# IMPROVING QUALITY OF STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING SONORITY INFORMATION




*BIDISHA SHARMA*

# IMPROVING QUALITY OF STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING SONORITY INFORMATION

*A*

*Thesis submitted*

*for the award of the degree of*

**DOCTOR OF PHILOSOPHY**

By

**BIDISHA SHARMA**

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

March 2018

# Certificate

This is to certify that the thesis entitled "**IMPROVING QUALITY OF STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING SONORITY INFORMATION**", submitted by **Bidisha Sharma** (136102017), 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 her 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

**Biren Sarma** and **Mira Devi**

for their love, encouragement and sacrifice

&

My brother

**Surajit Sharma**

for being with me whenever I needed

# Acknowledgements

This thesis would not have been possible without the constant support and motivation from my research supervisor Prof. S.R.M. Prasanna. His insightful comments and regular discussions immensely helped me to carry out the work. His discipline of work and motivational words can inspire anybody a lot. I am wholeheartedly grateful to him for providing enormous facilities and creating an excellent atmosphere for doing research in the EMST and signal informatics lab.

I am thankful to my doctoral committee members Prof. Samarendra Dandapat, Prof. Rohit Sinha and Dr. Priyankoo Sarmah for their encouragement and valuable suggestions on my work. Their constructive criticism during my seminars helped me to improve the work. My frequent interaction with Dr. Priyankoo Sarmah sir helped me with his suggestions. 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.

During the period I worked in project, I had opportunity in the review meetings and workshops to talk to Prof. Hema A Murthy from IIT Madras, who was the consortium leader of the project. Her enthusiastic speech really encouraged me towards doing research and provided us opportunity of collaborative learning. Further I would like to thank the funding agency of the project Department of Information Technology (DeitY), New Delhi for providing funding for attending workshops and the computational facilities in the lab. I would like to convey my gratitude to Dr. Sansam Ranbir Singh from CSE branch for his support, constructive technical suggestions and ideas. I sincerely thank Dr. L.N. Sharma for his smooth maintenance and timely response whenever we face any issue. I would also convey my gratitude to the Head of the Department, EEE, Prof Chitralekha Mahanta and the Office Staffs for timely and flawless processing of different applications.

My sincere thanks goes to my seniors Dr. Deepak K.T., Dr. Biswajit Dev Sarma, Dr. Nagaraj Adiga, Banriskhem K. Khonglah for various technical discussions and suggestions, they provided whenever I needed. I convey my gratitude to Dr. Rohan Kumar Das for carefully correcting my papers, thesis and inspiring me for doing research. I would also like to thank my seniors Dr. Govind D, Dr. S. Shahnawazuddin, Ramesh sir, Dr. Sunil Y for their help at different times. In my daily work I have been blessed with a friendly and cheerful group of lab mates. I would never forget my close friends Himakshi and Protima, for being so much patient to listen to me whenever I needed

them. Starting from the frustrated moments till the party time they constantly stood by me. I am fortunate to have friends Suman, Subhasish, Nagendra, Sreeram who helped me at different times. The constructive technical discussions with Sishir, Vikram and Akhilesh helped me in my work. I convey my thank to them. A thankful note to past/present members of the lab Anurag, Jiss, Padhy, Biju, Mawsumi, Bhukya, Parishmita, Abhishek, Sarfaraz, Tilendra, Sikha, Moa, Mrinmoy, Saswati, Sandeep, Sukanya, Prabhakar, Vineeta, Alex, Ato and the rest for their direct/indirect contributions during my stay at IITG. I wholeheartedly thank all of them for taking part in the tedious listening tests, that extremely helped me to evaluate my work. Also a heartful thank to my past and present project group Indrajit, Hridoy, Madhurya, Gyanendero, Rajlakshmi, Nanaobi, Deepshikha, Irani, Anupama for their helping hand in smoothly carrying out the project.

During my PhD I attended Interspeech conference abroad with funding received from International Speech Communication Association (ISCA) and Science & Engineering Research Board (SERB), Govt. of India, for which I will be thankful forever. I convey my sincere thank to MHRD, Govt. of India for providing fellowship for my PhD thesis work.

