.7. The bar plot shows the MOS with a standard deviation of all the 4 systems. From the bar plot, it can see that proposed method for the female speaker is significantly better than the mixed excitation, whereas, for male speaker proposed and mixed excitation performs almost similarly. This is due to the fact that the number of RMCEP parameter used for both male and female speakers has a constant value of 20. For the male speaker, the pitch period is high and more number of harmonics will be present within voiced frequency, increasing the RMCEP parameter may improve the synthesis quality, however, in this chapter only fixed RMCEP are used for comparison purpose. In the preference test, for each sentence subject were asked to listen two system shuffled randomly from 4 systems at a time and asked to choose any one system or prefer none of them as their preference. The percentage of preference scores can be viewed in Table 6.3. A clear improvement of the proposed method over the traditional impulse/noise and mixed excitation source model can be observed from the table, whereas it performs equally effective with respect to STRAIGHT method.

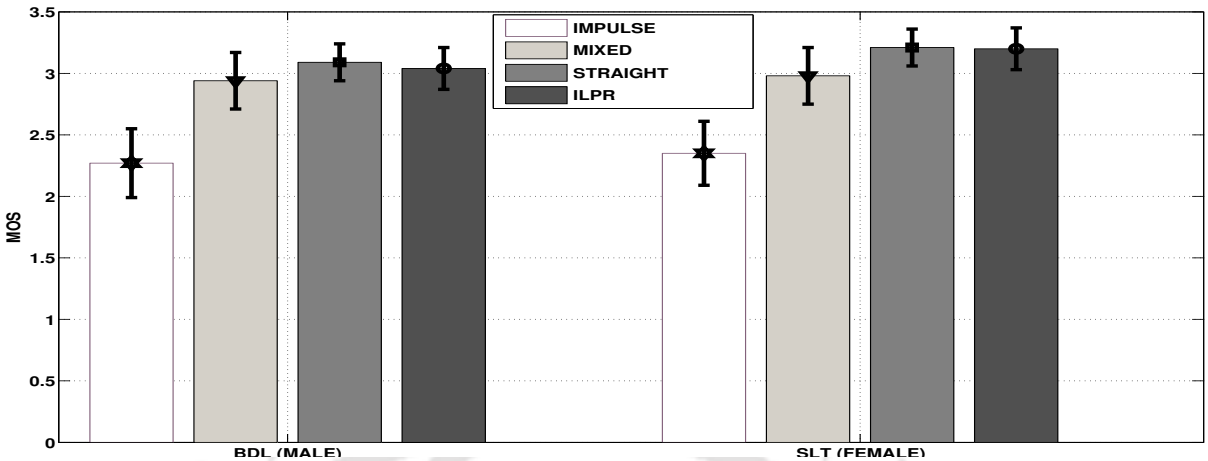

Figure 6.7: Average MOS of 4 HTS systems, impulse/noise, mixed, STRAIGHT and ILPR, respectively, for SLT and BDL speaker

Table 6.4: Objective evaluation results of PESQ and LSD

Experimental Evaluation	Source model using different types of excitation			
	Impulse/noise	Mixed	STRAIGHT	ILPR
PESQ	1.21±0.03	1.32±0.04	1.47±0.05	1.45±0.04
LSD	2.20±0.24	2.13±0.25	2.01±0.23	2.03±0.24

Objective evaluation

In this chapter, two objective measure is used, namely, perceptual evaluation of speech quality (PESQ) and log spectral distance (LSD) [101]. The PESQ scores obtained for 4 types of source modeling are tabulated in Table 6.4. It can be observed from the table that proposed ILPR based source model having a relatively low PESQ score of 1.45 with the standard deviation of ±0.04. However, even the scores obtained by the impulse and mixed excitation source model itself are relatively lower than the proposed excitation, which signifies the improvement in the synthesis quality of the proposed method. The STRAIGHT method performed slightly better than the proposed method, this is due to the fact that phase information is also modeled in STRAIGHT, which is ignored in the proposed method.

The LSD measure is evaluated between the original speech and the synthesized speech for the same text. The average LSD for all the 4 source model are given in Table 6.4 along with standard deviation. The LSD of the proposed excitation is lesser with distortion of 2.01, indicating the better spectral modeling of the proposed method comparable to that of impulse and mixed excitation. Whereas it performed almost equal with STRAIGHT method.

6.4 Phase modeling

In the previous section, importance of periodic and aperiodic representation for glottal activity region is shown using ILPR signal. In this section, importance of phase component present in the glottal activity region is shown.

6.4.1 Phase modeling using Integrated linear prediction residual

In the conventional way, STFT phase can be used for modeling phase. However, STFT phase has a unwrapping issue, whereas, cosine phase overcome this problem. Cosine phase is computed by forming analytic signal from ILPR. The obtained cosine phase obtained from the analytic signal contains only the phase component. The periodicity effect present in the residual signal is deemphasized.

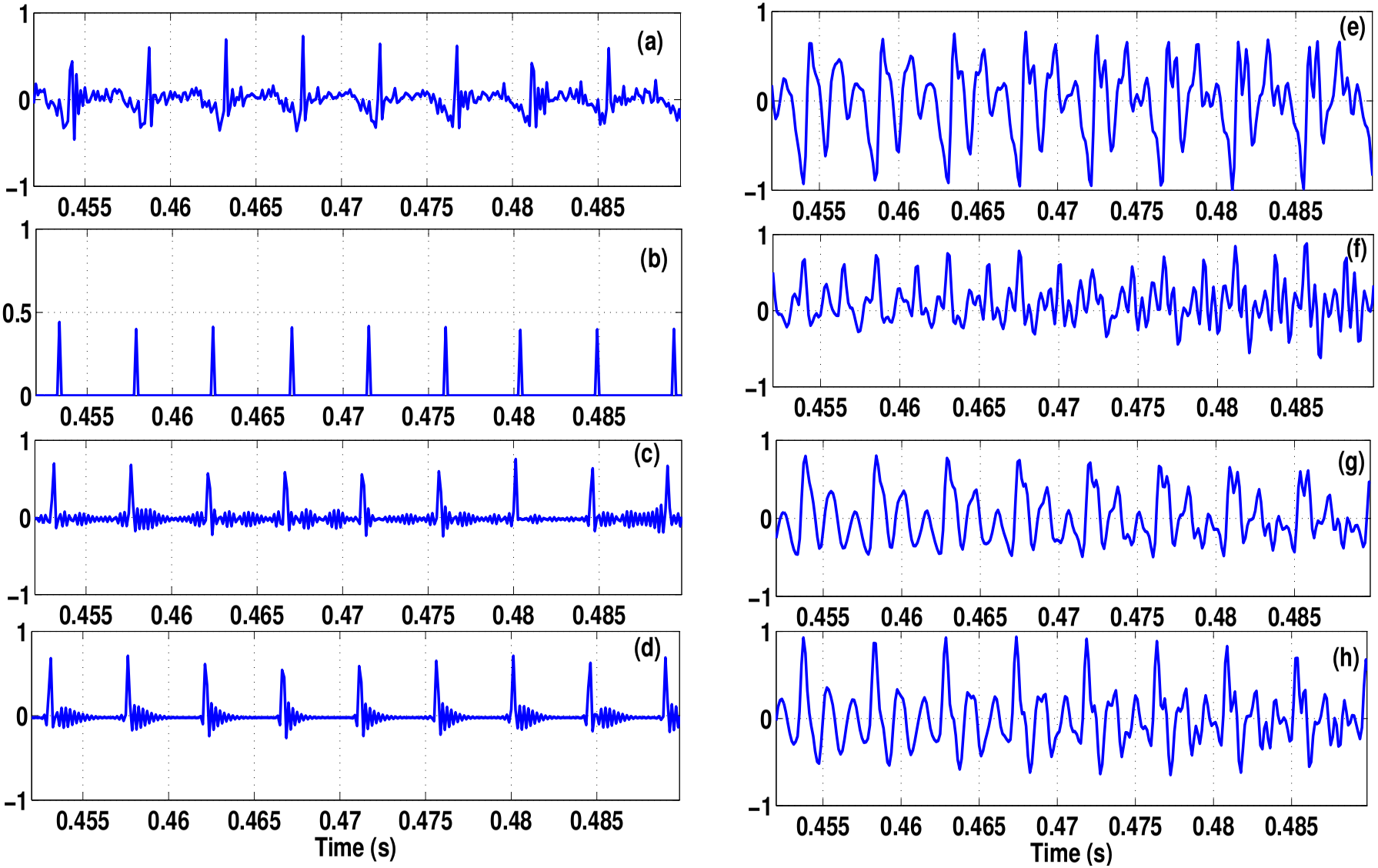

Figure 6.8: Comparison of different type of excitation with its synthesized speech: ((a)-(d)) represents the residual, impulse, cosine phase, and group delay phase excitation , respectively; ((e)-(h)) represent synthesized speech for the excitation signal shown in ((a)-(d)), respectively

To illustrate the significance of cosine phase, it is used as an excitation signal by adding it with epoch weighted impulse excitation [123, 162, 163]. Further details on the generation of the epoch weighted impulse excitation is discussed in our earlier work [156]. Figure 6.8(c) shows the excitation signal obtained after adding cosine phase with the epoch weighted impulse excitation. This figure is shown for the voiced portion of one utterance of SLT speaker taken from the CMU-ARCTIC database and its corresponding synthesized speech is shown in Figure 6.8(g). Here, for synthesis, an excitation signal is passed through LP filter.

Phase using group delay

In order to gauge the significance of cosine phase, it is compared with the group-delay based phase estimation used in the STRAIGHT [18]. In the STRAIGHT approach, random phase with a fixed group delay is added to impulse excitation using an all-pass filter. In this process, the random phase is generated and a fixed group delay is added to the higher frequency above 4 kHz. In [18], it is shown that the random phase addition to the zero-phase impulse excitation is useful in terms of the perceptual quality. Figure 6.8(d) shows the group delay phase excitation obtained by passing impulse excitation with an all-pass filter. Here, a fixed group delay is added with a standard deviation of 0.5 ms. However, the nature of the signal is not similar to the residual signal shown in Figure 6.8(a). Hence, there is a need for separately modeling the phase instead of using random phase.

Significance of cosine phase

The excitation signal obtained from the cosine phase is better than fixed group delay based phase excitation. In addition to that, even the synthesized speech is quite similar to that obtained using ILPR excitation. Further, to know the significance of the cosine phase for synthesis, its quality is evaluated using the PESQ measure. The score for the cosine phase excitation signal is shown in Table 6.5. For comparison, synthesis quality of the proposed excitation signal from the cosine phase is compared with that of the impulse excitation and group-delay based phase excitation (Figure 6.8(h)). From the PESQ score, it can be seen that the perceptual quality of the cosine phase excitation signal is better than both the zero phase impulse excitation and the group-delay based phase excitation.

Table 6.5: PESQ for different types of excitation

Excitation signal	PESQ
Integrated LP residual	4.5
Impulse	1.81
Cosine phase	2.54
Group delay phase	2.27

6.4.2 Cosine phase modeling

In this section, modeling of cosine phase using the all-pass filter is described. The all-pass filter is having unit magnitude response and captures the phase component. Further, obtaining the all-pass filter coefficients (APC) from the cosine phase signal is shown. Finally, synthesis framework to derive the excitation signal from the filter coefficients is showed to get the mixed phase response of speech signal.

Analysis

The cosine phase is modeled as an output of an all-pass filter excited by white Gaussian input sequence. An all-pass filter response ($H_{ap}[z]$) has a unit magnitude with poles of the system that are complex conjugate reciprocal locations of its zeros. The filter response is as follows:

$$H_{ap}[z] = \frac{a_N + a_{N-1}z^{-1} + a_{N-2}z^{-2} ..... + a_1 z^{-N+1} + z^{-N}}{1 + a_1 z^{-1} + ... + a_{N-1}z^{-N+1} + a_N z^{-N}} \tag{6.4}$$

where $[a_1, a_2, ..., a_{N-1}, a_N]$ are the desired APC. Further, the cosine phase is modeled as an output of an all-pass filter $H_{ap}(z)$ excited by Gaussian input sequence $x[n]$.

To estimate both APC and $x[n]$, some prior information about either $x[n]$ or APC are required. In this work, the estimation of APC is done similar to the approach given in [164] with the output of the all-pass filter being assumed as cosine phase. The energy in $x[n]$ should be concentrated around a few samples. Therefore, the input-output relation between $x[n]$ and $c[n]$ can be written in-terms of APC as follows:

$$c[n] = -\sum_{k=1}^{N} a_k c[n-k] + x[n-N] + \sum_{k=1}^{N} a_k x[n-N+k]. \tag{6.5}$$

The input signal to the all-pass filter is obtained by the stable and non-causal inverse filtering of phase

signal as reported in [164]. The input signal $x[n]$ can be written as

$$x[n] = -\sum_{k=1}^{N} a_k x[n+k] + c[n+N] + \sum_{k=1}^{N} a_k c[n+N-k]. \tag{6.6}$$

Here, the energy of both $c[n]$ and $x[n]$ is same since $x[n]$ is passed through the all-pass filter to give $c[n]$. The energy of the input sample $x[n]$ is given by:

$$e[n] = x^2[n]. \tag{6.7}$$

In order to concentrate the energy around a few samples, the entropy of $e[n]$ is minimized. The entropy, in turn being a function of APC, can be expressed in terms of an objective function ($J[a_k]$) given by,

$$J[a_k] = -\sum_{n=1}^{N} e[n] log e[n]. \tag{6.8}$$

To minimize the entropy, the gradient based minimization is carried out with respect to a_k's. The function $J[a_k]$ will be minimized for a particular set of $[a_k]$'s. In this chapter, the gradient descent algorithm is used as mentioned in [165] to minimize the entropy. The APC are iteratively updated until the minimum error is achieved below some epsilon value. In our work, epsilon value is chosen as 10^{-6} determined empirically. All the APC are in the form of a Gaussian function, thus implying that they converge and become suitable for training in HMM.

Synthesis using cosine phase

In synthesis stage, APC are used to synthesize speech by adding it as a phase to the excitation signal. The fundamental frequency (F0), SoE, and voicing decision are obtained from zero-frequency filter [110, 119]. In the voiced signal, an impulse excitation is generated according to F0 and passed through the all-pass filter with filter coefficients derived from the modeled APC to get the excitation signal. The resulting signal consists of zero phase impulse excitation signal added with phase. To derive the unvoiced excitation, white Gaussian noise is used. The excitation input signal is convolved with a time-varying filter, which in this chapter is the MLSA filter [13]. The coefficients of this MLSA filter are MCEP obtained from short-term Fourier transform. The block diagram for the synthesis module is shown in Figure 6.9. In order to get the harmonic and noise representation, in our earlier work [166], magnitude spectrum of ILPR signal is modeled with RMCEP parameters and to model the noise component, white Gaussian noise is weighted by the pitch adaptive triangular envelope. To

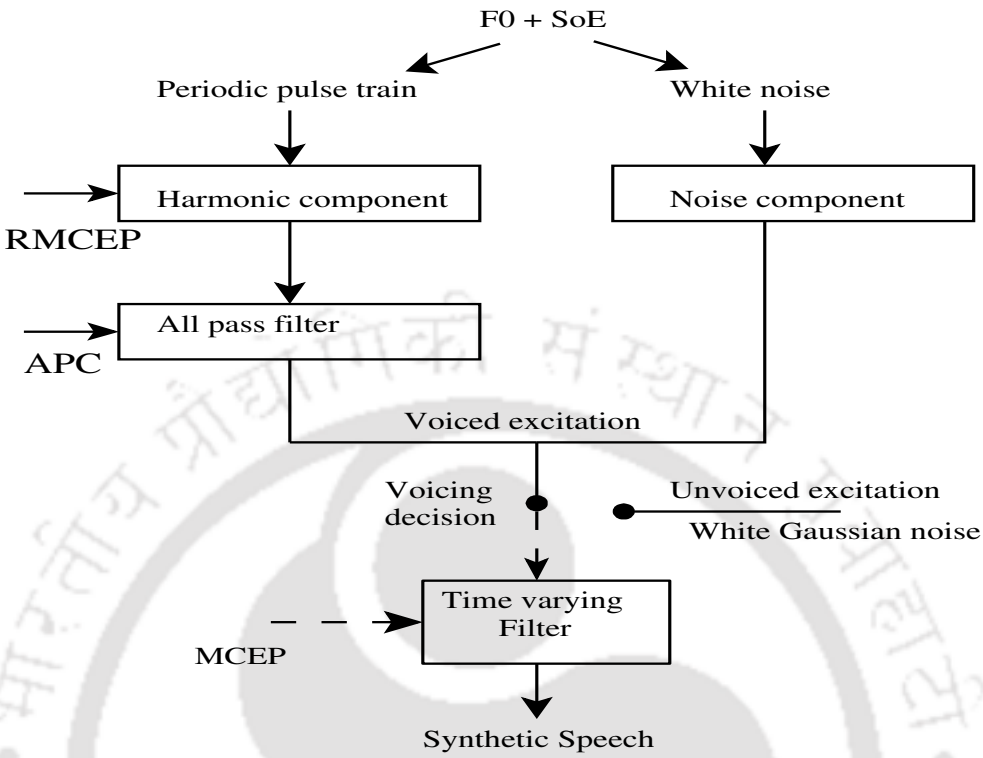

Figure 6.9: Integration of all-pass model to the proposed vocoder

this framework, cosine phase is integrated.

6.4.3 Experimental evaluation

The evaluation of the proposed phase modeling in SPSS is done using open source toolkit HTS [34]. Two speakers from ARCTIC database (1 male BDL speaker + 1 female SLT speaker) were used for evaluation [22]. The database consists of 1132 sentences for both speakers. During training the HTS system, 1000 sentences were used while remaining 132 sentences were used for evaluation. The parameters F0, MCEP, RMCEP, SoE, and APC are extracted for a frame rate of 5 ms. These parameters are trained in the HMM framework. The extracted parameters for each phoneme is modeled using 5 states, with each state consists of 7 streams. Basic features MCEP and F0 are modeled in first 4 streams. MCEP are modeled as continuous distribution, whereas, F0 is modeled as MSD [58]. In the fifth and sixth stream, RMCEP and SoE along with its derivatives are modeled, respectively. Along with these 6 streams, 20 APC parameters and its derivatives are modeled in the seventh stream to represent the cosine phase of the ILPR signal.

For comparison purpose, impulse excited (with MLSA filter) and STRAIGHT (fixed group-delay

Table 6.6: Objective evaluation results of only phase modeling in proposed method

Experimental Evaluation	Different methods		
	Impulse	**Group delay phase**	**Cosine Phase**
PESQ	1.26±0.04	1.38±0.06	1.51±0.03
LSD	2.57±0.27	2.35±0.28	2.07±0.24

based phase) systems are also developed in the HMM framework. In this work, version 40 of STRAIGHT is used to generate fixed group delay based excitation. To do a fair comparison of proposed phase modeling, in all the methods F0 is extracted from the zero-frequency filter and synthesis is done using MLSA filter. The synthesized speech from different methods can be listened from the following link[3].

Objective evaluation

To know the significance of phase for synthesis, objective measures, namely, PESQ [106] and LSD [101] are conducted. Before doing the evaluation, the duration of the synthesized speech and the original speech is aligned using the dynamic time warping algorithm. PESQ scores, computed for different phase modeling techniques are tabulated in Table 6.6.

Table 6.7: Objective results of proposed phase and aperiodicity modeling

Experimental Evaluation	STRAIGHT vs ILPR	
	STRAIGHT	**ILPR**
PESQ	1.38±0.06	1.74±0.03
LSD	2.35±0.28	1.83±0.21

It can be noticed that proposed cosine phase model has a relatively higher PESQ score of 1.51 with a standard deviation of 0.03 when compared to zero phase excitation and group-delay based phase excitation. This signifies the importance of preserving cosine phase in APC for improving the naturalness of synthetic speech. The LSD evaluation gives the distortion present in the frequency domain. This test is computed between reference original file and the synthetic file for the same text. The overall LSD evaluated for all the three methods is given in Table 6.6. The LSD of the cosine

[3] http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/apcshts.php

Table 6.8: Subjective evaluation results of only phase modeling in proposed method

Experimental Evaluation	HTS system using different types of phase information			
	Impulse	Group delay phase	Cosine phase	Same
MOS	2.71±1.01	3.15±0.92	3.22±0.81	-
PT	11%	-	80%	9%
	-	37%	46%	17%

phase based system has lesser distortion of 2.07. This infers that the better spectral reconstruction in case of cosine phase based excitation technique when compared to the excitation used in impulse and STRAIGHT techniques.