Without the blessings, support and love from my grandmother and my parents this work would not have been possible, for which I will be indebted to them forever. Finally, I would like to thank my close friends for their care and love that made my days beautiful.

*Bidisha Sharma*

# Abstract

This thesis aims towards improving naturalness and intelligibility of synthesized speech obtained from statistical parametric speech synthesis (SPSS). Along with the conventional source and spectral information, some additional significant features can also be derived from the speech signal to preserve its characteristics in parametric form. The sonority information represents spectral prominence, higher energy and periodicity aspects, which are related to human speech perception, that change with the varying vocal-tract constriction and glottal source amplitude during speech production. Therefore, this information is extracted from the speech signal in terms of sonority feature. It is capable to delineate the degree of sonority associated with a sound unit. The sonority feature is incorporated in the SPSS framework to use it in the studies related to this thesis.

To alleviate the over-smoothing effect from parameter sequences generated from SPSS, post-filtering mechanisms are found to be effective. By considering the fact that the characteristics of the speech parameters may extensively vary based on the broad categories of sound units, a class based dynamic post-filtering method is proposed. The excitation source (fundamental frequency and strength of excitation (SoE)) and spectral parameters (sharpness of peaks and valleys of the spectrum) corresponding to each frame are enhanced using post-filtering factors that change with sonorant sound categories. The sonorant class information is derived from a support vector machine based classifier trained using sonority feature associated with each frame. This method improves the temporal variation, fine spectral structure as well as reduces the deviation with the natural counterpart leading to improvement in synthesized speech quality.

Spectral slope is another aspect that influences on perception of synthesized speech. From the analysis of natural and synthesized speech, it is observed that the spectral slope of synthesized speech is more negative compared to that of the natural. Therefore, a novel method is proposed to modify the spectral slope of synthesized speech that reduces the de-

viation in spectral tilt between natural and synthesized speech. The enhanced synthesized speech with modified flatter spectrum sounds clearer. It shows improvement in terms of naturalness, intelligibility and speaker similarity.

The voicing decision plays a significant role in excitation source generation module of SPSS. Along with the excitation source feature, the spectral prominence aspect in the sonority information makes it useful for voicing decision. The sonority feature is employed to develop a voiced/unvoiced classifier, that improves the naturalness compared to the existing methods of voicing decision. A common framework is developed that models the sonority feature and incorporates it in post-filtering and voicing decision followed by spectral tilt modification. The application of individual modules and their combination brings significant improvement to quality of synthesized speech obtained from SPSS.

The major contributions of this thesis are as follows:

- Proposing an efficient feature set from system, source and suprasegmental aspects of the speech signal having capability to represent degree of sonority.
- As sonority is associated with production aspects of sound units, its subsequent application in improved phoneme recognition and vowel onset point detection.
- A sonorant class based dynamic post-filtering method is proposed to reduce over-smoothing and the deviation between natural and synthesized speech parameters.
- To minimize the difference in spectral tilt of natural and synthesized speech, a spectral tilt modification method is proposed that leads to better quality of synthesized speech.
- Use of sonority feature in voicing decision during excitation source generation in the SPSS framework.
- A combined framework is proposed that models the sonority feature and further employs it in post-filtering of spectral peaks and valleys, $F_0$; voicing decision and spectral tilt modification at the same time.

**Keywords:** Statistical parametric speech synthesis, sonority, sonority hierarchy, source, spectral, suprasegmental, post-filtering, spectral tilt, speech enhancement, voicing decision.

# Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| ASR | Automatic Speech Recognition |
| BAP | Band Aperiodicity Parameter |
| BW | Baum-Welch |
| CCA | Canonical Correlation Analysis |
| CDF | Cumulative Distribution Function |
| DCT | Discrete Cosine Transform |
| DEGG | Differenced Electro-Glotto-Graph |
| DFT | Discrete Fourier Transform |
| DNGD | Differentiating Numerator Group Delay Two Times |
| DNN | Deep Neural Network |
| DRF | Dominant Resonance Frequency |
| DTW | Dynamic Time Warping |
| EM | Expectation Maximization |
| FAR | False Alarm Rate |
| FOGD | First Order Gaussian Differentiator |
| FFT | Fast Fourier Transform |
| GCI | Glottal Closure Instant |
| GD | Group Delay |
| GMM | Gaussian Mixture Models |
| GV | Global Variance |
| HE | Hilbert Envelope |
| HMMs | Hidden Markov Models |
| HNGD | Hilbert Envelope of Numerator of Group Delay |
| HNM | Harmonic Plus Noise Model |

| | |
|---|---|
| HPF | High Pass Filter |
| HSMM | Hidden Semi-Markov Model |
| HTS | HMM based Speech Synthesis System |
| IFT | Inverse Fourier Transformation |
| KLD | Kullback Leibler Divergence |
| LF | Liljencrants-Fant |
| LP | Linear Prediction |
| LPC | Linear Predictive Coefficients |
| LPCC | Linear Predictive Cepstral Coefficients |
| LQO | Listening Quality Opinion |
| LSD | Log Spectral Distance |
| LSP | Line Spectral Pair |
| MAP | Maximum A Posteriori |
| MFCC | Mel-Frequency Cepstral Coefficients |
| MGC | Mel-Generalized Cepstral |
| MGE | Minimum Generation Error |
| ML | Maximum Likelihood |
| MLLR | Maximum Likelihood Linear Regression |
| MLSA | Mel-log Spectrum Approximation |
| MOS | Mean Opinion Score |
| MS | Modulation Spectrum |
| MSD | Multi Space Distribution |
| NGD | Numerator of Group Delay Function |
| PDF | Probability Density Function |
| PER | Phone Error Rate |
| PESQ | Perceptual Evaluation of Speech Quality |
| PF | Post-Filtering |
| PVR | Peak-to-Valley Ratio |
| RAPT | Robust Algorithm for Pitch Tracking |
| RBF | Radial Basis Function |

| | |
|---|---|
| REAPER | Robust Epoch And Pitch EstimatoR |
| SoE | Strength of Excitation |
| SNR | Signal to Noise Ratio |
| SRH | Summation of Residual Harmonics |
| STFT | Short Term Fourier Transform |
| STRAIGHT | Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum |
| SVM | Support Vector Machine |
| TM | Tilt Modification |
| TPR | True Positive Rate |
| TTS | Text-to-Speech |
| USS | Unit Selection Synthesis |
| VOP | Vowel Onset Point |
| VTS | Vocal-Tract Spectrum |
| ZFF | Zero Frequency Filtering |
| ZFFS | Zero Frequency Filter Signal |
| ZFR | Zero Frequency Resonator |
| ZTW | Zero Time Windowing |

# 1

# Introduction



Contents

### Objective

*The SPSS has been widely used due to its advantages like flexibility in terms of adaptation of speakers, emotions, lower footprint and robustness. However, the muffled quality of synthesized speech remains as one of the major drawbacks of SPSS. This can be often attributed as the outcome of poor representation of speech signal using limited number of features, over-smoothed generated feature contour, poor representation of excitation source and simple vocoder. The sonority information is associated with formant prominence, excitation source strength and periodicity aspects of speech signal that exhibit significance in terms of speech perception. The objective of this thesis is to overcome some of the limitations associated with synthesized speech obtained from SPSS, by exploiting the usefulness of sonority information. The sonority feature extracted from the speech signal has the capability to correctly represent the degree of sonority associated with different sound units. Along with the conventional source and spectral features, the sonority feature can efficiently represent the speech signal in parametric form. With the analysis of deviation of different features between natural and synthesized speech, it can be hypothesized that the feature contours generated from statistical models can be modified to overcome over-smoothing effect and make it closer to that of the natural speech. This leads to the idea of sonorant class-based post-filtering (PF) method, that increases the dynamic range of feature contours. The spectral tilt associated with the speech signal is also found to have impact on its perception, which is analyzed and modified to improve the quality of synthesized speech. The voicing decision corresponding to each frame plays a pivotal role in excitation source generation module. The exploration of sonority feature in improved voicing decision further contributes to the enhancement of synthesized speech obtained from SPSS.*