Subjective evaluation

To evaluate the perceptual quality of the proposed phase technique, MOS and PT measures are conducted. In MOS test, 20 sentences were used and asked the subjects to give their opinion score on the scale of 1 to 5 (1: poor and 5: excellent) by comparing with the original file. A total of 10 subjects were used in the listening test. For evaluations, listeners were asked to examine the naturalness of each file compared to the original file and give their scores. The mean and standard deviations of the scores obtained from the subjects are given in Table 6.8. From this table, it can be seen that the proposed phase modeling based on cosine phase outperform both the impulse and the STRAIGHT methods. In the preference test, subjects have to prefer between the two versions of the system selected from three developed systems at a time, by listening to the different sentence from each system. Further, they can prefer both the system as same as their choice. The percentage of preference scores with p-value from listeners can be viewed in Table 6.8. It can be seen that subjects preferred the proposed technique over both zero-phase and random phase methods with statistically significant p-values given by hypothesis tests. This indicates the importance of phase for improving the naturalness of synthetic speech. Further, proposed phase model is integrated with periodic and aperiodic representation discussed in Section 6.3. From Table 6.7 and 6.9, it is clear that both in-terms of objective and subjective evaluations, proposed ILPR phased source model in the glottal activity region is helped in improving the naturalness of speech. Further, the excitation model proposed method is better than STRAIGHT

Table 6.9: Subjective results of proposed phase and aperiodicity modeling with 95% confidence interval

Experimental Evaluation	Comparison of STRAIGHT and ILPR			p-value
	STRAIGHT	ILPR	Same	
MOS	3.16±0.88	3.31±0.95	-	
PT	32%	39%	29%	<0.01

excitation. The main reason for the improvement comes from explicit phase modeling in the proposed method, whereas, in the STRAIGHT method, the random phase is used for excitation.

6.5 Summary

In this chapter, different methods for deriving the excitation signal for source modeling is explored. Initially, an epoch based excitation by considering epochs as anchor points to select the samples from the residual signal is proposed to represent aperiodic and phase components. The experimental studies indicate that a small set of samples around the epochs are sufficient to preserve the naturalness and prosodic information. Further, the results indicate that both the aperiodic and phase knowledge are important from the point of preserving the naturalness and prosodic information.

The source modeling using residual signal by storing samples around epochs in a statistical framework is difficult. Hence, a method is proposed by dividing the magnitude spectrum of an ILPR signal into two bands, harmonic and noise component. The harmonic component is represented by RMCEP and the noise component by SoE weighted triangular shaped random noise. The proposed ILPR excitation source model is compared with impulse, mixed, and STRAIGHT excitation source modeling. Through the subjective and objective tests, the proposed method clearly outperforms the baseline version of the HTS system and mixed excitation source model both in terms of naturalness and speaker similarity. Further, proposed method gave an almost similar performance with STRAIGHT method.

Finally, the phase information present in ILPR signal is analyzed to improve the naturalness of synthetic speech. The phase signal is obtained from the cosine phase of ILPR signal and modeled using the all-pass filter coefficients. Here, the filter coefficients are optimized using the gradient descent algorithm. Initially, in the analysis-synthesis framework, the significance of cosine phase for speech synthesis is compared with zero phase impulse excitation and group delay based phase excitation. Then the APC are trained in HMM along with excitation and vocal tract spectrum coefficients. The

obtained phase signal from the modeled APC is compared with zero phase impulse excitation and group delay phase excitation used in STRAIGHT method along with the aperiodic excitation. The synthesis quality of the proposed method is better than both the methods in terms of naturalness.

All the previous chapters till now is focused on different modules present in speech production mechanism like voicing decision, vocal tract, and source modeling for speech synthesis. In the next chapter, we tried to combine all these modules in a single framework for the task of speech synthesis. The significance of each of these components for improving the quality of speech is also discussed. Further, in the next chapter, proposed framework is also tested for different Indian languages and DNN framework.

7

Suprasegmental, System, and Source features for Speech Synthesis

Contents

7.1 Introduction 140
7.2 Glottal activity region based processing 141
7.3 Proposed analysis/synthesis framework 144
7.4 Experimental evaluation 145
7.5 Summary and Discussion 151

Objective

The objective of this chapter is to demonstrate the significance of combining different features present in glottal activity region for SPSS. Glottal activity regions constitute the majority of the speech sound units and these regions are perceptually very important to decide the high quality of speech. Different features present in the glottal activity regions are broadly categorized as suprasegmental, system, and source features, which essentially represent the prosodic, intelligibility, and naturalness of speech, respectively. Combining these features helps in bringing the advantages present in each of these features to the synthesizer. Further, this also helps in getting best features to the framework results in the enhancement of the overall perceptual quality of SPSS.

7.1 Introduction

In the previous chapters, the focus is on the extraction of different acoustic features present in the glottal activity region and their usage for the speech synthesis. In particular, the effectiveness of these features is shown for speech synthesis using SPSS framework. In Chapter 3, glottal activity regions are detected using the source features like SoE, NAPS, and HOS, which represent amplitude, periodicity, and asymmetry of the glottal pulse signal, respectively. In Chapter 4, the significance of glottal activity features for speech synthesis is shown. Further, in that chapter, the voicing decision is improved using classifiers and its significance is shown for SPSS framework. The 2-D based processing of speech signal in the glottal activity region using Riesz transform is presented in Chapter 5. Riesz transform provides smoothed vocal tract spectral envelope, which is comparable to STRAIGHT vocal tract spectrum. In addition, Riesz transform based processing of speech also provides voicing decision, F0, and aperiodicity component. In Chapter 6, aperiodic and phase component representation for source modeling is derived using integrated linear prediction residual (ILPR). In this chapter, we propose a unified framework for speech synthesis by processing glottal activity region of speech signal. In the proposed framework, different features computed previous chapters are combined to improve the quality of the synthesizer. Further, proposed framework is tested in statistical framework using HMM for different Indian languages like Assamese and Manipuri. Finally, proposed method is also tested in deep neural network (DNN) framework.

The structure of this chapter is organized as follows. Section 7.2 details the suprasegmental, system, and source features present in the glottal activity region. Section 7.3 explains the proposed architec-

ture for speech synthesis using glottal activity region based processing. The proposed framework is evaluated in section 7.4. Finally, Section 7.5 summarizes the work and concludes.

7.2 Glottal activity region based processing

The glottal activity region constitutes the voiced sounds which are characterized by different acoustic cues [167]. These different cues can be categorized as suprasegmental, system, and source features, representing prosodic, intelligibility, and naturalness of speech, respectively. The brief overview of different features present in glottal activity region, which are explored in our earlier chapters is explained in this section. Further, their significance to the context of improving the quality of speech synthesis is also discussed here.

7.2.1 Suprasegmental features

Prosody or melody of the speech is usually lost for more than one segments and the features derived in these regions are called as suprasegmental. Basically, F0 , loudness, and duration represent the suprasegmental features [168]. Moreover, suprasegmental features differ from segmental features by the fact that suprasegmental features are defined by a comparison of segments in a sequence, whereas segmental features are recognized by the specific segment itself. In this thesis, only F0 and loudness features are studied for speech synthesis task. The duration features are not explored in this thesis. Even though F0 and voicing decisions are computed segment wise, but these features are will be there for longer frames. Hence, these two features are called as suprasegmental features in this work. For duration features, not much exploration is done.

Fundamental frequency (F0)

F0 is a tonal or prosodic feature which represents the prosody of the speech signal. Further, F0 feature also characterizes a particular region of speech is voiced or not. Based on F0, excitation signal is generated. Hence, accurate estimation of F0 is required to get the high quality speech. In our work, two F0 estimation methods are used. The first method is based on zero-frequency filter (ZFF) and the second one is based on Riesz carrier [110]. Relatively, F0 from ZFF gives an accurate estimation when compared to Riesz carrier. F0 estimation from the ZFF method gives more smoothed F0 estimation when compared to Riesz method. Hence, in this combined work, F0 is estimated from ZFF method.

Voicing decision

Voicing decision is very important for speech synthesis, which says a particular speech frame is tonal or not. Hence, it is important to have an accurate detection of voicing decision. In chapter 3, different features present in glottal activity region like SoE, NAPS, and HOS are used for voicing decision. Further, in Chapter 4, voicing decision from glottal activity features are enhanced using support vector machine (SVM) classifier. In Chapter 5, coherence spectrum and pitch map are used for the voicing estimation. The SVM classifier is found to be more accurate for voicing decision compared to other methods. Hence, in this combined framework, SVM classifier method is used as voicing decision.

Loudness

Loudness or amplitude in glottal activity region usually occurs due to variation in the transglottal air pressure during the production of the voiced sounds [2]. In this work, to measure this activity present in the glottal activity region, SoE derived from ZFF is used. In [108], it is shown that SoE derived from ZFF is better than EGG signal. It is also shown that SoE represent the loudness attribute. Hence, this feature is used as a parameter to represent the loudness in this chapter.

7.2.2 System feature

System feature represents the vocal tract shape for each sound units, which is characterized by the resonance structure. The conventional methods for vocal tract feature extraction usually attempt to separate source and system features [77, 80]. However, computed speech spectrum from those methods contains temporal and spectral fluctuations, which results in less intelligibility in the synthesized speech [18]. To get the smoothed spectral envelope many algorithms have been applied in the literature [18, 77–80]. In Chapter 5, Riesz transform based processing of speech spectrogram in the 2-D domain is proposed to get smoothed vocal tract envelope. The envelope is almost similar to STRAIGHT method, hence, in this work, Riesz transform based features are used for vocal tract representation. The smoothed envelope computed from Riesz transform is represented in parametric form using MCEP parameter.

7.2.3 Source features

Speech signal can be synthesized by exciting vocal tract transfer function with source model. Simple excitation schemes will not give a natural speech. Hence, while generating an excitation signal, first different components present in excitation signal like aperiodic and phase component are extracted.

Periodic and aperiodic components

The voiced speech is usually assumed to be produced from periodic excitation over a short frame segment. However, even within a short frame segment variations in the strength of excitation (shimmer) and periodicity (jitter) are observed due to constant movements of vocal tract organs [25, 91]. Further, turbulence noise will also be present in voiced speech [2]. Hence, speech signal will not be perfectly periodic in voiced speech. The simple F0 based impulse excitation will not be able to accommodate this aperiodic component present in the voiced speech. Hence, we need to incorporate this aperiodicity while generating the excitation signal. In the recent literature, two types of excitation are generated within voiced speech segment to represent the periodic and aperiodic component of speech. In our work, two methods are proposed for aperiodic representations. First one is using ILPR signal, where the harmonic component is modeled using MCEP computed on the residual signal, whereas, noise component is modeled using white Gaussian noise is weighted pitch adaptive strength weighted triangular envelope. In our second work, periodic and aperiodic components are computed from the coherence map, which gives an indication of the amount of harmonic component present in the different frequency bands. The coherence map from Riesz transform gives a better representation of aperiodic component. Hence, in this work, coherence map is used as an indicator for periodic and aperiodic decomposition. Further, this coherence map is represented in parametric form using band aperiodicity (BAP) parameters.

Phase component

In the conventional speech processing, normally, only magnitude characteristics of speech is employed by ignoring phase characteristics. However, phase plays a prominent role for high-quality speech synthesis. In chapter 6, the phase signal is captured using ILPR signal for synthesis application. It is shown that cosine phase of ILPR signal is represented in parametric form using all-pass filter coefficients (APC). The APC are computed from cosine phase using the iterative procedure, which improves the perceptual quality. Henceforth, in this combination of different acoustic features, phase

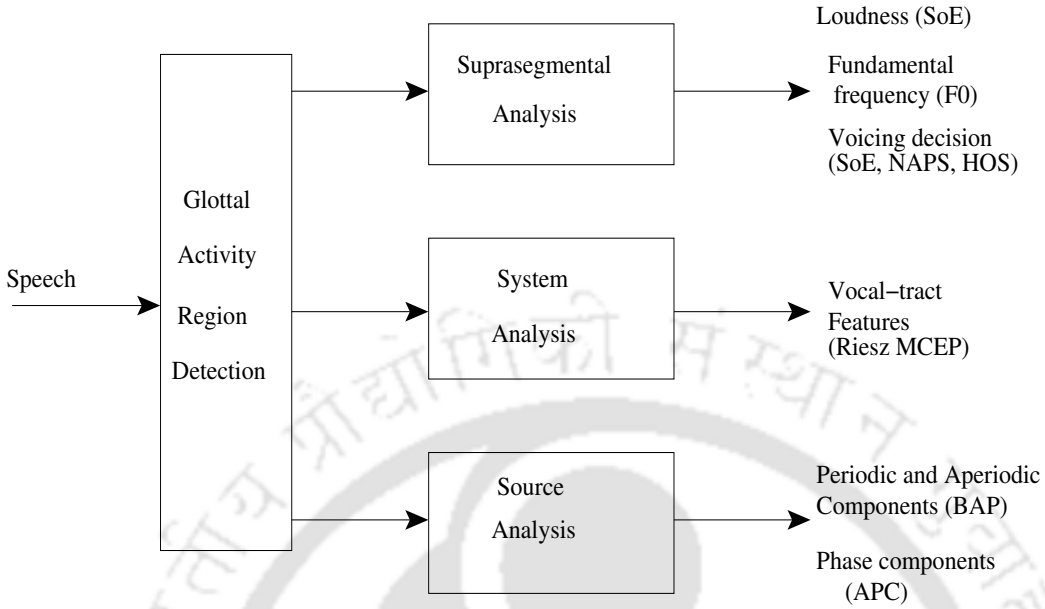

Figure 7.1: Proposed Analysis framework for Glottal activity based processing for SPSS

component is also modeled.

7.3 Proposed analysis/synthesis framework

The main motivation for proposed framework is to explore different acoustic features present in the glottal activity region and showing its significance for speech synthesis. The glottal activity regions are characterized by a different set of features, which can be categorized as suprasegmental, system, and source features.

The proposed framework is categorized as analysis and synthesis stage. The analysis and synthesis stage block diagram is shown in Figure 7.1 and 7.2, respectively. During the analysis stage, the speech frames are analyzed and classified as glottal or non-glottal regions based on whether the activity of glottis is present or not. After classifying the speech frames to glottal activity region, suprasegmental, system , and source analysis are made to get different features as mentioned in the previous section. These features (MCEP, F0, BAP, APC, SoE, NAPS, and HOS) are trained in HMM using EM algorithm [11].

During the synthesis stage, frame wise features (MCEP, F0, BAP, APC, SoE, NAPS, and HOS) are computed from statistical models for a given text. From these parameters, speech is synthesized using the vocoder shown in the block diagram 7.2. The synthesis equation is given as follows:

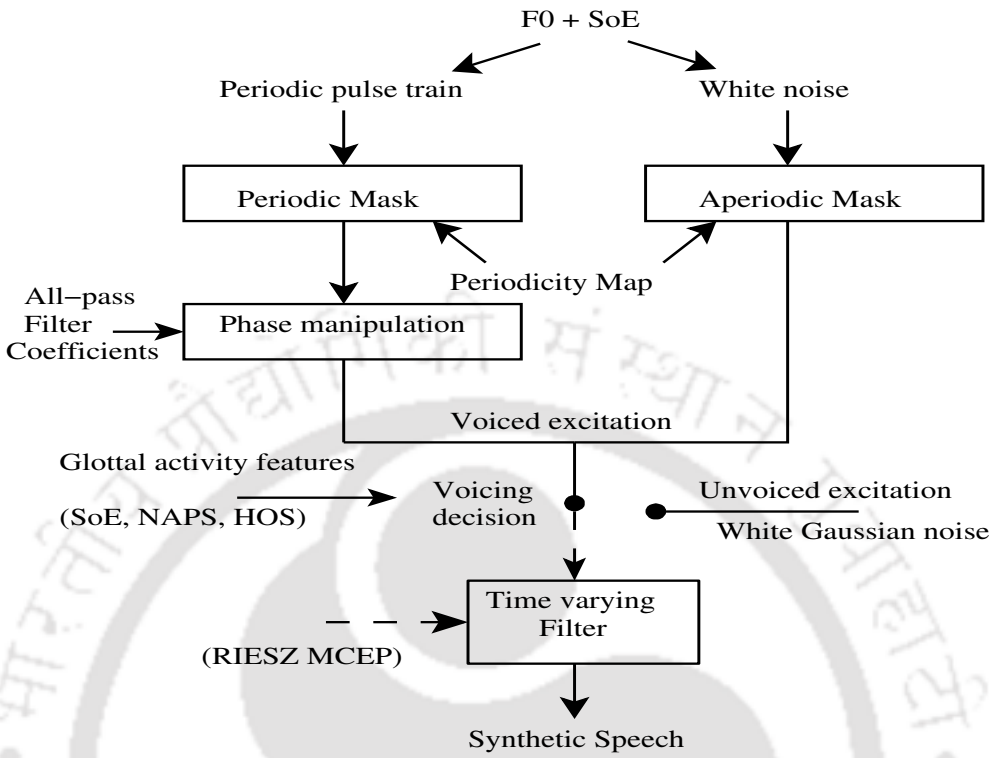

Figure 7.2: Proposed Synthesis framework for Glottal activity based processing for SPSS

$$S(\omega) = E(\omega)V(\omega) \tag{7.1}$$

where $V(\omega)$ smoothed vocal tract envelope computed from Riesz transform and $E(\omega)$ is the excitation spectrum given by:

$$E(\omega) = [1 - A(\omega)]I(\omega)\Phi(\omega) + A(\omega)W(\omega) \tag{7.2}$$

where $A(\omega)$ is the aperiodic spectra computed from Riesz transform, $I(\omega)$ is the spectra computed from impulse excitation, $\Phi(\omega)$ is the phase spectrum computed from the all-pass filter, and $W(\omega)$ is the white Gaussian noise spectrum computed from the random noise weighted by sigmoid function in the higher frequency. The speech signal is reconstructed back by taking the inverse Fourier transform of the $S(\omega)$ for each frame and overlap-add method to get the synthetic speech.

7.4 Experimental evaluation

In this section, proposed combination of features is tested in SPSS framework with HTS software using ARCTIC database. Along with ARCTIC database, HMM based speech synthesis systems

are also developed with low resource Indian languages like Assamese and Manipuri. The proposed framework is also tested in the DNN based speech synthesis system.

7.4.1 Database

In all the previous chapters, for evaluation of different features, ARCTIC database which consists of English sentences are used. Similarly, in this work also ARCTIC database is used. Additionally, Assamese and Manipuri speech database are used for evaluating the proposed framework [146, 169]. Assamese and Manipuri are two Indian languages spoken in north-east of India.

Assamese is an eastern Indo-Aryan language spoken mainly in the state of Assam. The database was collected from 2 native Assamese speakers, with one male speaker and one female speaker. Both are professional speakers and their voice quality is tested before taking recordings. The subjects are asked to read articles from Assamese newspaper. Recordings are carried out in a soundproof recording booth with data recorded at 48 kHz sampling frequency and 16 bits/sample resolution. The speech files are saved in wave format and are split manually into speech segment of about 5-6 s length using the Wavesurfer toolkit [121]. Duration of the speech corpus is about 10 hours. The details of the database are summarized in Table 7.1. Two systems are developed for one male and one female speaker taken from Assamese database. In this work, only 5 hrs of recording from both the speakers are taken, as adding more data is not helped in improving the synthesis quality.

Manipuri is one of the Tibeto-Burman language spoken in Northeast India, which has its own script Meetei Mayek and literature. At present, Manipuri uses Bengali script for writing. It is also one of the low resource languages. Hence, Manipuri data is collected from two speakers specifically aimed to build TTS systems. For the recordings, Manipuri children stories are taken and while recording, the speaker is instructed to narrate the story under controlled environment i.e., controlled word rate and controlled amplitude. The speaker is of native Manipuri professional story reader. In this database also recording are done at 16 bits mono channel with sampling rate 48 kHz. Further, details of Manipuri corpus is given in Table 7.1. For developing SPSS system, similar to Assamese database, only 5 hr of data is used from Manipuri database.