## 1.1 Overview of text-to-speech synthesis

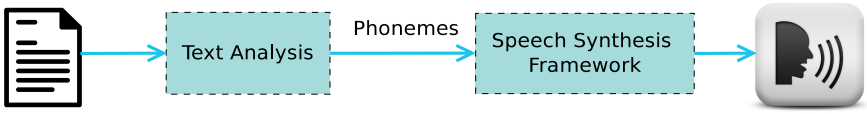


**Figure 1.1:** Block diagram representing text-to-speech synthesis system.

A text-to-speech (TTS) system aims to convert a discrete sequence of text input to corresponding continuous speech signal. With the increase in contribution of technology in day-to-day life of human

being, a natural sounding compact TTS synthesis system becomes of paramount significance. As shown in Figure 1.1, the text to be synthesized is first passed through a text analysis module to derive corresponding language specific phonemes. The phoneme sequence is then given to the synthesiser which produces the synthetic speech. Provided the front end remains same, the synthesis module can be based on different algorithms. The state-of-the-art approaches to develop TTS synthesis system are concatenative speech synthesis using unit selection algorithm and statistical generative model based approach. The concatenative speech synthesis joins speech segments corresponding to different units based on the text labels and cost function. In contrast to that the SPSS approach generates source and spectral parameters from some generative models corresponding to text labels developed in the training stage. The parameters are further used in a vocoder to restore the speech signal. Due to the growing demand of using low resource, low cost, flexibility to adapt different speakers, different speaking styles and emotions [1,2], SPSS has grown in popularity in the last decade over other speech synthesis techniques like unit selection synthesis (USS) [3–5].

## 1.2 Introduction to statistical parametric speech synthesis

In SPSS, instead of storing the speech segments directly, the average characteristics of excitation source and spectral features of similar speech segments are retained by using statistical generative models. In this case, although different generative models can be used, hidden Markov models (HMMs) are used predominantly. The overall architecture of HMM based SPSS is depicted in Figure 1.2, which can be divided into training and testing phases. In the training phase, the acoustic parameters, log fundamental frequency ($\log F_0$) representing excitation source and mel-generalized cepstral (MGC) coefficients representing spectral parameters are extracted from the speech database, with reference to corresponding phoneme labels. Context dependent phoneme HMMs are trained with these parameters using the following maximum likelihood criterion.

$$\hat{\boldsymbol{\lambda}} = arg\ \max_{\boldsymbol{\lambda}}\{P(\boldsymbol{O}|\mathcal{W},\boldsymbol{\lambda})\}, \tag{1.1}$$

where, $\boldsymbol{\lambda}$ is a set of model parameters, $\boldsymbol{O}$ is the set of training parameters and $\mathcal{W}$ is the set of word sequences corresponding to $\boldsymbol{O}$. $\hat{\boldsymbol{\lambda}}$ is the estimated model. During testing, given the text to be synthesized, from the knowledge of corresponding phoneme sequence, the HMMs are retrieved and concatenated to form the sentence HMM. Then, the speech parameter generation algorithm generates

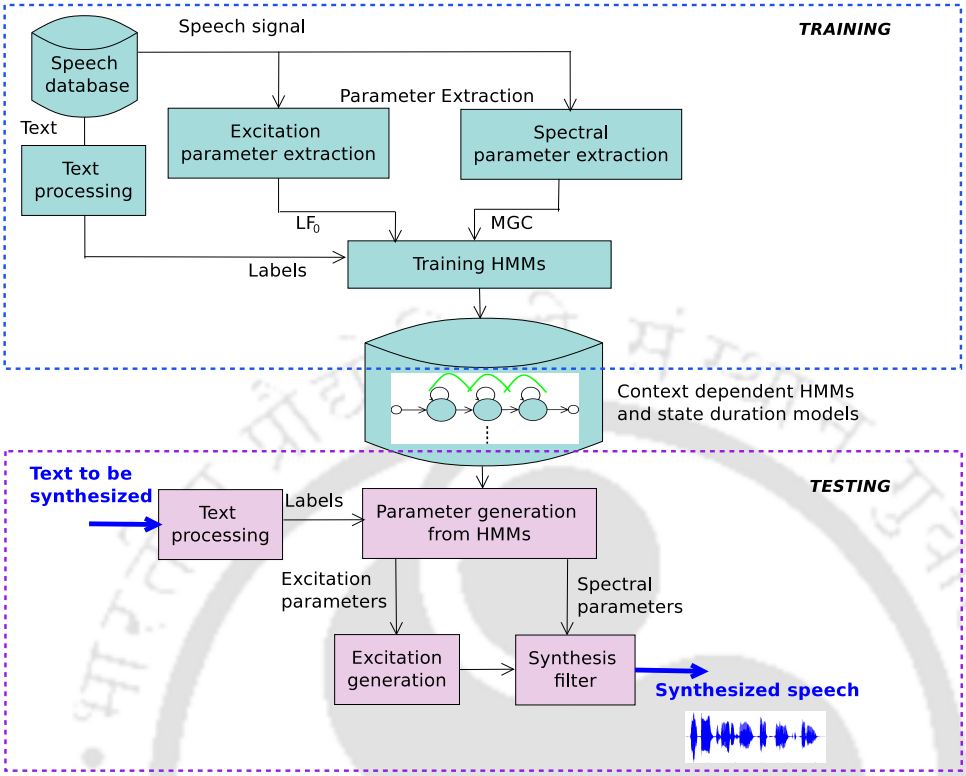


**Figure 1.2:** Block diagram representing HMM based SPSS framework.

the corresponding spectral and excitation parameters which is further passed through a vocoder to derive the synthesized speech.

## 1.3 Issues in statistical parametric speech synthesis

The wide range of applications and supremacy of SPSS demand further development in terms of perceptual quality by alleviating the difference between natural and synthesized speech. The factors that may affect naturalness and intelligibility of synthesized speech may arise from different modules. The possibilities include inefficient statistical modeling technique, parameters used to represent the speech signal, over-smoothening of generated parameters, poor excitation source generation method and simple vocoder structure, each contributing quite significantly to the synthesized speech. Several approaches have been proposed in the literature that aim to improve different modules. This work attempts to address the following concerns in detail and includes explorations to bring out novel approaches to overcome these.

### 1.3.1 Poor representation of speech signal

As shown in the block diagram 1.2, in the training process of SPSS the first module deals with parameter extraction from speech signal, which are basically excitation source and spectral parameters. The source parameters used in conventional SPSS are fundemental frequency ($F_0$), its delta and delta-delta and spectral parameters include Mel generalized cepstral coefficients (MGCs), its delta and delta-delta. After the parameter extraction the entire acoustic information in the speech database is represented in terms of these parameters with respect to corresponding label sequence obtained from the text analysis module. Only $F_0$ and MGCs may not be able to capture all information related to perceptual quality of speech signal. Therefore, incorporation of efficient additional features related to naturalness, intelligibility or perceptual quality of speech signal may be helpful to preserve rich acoustic information of speech signal in parametric form. The additional features may be modeled using HMMs and further employed in different modules in synchronous to conventional source and spectral features to improve the perceptual quality of synthesized speech.

### 1.3.2 Over-smoothing of generated parameter sequence

The widely used generative model in SPSS framework is HMM. As mentioned earlier, based on the context dependent label sequence, the parameter generation algorithm generates static features by maximizing the likelihood of a given HMM for static and dynamic features under an explicit constraint between those two features as given in (1.2).

$$\hat{\boldsymbol{c}} = arg\ \max_{\boldsymbol{c}}\ P(\boldsymbol{W}\boldsymbol{c}|\boldsymbol{\lambda}), \tag{1.2}$$

where, $\boldsymbol{c} = [\boldsymbol{c_1^T}, ..., \boldsymbol{c_t^T}, ..., \boldsymbol{c_T^T}]^T$ is speech parameter vector sequence of $T$ frames. $\boldsymbol{c_t}$ is a $D$-dimensional feature vector of corresponding to frame $t$, $\boldsymbol{W}$ is a weighting matrix to calculate dynamic features and $\boldsymbol{\lambda}$ is the HMM set [6]. The inclusion of dynamic features as well as statistical processing result in over-smoothed parameter contours that are inefficient in preserving the fine temporal and spectral structure intact. This reduction in variation in parameter sequence leads to reduction in naturalness of synthesized speech. Apart from this, within each frame the spectral prominence of formant structure is also smoothed out resulting in less intelligibility of generated speech. The effort to preserve the fine spectral and temporal structure in terms of the source and spectral parameters may help better quality speech.