7.4.2 HMM based speech synthesis system

SPSS system is developed from the proposed glottal activity region based processing by combining suprasegmental, source, and system features. In ARCTIC database, two speakers SLT and BDL are

Table 7.1: Assamese and Manipuri database showing the number of unique words, syllable and duration of each word

Database	Hrs	Unique words	Unique syllables	words/sec			
				min	max	avg	std
Assamese	17	90,444	3474	0.81	5.01	3.48	1.20
Manipuri	10	26,203	3817	0.512	2.85	1.09	0.25

Table 7.2: Speech parameters used per frame for the proposed Glottal activity region based system vs STRAIGHT

Feature	**Number of parameters**	
	Proposed	**STRAIGHT**
Mel-cepstral coefficients (MCEP)	35	35
Fundamental frequency (F0)	1	1
Band aperiodicity (BAP)	25	25
Glottal activity parameters (SoE+NAPS+HOS)	3	-
APC	20	-

used. Both the speakers consist of 1132 sentences, out of which 1000 sentences are used for training and remaining are used for testing. The features are computed from the database for 25ms with a frame rate of 1 ms. The number of parameters computed for each frame is mentioned in Table 7.2. For comparison purpose STRAIGHT system is also developed with a number of parameters mentioned in Table 7.2. In STRAIGHT system, F0 feature is computed from TEMPO algorithm is trained as MSD in HMM framework, whereas other parameters MCEP and Band aperiodicity (BAP) are modeled as a continuous probability distribution. In the proposed framework, total five streams are used and all the streams are modeled with the continuous probability distribution.

The details of each stream are given below:

- The first stream for MCEP and its delta, and delta-delta.
- The second stream for F0 and its derivatives.
- The third stream consist of band aperiodicity computed from Riesz transform and their derivatives.

- The fourth stream for SoE, NAPS, and HOS features and its derivatives.
- The fifth stream consist of APC and its derivatives to represent phase component.

The speech synthesis procedure is given in the block diagram 7.2. To build the Assamese and Manipuri systems, we followed the similar procedure except that 5 hr data is used for developing systems. Some of the synthesized files from the proposed and STRAIGHT method can be accessed from the following link [1].

The synthesized files are evaluated by both objective and subjective evaluations. The objective results of the two systems using the proposed and STRAIGHT framework are presented in Table 7.3. It is observed that proposed method performs better than STRAIGHT method in terms of lesser Mel-cepstral distance (MCD), F0 root mean square error (RMSE), and voicing error (V/UV). In the case of aperiodicity parameter, STRAIGHT method gives lesser band aperiodicity (BAP) spectrum error. In general, objective results confirm that the proposed method is better than STRAIGHT method.

Table 7.3: Comparison of objective results of STRAIGHT and proposed analysis/synthesis framework for HTS

Database	**Objective Measure**	**MCD (dB)**	**BAP (dB)**	F0 **RMSE Hz**	**V/UV %**
ARCTIC	STRAIGHT	4.36	1.40	9.31	11.66
	Proposed method	4.12	2.05	8.55	4.18
Assamese	STRAIGHT	5.12	2.45	10.30	12.31
	Proposed method	4.95	3.02	9.07	6.38
Manipuri	STRAIGHT	5.96	1.95	11.21	13.87
	Proposed method	5.41	2.12	10.04	9.49

Similar to previous chapters, two subjective tests are done to evaluate the proposed glottal activity region based framework. The first test is mean opinion score (MOS), which is evaluated with 25 sentences synthesized from proposed framework for ARCTIC, Assamese, and Manipuri database. Similarly, for comparison, subjective evaluation is done for STRAIGHT also. The MOS of the both the system is shown in Figure 7.3. From the scores, it is clear that proposed system slightly better than STRAIGHT system for all three languages. The reason for the improvement is due to the two factors:

[1]:http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/combinedhts.php

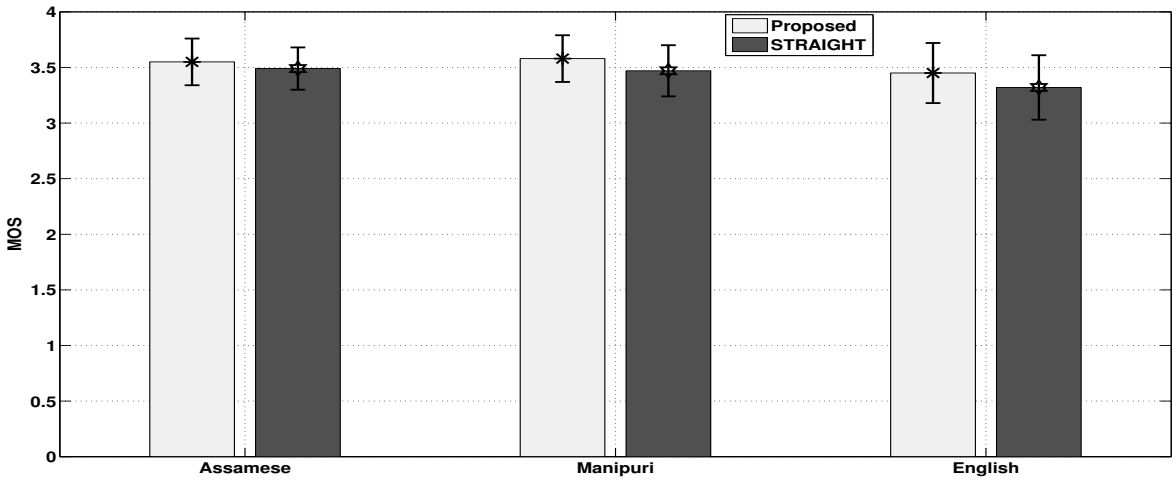

Figure 7.3: Average MOS of Proposed and STRAIGHT system for Assamese, Manipuri and English Database in HTS framework

Table 7.4: Preference test (PT) results

Experimental Evaluation	Different techniques		
	Proposed	STRAIGHT	Same
ARCTIC	37%	30%	33%
Assamese	39%	32%	29%
Manipuri	40%	34%-	26%

one is explicit phase modeling is done in the proposed framework, and secondly, voicing decision of the proposed system is slightly better than STRAIGHT framework. In the preference test, subjects have to prefer between the proposed and STRAIGHT systems at a time, by listening to the different sentence from each system. Further, they can prefer both the system as same as their choice. The percentage of preference score from listeners can be viewed in Table 7.4. It can be seen that subjects preferred the proposed technique over both STRAIGHT method with significant p-values (< 0.01). This indicates the importance of glottal activity region based processing for speech synthesis task.

7.4.3 DNN based speech synthesis system

To build the DNN based speech synthesis system, Merlin toolkit is used [35]. In Merlin system, linguistic features are taken as input and tried to predict acoustic features, which are then passed to a vocoder to produce the speech waveform. Many neural network architectures are implemented in Merlin software, in this work, we have chosen recurrent long short-term memory (LSTM) based

recurrent neural network (RNN). For evaluation, only two speakers (SLT and BDL) from ARCTIC database are used. To evaluate for other languages like Assamese and Manipuri, linguistic features have to be given, which is not yet explored in our study. Further, the main aim of the combination framework is to know the suitability of the proposed features to DNN.

The speech signal is sampled at a rate of 48 kHz. 1000 utterances are used for training, 70 as a development set, and 72 as the evaluation set. The input features for the system consisted of 491 features. 482 of these are derived from linguistic context, including quinphone identity, part-of-speech, and positional information within a syllable, word, and phrase, etc. The remaining 9 are within-phone positional information: frame position within HMM state and phone, state position within phone both forward and backward, and state and phone durations. The frame alignment and state information are obtained from forced alignment using a monophone HMM based system with 5 emitting states per phone [35].

For comparison, STRAIGHT vocoder is used in this experiment. In STRAIGHT method, 60 MCEP parameter to represent STRAIGHT spectrum, 25 BAP parameters to represent aperiodicity, and fundamental frequency on the log scale at 5 msec frame rate is used. Similarly, for the proposed method also, 60 dimensions MCEP computed from Riesz transform, 25 BAP parameters computed from coherence map, and fundamental frequency in log scale computed from ZFF is used. However, phase component is not modeled in this experiment. Before training, the input features are normalized using min-max to the range [0.01, 0.99] and output features are normalized to zero mean and unit variance. At synthesis time, Maximum likelihood parameter generation (MLPG) is applied to generate smooth parameter trajectories from the de-normalized neural network outputs. The objective results of the proposed method are better than the STRAIGHT method is shown in Table 7.5. In particular, in the proposed framework, error in V/UV, F0, and MCD with respect to original signal are low compared to STRAIGHT. Whereas, in STRAIGHT system, aperiodicity component is modeled better. Some of the synthesized files from the DNN systems can be accessed from the following link [2].

For DNN systems also, MOS and PT tests are conducted. Total of 10 subjects were used in the listening test. 25 sentences from both proposed and STRAIGHT method is given to subjects. The MOS test gives a score of 3.65 compared to STRAIGHT system with a score of 3.61. The relatively DNN based system is better than HTS system. Further, the improvement in the proposed system is not much, since phase is not modeled in DNN system. In the preference test, 41% waveforms felt as

[2] :http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/combinedhts.php

Table 7.5: Comparison of Objective results using STRAIGHT and proposed analysis/synthesis framework for SPSS

Database	Objective Measure	MCD (dB)	BAP (dB)	F0 RMSE Hz	V/UV %
ARCTIC	STRAIGHT	3.27	1.01	8.39	5.12
	Proposed method	3.16	1.78	7.27	4.06

same indicating the similarity in both proposed and STRAIGHT method. The preference to proposed method is around 36% indicating that subjects were not able to distinguish between two systems. However, proposed method works even in the DNN system.

7.5 Summary and Discussion

In this chapter, glottal activity region based processing of speech signal for speech synthesis is shown. The different features present in the glottal activity regions are broadly categorized as suprasegmental, system, and source features, which essentially represent the prosodic, intelligibility, and naturalness of speech, respectively. The quality of synthesized files is measured using objective and subjective evaluations. The results show that proposed method performed better than STRAIGHT system. The effectiveness of proposed method is studied in statistical framework using HMM and DNN. Further, proposed framework is tested in statistical framework using HMM for different Indian languages like Assamese and Manipuri. The results show that using different components derived from glottal activity region from the proposed method is equally effective like STRAIGHT method.

In this work, different components present in the speech production mechanism are modeled to improve the intelligibility and naturalness of SPSS speech. In particular, different factors which are related to naturalness and intelligibility are shown.

- Intelligibility mainly refers to message information and which is represented by the vocal tract envelope. To get the better intelligibility, the vocal tract envelope has to capture spectro-temporal dynamic variations without having any source components. 2-D based processing of speech spectrum using Riesz transform gives an alternative mechanism to get smoothed vocal tract envelope without any temporal and spectral fluctuations. This method helped in enhancing the intelligibility of speech.

- Naturalness is a perceptual phenomenon and related how closely synthesized speech matches with original speech. Based on our work, the naturalness of SPSS speech is present in different aspects, like voicing decision, aperiodicity, and phase components. In this work, it is shown that glottal activity region based processing is used to capture these different components, which internally helped in improving the naturalness of synthesized speech.
- In addition, to have emotional or prosodic speech, we need to have suprasegmental features like duration, strength, and pitch information. In this work, the explicit study of suprasegmental features is not done. However, the effect of suprasegmental parameters like the strength of excitation and accurate pitch information is studied to improve the quality of speech. Further, duration analysis can be done as future work to improve the prosody of the proposed framework.

In the next chapter, a summary of all the chapters of the thesis is presented.

8

Summary and Conclusions

Contents

8.1 Summary of the work 154
8.2 Contributions of this thesis 156
8.3 Directions for future work 157

8.1 Summary of the work

In this thesis, we made an attempt to demonstrate the significance of glottal activity region based processing for speech synthesis using the statistical framework. To achieve this, first, a method is proposed for the detection of glottal activity region. Next, three different categories of acoustic features are analyzed from the glottal activity region. The three different categories of features include suprasegmental, system, and source representation. To get good results from these features, a combined framework for speech synthesis is proposed. Review of different works introduced in this thesis is presented below.

Some of the important results of the thesis are as follows:

(i) **Glottal activity region detection:** The major activity during speech production is glottal activity, which was earlier detected using the SoE. In this thesis, the NAPS and HOS are used as additional features for detecting glottal activity region. The three features, namely, SoE, NAPS, and HOS, are, respectively indicators of different attributes of glottal activity region, namely, energy, periodicity, and asymmetrical nature of the source signal. The effectiveness of these features is analyzed using the differential electroglottograph signal, zero-frequency filtered signal, and integrated linear prediction residual (ILPR), as representatives of the source signal. The combination of glottal activity information from the three features outperforms the other state-of-the-art algorithms for voicing detection, which demonstrates the different information represented by each of these features.

(ii) **Glottal activity features for speech synthesis:** The different glottal activity features like epoch locations, epoch strength, aperiodicity, phase information, and voicing decision are useful for high quality speech. In existing methods, voicing detection relies mostly on fundamental frequency F0, which may result in errors when the prediction is inaccurate. The voicing decision is computed from the different glottal activity features present in the excitation source signal. The glottal activity features SoE, NAPS, and HOS are used for the voicing decision. To improve the voicing decision and to avoid the heuristic threshold for classification, glottal activity features are trained using different statistical learning methods such as a k-nearest neighbor, support vector machine (SVM), and deep belief network. The voicing detection performs best with SVM classifier and its effectiveness is tested by using it as voicing decision for SPSS. The glottal

activity features SoE, NAPS, and HOS are trained in HMM along with F0 and MCEP to get the voicing decision. The objective and subjective evaluations demonstrate that the proposed method improves the naturalness of synthetic speech.

(iii) **Riesz transform for speech synthesis:** The traditional analysis/synthesis methods are based on the fixed short time frame analysis, resulting errors in formant estimation. Hence, 2-D spectro-temporal analysis/synthesis method is proposed using Riesz transform. The 2-D spectro-temporal analysis is motivated by the fact that the human auditory cortex is tuned to localized spectro-temporal modulations. The spectro-temporal receptive fields of these cortical cells look like 2-D spectro-temporal Gabor filters. The demodulation of 2-D spectro-temporal patches using Riesz transform yields smoothed spectral envelope, carrier signal, and coherence map, representing vocal tract spectrum, source signal, and periodicity using a single framework. The analysis/synthesis representation gives better synthesis quality than the state-of-the-art STRAIGHT vocoder. Further, smoothed spectral envelope is compactly represented using MCEP and trained on the HMM framework and then these parameters used as input to the synthesis module of vocoder framework. The synthesized speech is compared with state-of-the-art STRAIGHT system, which is based on the pitch-synchronous analysis. The synthesized files are measured using objective and subjective evaluation. The results show that for voiced sound, proposed method performed better than the STRAIGHT system and overall for continuous speech it performed equally well with STRAIGHT framework. The effectiveness of Riesz transform is further studied in statistical framework using HMM based speech synthesis and results show that decomposition of the spectrum into an envelope, carrier, and coherence map is equally effective like STRAIGHT method.

(iv) **Integrated linear prediction residual for source modeling:** Source modeling for SPSS is proposed using ILPR. The nature of ILPR waveform resembles the glottal flow derivative signal and keep the speaker characteristics in a better way. The different events present in the source signal, namely, glottal closure, glottal opening, onset of burst, frication and a small number of excitation instants around them known as epochs. The speech signal is processed independently by the zero-frequency filter to obtain epoch locations. These events are used as anchor points for extracting the different representation present in glottal activity regions like strength around epochs, aperiodicity, and phase component. In the analysis/synthesis frame-

work using these components helped in improving the naturalness of speech. However, training these components directly in the statistical framework is difficult. Hence, parametric representation is explored. First, ILPR signal is modeled in the frequency domain by dividing the spectrum into two bands to characterize periodic and aperiodic components of the voice speech segment. The periodic component of ILPR signal below the maximum voicing frequency (f_m) is modeled using residual Mel-cepstral coefficients called as RMCEP, whereas aperiodic component above f_m is modeled by pitch adaptive triangular noise envelope weighted by the SoE. The RMCEP and SoE are trained using HMM framework along with MCEP and F0 representing vocal tract information and fundamental frequency, respectively. The synthesized speech by the proposed source modeling reduces the buzziness and improves the speaker similarity compared to the conventional impulse/noise and mixed excitation source modeling and it is comparable with STRAIGHT based excitation. Next, phase component is modeled using cosine phase by compactly represented using all-pass filter coefficients. These coefficients are obtained from the iterative procedure by assuming cosine phase as the output of the all-pass filter. The addition of phase improve the naturalness and gives synthesis quality better than STRAIGHT framework.

(v) **Combination of suprasegmental, system, and source features for Speech synthesis:** A combined framework for speech synthesis is demonstrated by processing speech in glottal activity region. These regions constitute the majority of the speech sound units and they are perceptually very important for high voice quality. The glottal activity features are broadly categorized as suprasegmental, system, and source features, which essentially represent the prosodic, intelligibility, and naturalness of speech, respectively. Combining various features present in the glottal activity region aids in bringing the advantages present in each of these features to the synthesis system and getting the best features results in the enhancement of the overall perceptual quality of SPSS.

8.2 Contributions of this thesis

The major contributions of the research work reported in this thesis includes

- Glottal activity region detection using three glottal source features, namely the SoE, NAPS, and HOS.

- Using glottal activity region detection as a voicing indicator and improving the accuracy of

voicing decision with the classifiers. Finally, applying voicing decision for speech synthesis in an SPSS framework.

- 2-D based processing of speech spectrogram using Riesz transform to get the smoothed vocal tract envelope. In addition, Riesz transform provides 2-D pitch map, voicing decision, and aperiodicity spectrogram. Finally, modeling these Riesz parameters in SPSS and showing its importance for improving the quality of SPSS.

- Source modeling by different aspects of glottal activity region like epoch location, epoch strength, aperiodic component, and phase information using ILPR. Periodic components are modeled using MCEP and aperiodic representation of ILPR signal in glottal activity region is modeled using white Gaussian noise modulated with the pitch adaptive triangular envelope weighted by SoE. Finally, processing of phase component present in the ILPR using all-pass filter coefficients and showing its significance to SPSS.

- Combining suprasegmental, source, and system features to improve the prosody, naturalness, and intelligibility of SPSS, respectively.

8.3 Directions for future work

Based on the outcome of this thesis work, this section provides some of the possible future directions for research.

(i) Glottal activity region based processing can be used for speaker verification, speaker adaptation, speaking style, and emotions change by modifying the glottal activity parameters. As an initial work, glottal activity region based processing is used for speaker verification task in [170]. Similarly, this work can be extended for other tasks.

(ii) Riesz transform analysis can be extended to individual speech sound analysis by using pitch map, coherence map, and smoothed spectral envelope. Further, an extension of this can be applied to speech recognition and speaker verification application.

(iii) For the high-quality speech, modeling unvoiced sounds is also important. Hence, processing of non-glottal activity region specific features for analysis/synthesis of unvoiced sounds may be helpful for perceptual improvement.

Bibliography

[1] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[2] J. L. Flanagan, *Speech analysis, synthesis and perception*. Springer Science & Business Media, 2013, vol. 3.

[3] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[4] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[5] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101-5, pp. 1234–1252, 2013.

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

[7] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 373–376, 1996.

[8] 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.

[9] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51(11), pp. 1039–1064, 2009.

[10] S.-J. Kim, J.-J. Kim, and M. Hahn, "HMM-based korean speech synthesis system for hand-held devices," *IEEE Trans. Consumer Electronics*, vol. 52-4, pp. 1384–1390, 2006.

[11] J. A. Bilmes *et al.*, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *Int. Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[12] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice hall, 1993.

[13] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 8, pp. 93–96, 1983.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech*, 2001.

[15] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the blizzard challenge 2005," *IEICE Trans. Info. Sys.*, vol. E90-D No.1, pp. 325–333, 2007.

[16] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation." in *Proc. ICSLP*, 1994.

[17] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[18] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Commun.*, vol. 27(3-4), pp. 187–207, 1999.

[19] K. U. Ogbureke, J. P. Cabral, and J. Carson-Berndsen, "Using noisy speech to study the robustness of a continuous F0 modelling method in HMM-based speech synthesis."

[20] P. Cabañas-Molero, D. Martínez-Muñoz, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Rodríguez-Serrano, "Voicing detection based on adaptive aperiodicity thresholding for speech enhancement in non-stationary noise," *IET Signal Process.*, vol. 8, no. 2, pp. 119–130, 2014.

[21] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Info. Sys.*, vol. E90-D, pp. 816–824, 2007.

[22] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224. [Online]. Available: http://festvox.org/cmu_arctic/index.html

[23] R. Maia, T. Toda, H. Zen, Y. Nankaku, and Tokuda.T, "An excitation model for HMM-based speech synthesis based on residual modeling," in *Proc. 6th ISCA Workshop Speech Synth.*, 2007.

[24] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "A trainable excitation model for HMM-based speech synthesis," in *Proc. Interspeech*, 2007.