### 1.3.3 Inadequacy in generation of excitation source

The naturalness of synthesized speech is mostly governed by the excitation source component. In the SPSS framework, along with other factors voiced/unvoiced decision plays a key role in excitation source generation module. Generally, $F_0$ information is employed to persuade voiced/unvoiced decisions for frames in the utterance to be synthesized. Improving the voicing decision may absolutely improve the excitation source and synthesis quality. In conventional SPSS or HMM based speech synthesis, MGC coefficients are used to represent vocal-tract spectrum (VTS) information and $F_0$ is used to model the excitation source information. The $F_0$ in HMM is modeled along with voicing decision using multi space distribution (MSD) and consequently error in $F_0$ estimation is propagated to the excitation source generation module [7,8]. In the voiced regions $F_0$ is modeled as continuous Gaussian distribution and discrete symbol in unvoiced regions [8]. As the voicing decision is dependent on $F_0$, errors in $F_0$ tracking leads to error in MSD-HMM model. This leads to detection of false voiced and unvoiced frames. Specifically the misclassification of voiced regions as unvoiced seems to be more in case of weekly voiced sounds with lower energy. For these regions instead of impulse sequence random noise is used as excitation source. Repeated wrong classification of voiced frames, especially when a sequence of frames are misclassified leads to poor intelligibility. On the other hand, for false voiced regions impulse sequence is used as excitation instead of random noise resulting in buzziness in the synthesized speech.

## 1.4 Significance of sonority information

Sonority refers to relative loudness of sound units resulting from higher energy and periodicity. Sonorants form the prominent regions in speech signal produced with less vocal-tract constriction and glottal vibration [9]. This results in regions of regular structure having high energy and high degree of periodicity. The sonorant regions are therefore prominent ones in the speech signal and important for many speech processing tasks [10]. Vowels are the most sonorous sounds, which mostly form the nucleus of a syllable. The term sonority hierarchy can be defined as the degree of change in sonority with respect to different sound units. Various sonority hierarchies are defined in the literature as mentioned in [9]. However, the most commonly referred sonority hierarchy for the six major classes of sonorants in the decreasing order of sonority is *low-vowels, mid-vowels, high-vowels, glides, liquids and nasals*. In [11], the sonority hierarchy for obstruents is defined in the decreasing order of sonority as

*voiced fricatives, voiced affricates, voiced stops, voiceless fricatives, voiceless affricates, and voiceless stops*. Sonority is used to explain both the perception of syllables and their phonetic structure [12]. The *sonority sequencing principle* states that in every syllable, the syllable nucleus has the highest sonority value [13]. According to *syllable contact law*, the junction between two syllables is well recognizable when coda of the present syllable has higher sonority value than the onset of the next syllable [14]. According to [15], the syllables with nuclei having more sonority value tend to have more stress compared to the syllables with nuclei having less sonority value. For example, syllables with [e] or [o] may be perceived as having more stress than those with [i] and [u]. The possible sequence of consonants present in the syllable onset and coda also depends on the sonority value associated with them. For example, consonant clusters present in syllable onset of the form [pl], [dr], [km] are very common, but the reverse order is rare. In this case, [l], [r], [m] are more sonorous than [p], [d], [k]. Similarly, [mp] and [nd] are very common as syllable codas than [pm], [dn], where [m], [n] are more sonorous than [p], [d]. Therefore, sonority of a sound unit has an impact on the basic production pattern of speech sounds. In several studies of phonology such as consonant cluster, sonorant-obstruent cluster, syllable onset and coda position, degree of sonority is used [16, 17]. The *Degree of sonority* can be defined as sequential variation in various attributes that correlate to sonority, with respect to distinctive category of sound units. The variation in degree of sonority associated with different sound units is due to the change in the behavior of different articulators during production. This is also manifested in the produced speech signal in terms of different attributes. The sonority information is correlated with basic production pattern of sound units. Moreover, degree of sonority is associated with spectral sharpness, strength of excitation (SoE) and periodicity. These three aspects greatly affect the perception of speech signal. Therefore, the representation of sonority can be incorporated as an additional information in the SPSS framework to improve perceptual quality of synthesized speech.