[25] H. Kawahara, J. Estill, and F. Osamu, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," *Proc. MAVEBA*, 2001.

[26] R. Maia, M. Akamine, and M. J. Gales, "Complex cepstrum for statistical parametric speech synthesis," *Speech Commun.*, vol. 55, no. 5, pp. 606–618, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639313000034

[27] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011.

[28] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2, pp. 109 – 118, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016763939290005R

[29] F. Soong and B. Juang, "Line spectrum pair (*LSP*) and speech data compression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 9, 1984, pp. 37–40.

[30] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Commun.*, vol. 51, no. 5, pp. 401–411, May 2009. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2008.12.005

[31] H. L. Lou, "Implementing the viterbi algorithm," *IEEE Signal Process. Magazine*, vol. 12, no. 5, pp. 42–52, Sep 1995.

[32] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19-1, pp. 153–165, 2011.

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

[34] HTS. [Online]. Available: http://hts.sp.nitech.ac.jp/

[35] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. SSW, Sunnyvale, USA*, 2016.

[36] 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.

[37] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[38] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 3, 2000, pp. 1315–1318.

[39] G. K. Anumanchipalli and A. W. Black, "Adaptation techniques for speech synthesis in under-resourced languages." in *SLTU*, 2010, pp. 51–55.

[40] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[41] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[42] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[43] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 10, pp. 2129–2139, 2013.

[44] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2013, pp. 7962–7966.

[45] T. Koriyama and T. Kobayashi, "Prosody generation using frame-based gaussian process regression and classification for statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 4929–4933.

[46] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 3844–3848.

[47] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Proc. Interspeech*, 2014, pp. 1964–1968.

[48] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Magazine*, vol. 32, no. 3, pp. 35–52, May 2015.

[49] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. IEEE EUSIPCO*, 2014, pp. 2290–2294.

[50] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016.

[51] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort." in *Proc. Interspeech*, 2014, pp. 1969–1973.

[52] A. Suni, T. Raitio, D. Gowda, R. Karhila, M. Gibson, and O. Watts, "The *s*imple4all entry to the *b*lizzard challenge 2014," in *Proc. Blizzard Challenge*, 2014.

[53] S. Tiomkin, D. Malah, S. Shechtman, and Z. Kons, "A hybrid text-to-speech system that combines concatenative and statistical synthesis units," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 5, pp. 1278–1288, July 2011.

[54] Z.-H. Ling and R.-H. Wang, "Minimum unit selection error training for HMM-based unit selection speech synthesis system," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, March 2008, pp. 3949–3952.

[55] M. Plumpe, A. Acero, H. Hon, and X. Huang, "HMM-based smoothing for concatenative speech synthesis," in *Proc. ICSLP*, 1998, pp. 2751–2754.

[56] V. Pollet and A. Breen, "Synthesis by generation and concatenation of multiform segments," in *Proc. Interspeech*, 2008.

[57] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," in *Proc. ISCA SSW5*, 2004, pp. 179–184.

[58] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, Mar 1999, pp. 229–232 vol.1.

[59] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 5, pp. 1071–1079, July 2011.

[60] K. Yu, T. Toda, M. Gasic, S. Keizer, F. Mairesse, B. Thomson, and S. Young, "Probablistic modelling of F0 in unvoiced regions in HMM based speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, April 2009, pp. 3773–3776.

[61] Q. Zhang, F. Soong, Y. Qian, Z. Yan, J. Pan, and Y. Yan, "Improved modeling for F0 generation and V/U decision in HMM-based TTS," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, March 2010, pp. 4606–4609.

[62] S. Kang, Z. Shuang, Q. Duan, Y. Qin, and L. Cai, "Voiced/unvoiced decision algorithm for HMM-based speech synthesis." in *Proc. Interspeech*, 2009, pp. 412–415.

[63] U. Ogbureke, J. Cabral, and J. Berndsen, "Using multilayer perceptron for voicing strength estimation in HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Information Science Signal Process. and their Applications*, July 2012, pp. 683–688.

[64] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[65] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, 1997, pp. 1303–1306.

[66] D. Talkin, "REAPER: Robust epoch and pitch estimator [online] available: https://github.com/google/reaper."

[67] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, 2011, pp. 1973–1976. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2011/i11_1973.html

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

[69] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[70] A. Camacho, "Swipe: A sawtooth waveform inspired pitch estimator for speech and music," Ph.D. dissertation, University of Florida, 2007.

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

[72] D. Arifianto, "Dual parameters for voiced-unvoiced speech signal determination," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 4, April 2007, pp. IV–749–IV–752.

[73] L. Janer, J. J. Bonet, and E. Lleida-Solano, "Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms," in *Proc. IEEE Int. conf. Spoken lang. process.*, 1996, pp. 1209–1212.

[74] N. Narendra and K. S. Rao, "Robust voicing detection and F0 estimation for HMM-based speech synthesis," *Circuits Syst. Signal Process.*, pp. 1–23, 2015.

[75] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24, no. 3, pp. 201–212, Jun 1976.

[76] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 4, pp. 697–710, April 2013.

[77] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[78] C. Ma, Y. Kamp, and L. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Commun.*, vol. 12, no. 1, pp. 69 – 81, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016763939390019H

[79] A. Roebel and X. Rodet, "Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation," in *Proc. Int. Conf. Digital Audio Effects*, Madrid, Spain, Sep. 2005, pp. 30–35, cote interne IRCAM: Roebel05b. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01161334

[80] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Am.*, vol. 45, no. 2, pp. 458–465, 1969.

[81] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 137–140, 1992.

[82] H. KAWAHARA and M. MORISE, "Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework," *Sadhana*, vol. 36, no. 5, pp. 713–727, 2011. [Online]. Available: http://dx.doi.org/10.1007/s12046-011-0043-3

[83] T. Drugman, A. Moinet, T. Dutoit, and G. Wilfart, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, 2009.

[84] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, 2009.

[85] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *Proc. 6th ISCA Workshop on Speech Synthesis*, 2007.

[86] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7830–7834.

[87] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal Spectral Separation for Parametric Speech Synthesis," in *Proc. Interspeech*, 2008.

[88] T. Raitio, L. Juvela, A. Suni, M. Vainio, and P. Alku, "Phase perception of the glottal excitation and its relevance in statistical parametric speech synthesis," *Speech Commun.*, pp. –, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639316000303

[89] O. Abdel-Hamid, S. M. Abdou, and M. Rashwan, "Improving arabic HMM based speech synthesis quality," in *Proc. Interspeech*, 2006.

[90] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, no. 4, pp. 744–754, 1986.

[91] I. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, Ecole Nationale Supérieure des Télécommunications, 1996.

[92] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, 2001.

[93] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP Journal on Audio Speech Music Process.*, no. 1, pp. 1–16, 2014. [Online]. Available: http://dx.doi.org/10.1186/s13636-014-0038-1

[94] B. Eleftherios, E. Daniel, B. Antonio, and M. Asuncion, "Flexible harmonic/stochastic modeling for HMM-based speech synthesis," in *V Jornadas en Tecnologa del Habla*, 2008.

[95] C. Hemptinne, "Integration of the harmonic plus noise model (HNM) into the Hidden Markov Model-Based speech synthesis system (HTS)." Master's thesis, Idiap Research Institute, 2006.

[96] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Process.*, vol. 8, no. 2, pp. 184–194, April 2014.

[97] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, March 2008, pp. 4609–4612.

[98] Y. Ijima, T. Asami, and H. Mizuno, "Objective Evaluation Using Association Between Dimensions Within Spectral Features for Statistical Parametric Speech Synthesis," 2016, pp. 337–341.

[99] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. conf. Acoust. Speech Signal Process.*, vol. 2. IEEE, 2001, pp. 749–752.

[100] J. Nurminen, A. Heikkinen, and J. Saarinen, "Objective evaluation of methods for quantization of variable-dimension spectral vectors in WI speech coding," in *Proc. Eurospeech*, 2001.

[101] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Commun.*, vol. 10-5, pp. 819–829, 1992.

[102] J. Vepa and S. King, "Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 5, pp. 1763–1771, Sept 2006.

[103] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations," in *Proc. Interspeech*, 2015.

[104] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," 2008.

[105] C. Benot, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Commun.*, vol. 18, no. 4, pp. 381 – 392, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016763939600026X

[106] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. European Congress on Acoust.*, 2005, pp. 2725–2728.

[107] A. van den oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *ArXiv e-prints*, Sep. 2016.

[108] K. S. R. Murthy, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal process. lett.*, vol. 16, no. 6, pp. 469–472, June 2009.

[109] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 217–231, Mar 2001.

[110] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, pp. 1602–1613, November 2008.

[111] A. Prathosh, T. Ananthapadmanabha, and A. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 12, pp. 2471–2480, Dec 2013.

[112] P. FFabre, "Un procede electrique percutane dinscrition de laccolement glottique au cours de la phonation: glottographie de haute frequence. premiers resultats," *Bull. Acad. Natl. Med*, vol. 141, p. 66, 1957.

[113] E. R. Abberton, D. M. Howard, and A. J. Fourcin, "Laryngographic assessment of normal voice: a tutorial," *Clinical Linguistics & Phonetics*, vol. 3, no. 3, pp. 281–296, 1989.

[114] A. Krishnamurthy and D. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, no. 4, pp. 730–743, Aug 1986.

[115] T. V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics. STL-QPSR 23," Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Tech. Rep., 1984.

[116] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3(5), pp. 325–333, 1995.

[117] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[118] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multi-pitch tracking scenario," 2011, pp. 1509–1512.

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

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

[121] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. Interspeech*, 2000, pp. 464–467.

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

[123] M. Rothenberg, "Glottal noise during speech," *Quarterly Progress Status ReportSpeech Transmission Laboratory of the Royal Institute of Technology, Stockholm*, p. 1, 1974.

[124] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Commun.*, pp. –, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639316000364

[125] N. Adiga and S. R. M. Prasanna, "Detection of glottal activity using different attributes of source information," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2107–2111, Nov 2015.

[126] K. Fukunaga and P. M. Narendra, "A branch and bound algorithm for computing k-nearest neighbors," *IEEE Trans. Computers*, vol. 100, no. 7, pp. 750–753, 1975.

[127] C. Cortes, M. Mohri, and A. Rostamizadeh, "Two-stage learning kernel algorithms," in *Proc. of the 27th Int. Conf. on Machine Learning*, 2010, pp. 239–246.