## 1.5 Motivation for the present work

The present work aims to improve naturalness and intelligibility of synthesized speech by exploring different directions to overcome some of the downsides of SPSS. The concerns mentioned in Section 1.3 associated with the quality of synthesized speech can be addressed by integrating sonority feature in the SPSS framework. The effect of formant prominence and excitation source strength on sonority associated with a sound unit, make it an important cue related to the perception. As found in the

studies made in [18, 19], the relative sonority notion between two adjacent sound units has a strong impact on human speech perception. It influences the perception of syllable and word structure and therefore may have impact speech intelligibility, which is less studied in the literature of speech processing. Therefore, in this thesis, we made an effort to extract the sonority information from the speech signal based on the different acoustic correlates of sonority as studied in the phonology literature. Along with conventional source and spectral information used in SPSS, the additional sonority information may provide us some knowledge to be incorporated for improving the naturalness and intelligibility of the synthesized speech. Based on this motivation, following are the three major contributions established in this thesis.

### 1.5.1 Representation of rich acoustic characteristics

The conventional excitation source and spectral parameters extracted from the speech signal in SPSS donot carry the information regarding rate of glottal closure, periodicity and formant prominence. These factors in addition with basic source and spectral features may result in a sufficient representation of natural speech signal in parametric form. As the sonority aspect preserves the informations regarding relative formant prominence, SoE and regularity in signal structure, a set of features representing these can be extracted from the training speech corpora in the parameter extraction module. This feature is derived from excitation source, vocal-tract system and suprasegmental aspects of the speech signal, which has capability to delineate the degree of sonority associated with each segment of speech signal. This parametric form of speech signal is modeled using HMMs in the SPSS framework. The sonority feature is further used in improving quality of synthesized speech by incorporating it in the improved post-filtering (PF) and voicing decision methods.

### 1.5.2 Alleviating over-smoothing of parameter sequences

The sonority associated with a sound is related to the degree of openness of the vocal-tract that eventually changes during production of different sound units. The effects of change in vocal-tract constriction are reflected in the speech signal in terms of VTS as well as the excitation source, with different categories of sound units [20]. As the vocal-tract constriction decreases, the SoE and formant prominence increase. Depending on the sonority associated with each frame of synthesized speech, the behavior of formant peaks, valleys and SoE may be different. Therefore, the nature of the VTS and excitation source of synthesized and natural speech signal can be analyzed with reference to the

sonorant classes. This kind of sonorant class dependent comparison may lead to development of PF methods based on sonorant categories. The PF approaches can be applied during synthesis in order to increase the variance as well as to preserve fine spectral structure in the synthesized speech signal. For this, different PF factors can be derived with respect to the sonority associated with segments of speech signal. With this motivation, the sonority feature is exploited in PF to improve the naturalness and intelligibility of synthesized speech.

### 1.5.3 Improved voiced/unvoiced detection

The quality of synthesized speech obtained from SPSS significantly relies on excitation source generation. Voiced/unvoiced decision is an essential component for generation of excitation source. It is obtained from fundamental frequency and other excitation source evidences in the literature. The discontinuity at the point of contact in the vocal-folds releases a sudden puff of air into the vocal-tract resulting voicing effect in the produced speech signal. The perceptual reflection of voicing over the sound produced is correlated with the sonority information which is related to both less vocal-tract constriction and significant glottal vibration. Therefore, the possible variation in voicing with the change in supraglottal pressure due to vocal-tract constriction is intact in the sonority associated with a sound unit. Voicing and less vocal-tract constriction are the two most effective correlates of sonority. Moreover, the voicing distinctions potentially contribute to the sonority hierarchy for sonorants and obstruents uniformly. Therefore, the voicing effect can be captured by the sonority measurement derived from system, source and suprasegmental information in the speech signal.