[128] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Sys. Tech.*, vol. 2, no. 3, p. 27, 2011.

[129] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[130] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Commun.*, vol. 53, no. 2, pp. 154–174, 2011.

[131] B. K. Khonglah and S. R. M. Prasanna, "Speech / music classification using speech-specific features," *Digital Signal Process.*, vol. 48, pp. 71 – 83, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200415002730

[132] F. Plante, G. Meyer, and W. Ainsworth, "A pitch extraction reference database," *Children*, vol. 8, no. 12, pp. 30–50, 1995.

[133] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching." 1993.

[134] S. Shamma, "On the role of space and time in auditory processing," *Trends in cognitive sciences*, vol. 5, no. 8, pp. 340–348, 2001.

[135] T. Wang and T. Quatieri, "Two-dimensional speech-signal modeling," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 6, pp. 1843–1856, Aug 2012.

[136] T. F. Quatieri, "2-D processing of speech with application to pitch estimation." in *Proc. Interspeech*, 2002.

[137] K. G. Larkin, D. J. Bone, and M. A. Oldfield, "Natural demodulation of two-dimensional fringe patterns. I. General background of the spiral phase quadrature transform," *J. Opt. Soc. Amer. A.*, vol. 18, no. 8, pp. 1862–1870, 2001.

[138] C. S. Seelamantula, N. Pavillon, C. Depeursinge, and M. Unser, "Local demodulation of holograms using the Riesz transform with application to microscopy," *J. Opt. Soc. Am. A*, vol. 29, no. 10, pp. 2118–2129, Oct 2012. [Online]. Available: http://josaa.osa.org/abstract.cfm?URI=josaa-29-10-2118

[139] D. Gabor, "Theory of communication. part 1: The analysis of information," *J. Inst. Elec. Eng.-Part III Radio and Commun. Eng.*, vol. 93, no. 26, pp. 429–441, 1946.

[140] E. Bedrosian, "A product theorem for Hilbert transforms," *Proc. of the IEEE*, vol. 51, no. 5, pp. 868–869, May 1963.

[141] A. Nuttall and E. Bedrosian, "On the quadrature approximation to the hilbert transform of modulated signals," *Proc. IEEE*, vol. 54, no. 10, pp. 1458–1459, 1966.

[142] M. Unser and D. Van De Ville, "Wavelet steerability and the higher-order Riesz transform," *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 636–652, 2010.

[143] J. P. Havlicek, D. S. Harding, and A. C. Bovik, "Multidimensional quasi-eigenfunction approximations and multicomponent AM-FM models," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 227–242, 2000.

[144] S. Arseneau and J. R. Cooperstock, "An improved representation of junctions through asymmetric tensor diffusion," in *International Symposium on Visual Computing*. Springer, 2006, pp. 363–372.

[145] H. Aragonda and C. Seelamantula, "Demodulation of narrowband speech spectrograms using the Riesz transform," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 11, pp. 1824–1834, Nov 2015.

[146] H. A. Patil, T. B. Patel, N. J. Shah, H. B. Sailor, R. Krishnan, G. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra *et al.*, "A syllable-based framework for unit selection synthesis in 13 indian languages," in *Proc. Oriental COCOSDA*. IEEE, 2013, pp. 1–8.

[147] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, May 2006, pp. I–I.

[148] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, 1986, pp. 93–99.

[149] H. Kawahara, 2010. [Online]. Available: http://www.wakayama-u.ac.jp/~kawahara/puzzlet/STRAIGHTtipse/'

[150] T. T. Wang and T. F. Quatieri, "Towards co-channel speaker separation by 2-D demodulation of spectrograms," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 65–68.

[151] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text to speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer, Speech And Language*, vol. 19, pp. 55–83, 2005.

[152] J. Nurminen, H. Silen, E. Helander, and M. Gabbouj, "Evaluation of detailed modeling of the LP residual in statistical speech synthesis," in *IEEE Int. Symposium Circuits and Systems*, 2013.

[153] K. K. Paliwal and L. D. Alsteris, "Usefulness of phase spectrum in human speech perception." in *Proc. Interspeech*, 2003.

[154] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, 2001, pp. 133–136.

[155] R. Maia, M. Akamine, and M. J. F. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, March 2012, pp. 4581–4584.

[156] N. Adiga and S. R. M. Prasanna, "Significance of instants of significant excitation for source modeling," in *Proc. Interspeech*, 2013.

[157] N. Adiga and S. M. Prasanna, "Epochs based compression of lp residual for source modeling in text-to-speech synthesis," in *Proc. NCC*. IEEE, 2014, pp. 1–5.

[158] R. Crochiere, "A weighted overlap-add method of short time fourier analysis/synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 1, p. 99102, February 1980.

[159] A.Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, pp. 1541–1582, 1994.

[160] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Communication*, vol. 16(3), pp. 225 – 244, 1995.

[161] L. Saheer, J. Dines, and P. N. Garner, "Vocal tract length normalization for statistical parametric speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 7, pp. 2134–2148, Sept 2012.

[162] P. J. Murphy, "Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2866–2881, 1999.

[163] D. O'shaughnessy, *Speech communication: human and machine*. Universities press, 1987.

[164] K. Vijayan, V. Kumar, and K. S. R. Murty, "Allpass modelling of Fourier phase for speaker verification," in *Proc. Odyssey*, 2014, pp. 112–117.

[165] C.-Y. Chi and J.-Y. Kung, "A new identification algorithm for allpass systems by higher-order statistics," *Signal Process.*, vol. 41, no. 2, pp. 239–256, 1995.

[166] N. Adiga and S. R. M. Prasanna, "Source modeling for HMM based speech synthesis using integrated LP residual," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016.

[167] N. Adiga and S. M. Prasanna, "A hybrid Text-to-Speech synthesis using vowel and non vowel like regions," in *Proc. INDICON*. IEEE, 2014, pp. 1–5.

[168] N. Lass, *Contemporary issues in experimental phonetics*. Elsevier, 2012.

[169] B. Sharma, N. Adiga, and S. M. Prasanna, "Development of Assamese Text-to-speech synthesis system," in *Proc. TENCON*. IEEE, 2015, pp. 1–6.

[170] A. Pandey, R. K. Das, N. Adiga, N. Gupta, and S. M. Prasanna, "Significance of glottal activity detection for speaker verification in degraded and limited data condition," in *Proc. TENCON*. IEEE, 2015, pp. 1–6.

List of Publications

Journal Publications

1. **N.Adiga** and S. R. M. Prasanna, "Characterization of Glottal Activity using Different Attributes of Source Information", *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 2107-2111, Nov 2015.

2. **N. Adiga**, B. K. Khonglah and S. R. M. Prasanna, "Improved voicing decision using glottal activity features for HMM based speech synthesis" under first revision in *Journal of Digital Signal Processing*.

3. **N.Adiga** and S. R. M. Prasanna, "Acoustic Features for Statistical Parametric Speech Synthesis: A Review" under first revision in *IETE Technical review*.

Conference and Workshop Publications

1. **N. Adiga** and S. R. M. Prasanna, "Significance of instants of significant excitation for source modeling," in Proc. INTERSPEECH, 2013.

2. **N. Adiga** and S. R. M. Prasanna, "Epochs Based Compression of LP Residual for Source Modeling in Text-to-Speech Synthesis", in Proc NCC, 2014.

3. **N. Adiga** and S. R. M. Prasanna, "A Hybrid Text-to-Speech Synthesis using Vowel and Non Vowel like regions", in Proc. INDICON, 2014.

4. **N. Adiga**, D.Govind and S. R. M. Prasanna, "Significance of Epoch Identification Accuracy for Prosody Modification", in Proc. SPCOM, 2014.

5. **N. Adiga** and S. R. M. Prasanna, "Source modeling for HMM based speech synthesis using Integrated LP Residual", in Proc. ICASSP, 2016.

6. **N. Adiga** and S. R. M. Prasanna, "Phase modeling using Integrated LP residual for Statistical Parametric Speech Synthesis" communicated to Interspeech 2017

Other Publications

1. Hemant A Patil, Tanvina B Patel, Nirmesh J Shah, Hardik B Sailor, Raghava Krishnan, GR Kasthuri, T Nagarajan, Lilly Christina, Naresh Kumar, Veera Raghavendra, SP Kishore, SRM Prasanna, **N. Adiga**, Sanasam Ranbir Singh, Konjengbam Anand, Pranaw Kumar, Bira Chandra Singh, SL Binil Kumar, TG Bhadran, T Sajini, Arup Saha, Tulika Basu, K Sreenivasa Rao, NP Narendra, Anil Kumar Sao, Rakesh Kumar, Pranhari Talukdar, Purnendu Acharyaa, Somnath Chandra, Swaran Lata, Hema A Murthy, "A syllable-based framework for unit selection synthesis in 13 Indian languages," *in Proc. Oriental COCOSDA*, 2013

2. Deepak K T, Ramesh K, **N. Adiga** and S. R. M. Prasanna, "Speech and EGG Polarity Detection using Hilbert Envelope," *in Proc. TENCON*, 2015.

3. Bidisha Sharma, **N. Adiga** and S. R. M. Prasanna, "Development of Assamese Text-to-Speech Synthesis System ," *in Proc. TENCON*, 2015.

4. Ashutosh Pandey, Rohan Kumar Das, **N. Adiga**, Naresh Gupta and S. R. M. Prasanna, "Significance of Glottal Activity Detection for Speaker Verification in Degraded and Limited Data Condition," *in Proc. TENCON*, 2015.

5. Vikram C.M, **N. Adiga**, and S. R. M. Prasanna, "Spectral Enhancement of Cleft Lip and Palate Speech," *in Proc. INTERSPEECH*, 2016.

6. Priyankoo Sarmah, Biswajit Dev Sarma, **N. Adiga**, and S. R. M. Prasanna, "Dual channel signal analysis for nasal and oral consonants," *in Proc. TENCON*, 2016.

7. Bidisha Sharma, **N. Adiga** and S. R. M. Prasanna, "Dynamic Post-filtering using Source and Spectral Features for Statistical Parametric Speech Synthesis," **under review in** IEEE Trans. Audio Speech Lang. Process.

8. **N. Adiga**, Vikram C M, Keerthi Pullela and S. R. M. Prasanna, "Zero Frequency Filter Based Analysis of Voice Disorders," communicated to Interspeech 2017.