## 1.6 Organization of the thesis

To overcome the above mentioned issues, different directions to be explored in different modules of SPSS are presented in rest of the thesis as follows. **Chapter 2** reviews existing efforts in the literature to improve naturalness of synthesized speech obtained from SPSS. The three major issues, inadequate parametric representation of speech signal, over-smoothing of generated parameter sequence and inefficient voicing decision during excitation source generation are discussed in detail. The limitations of existing methods along with the scope of improvement is also discussed in this chapter.

**Chapter 3** explains the sonority feature extracted from speech signal. The significance of sonority information in speech signal along with production behavior of different sonorant sound units is also described. Then, a potent feature having the capability to correctly delineate the degree of sonority

is proposed. The sonority feature is extracted using evidence obtained from excitation source, vocal-tract system and suprasegmental aspects of speech signal. The efficacy of the proposed feature is established by applying it in sonorant classification. Further, the sonority feature also improves the performance of phoneme recognition and vowel onset point (VOP) detection.

The incorporation of the extracted sonority feature in the SPSS framework is explained in **Chapter 4**. The work proposed in this chapter primarily aims to overcome over-smoothing of parameter sequence. Initially, a comparison of different attributes is made between natural and synthesized counterpart. Based on this, a sonorant class based post-filter is designed to improve the dynamic range of generated parameter sequence. Along with the conventional source and spectral PF technique, a novel method for modification of spectral slope of synthesized speech is also proposed.

The exploration of sonority feature in voicing decision is carried out in **Chapter 5**. The voicing decision plays an important role in excitation source generation, which in turn affects naturalness of synthesized speech. A voiced/unvoiced classifier is modeled using the sonority feature that is further employed during excitation source generation. The incorporation of this efficient voicing decision algorithm improves the accuracy of excitation source generation during speech synthesis.

Each of the above chapters contributes to a specific module of the SPSS framework. **Chapter 6** presents a combined framework that integrates the sonority feature in the SPSS and uses this information for PF and voicing decision at the same time. The combined effect of the individual methods brings further improvement in quality of synthesized speech.

A summary of the present work is reported in **Chapter 7** by highlighting the contributions made in this thesis. A short description of different directions towards development of a natural sounding SPSS system are also presented in this chapter. Finally, the directions of future work related to this thesis are explained.

# 2

# Improvements in SPSS - A Review

Contents

### Objective

*In SPSS, speech specific features are modeled with respect to the linguistic representation. It uses a parameter generation algorithm to derive these features from the statistical models for a given text input, which are further applied to a vocoder to render the synthesized speech waveform. This approach has become popular due to its flexibility in altering the statistical behavior of models depending on the requirement. However, one of the major drawback of SPSS is muffled quality of synthesized speech. There are several works reported in the literature to alleviate this issue. These approaches are developed from different aspects, such as representation of speech signal in terms of different parameters, various modeling techniques, parameter generation methods, modification of the vocoder, PF and enhancement of synthesized speech. Each of these methods has a significant impact on the quality of synthesized speech along with the complexity involved. The objective of this chapter is to review the development of three major approaches that include incorporation of new features, PF and voicing decision algorithms in the SPSS framework. Each of these broad categories of techniques has compelling contribution towards improving naturalness and intelligibility of synthesized speech. Along with discussion of pros and cons of each method, the possible directions for further advancement are also presented in this chapter.*

## 2.1 Introduction

A TTS system deals with conversion of input text message to equivalent speech [21]. Typically a TTS system has two main modules, text analysis and speech waveform generation. In the text analysis module, the input text is passed through a language detection algorithm, language specific grapheme to phoneme conversion, and prosodic information like duration, pitch and stress. In the speech waveform generation module, speech waveform is generated from the produced linguistic specification. The goodness of synthesized speech from TTS is often measured in terms of intelligibility and naturalness. Intelligibility refers to how well the message content is comprehensible to the listeners, while naturalness refers to how well the synthesized speech is similar sounding to human speech. Thus highly intelligible and natural speech is mostly desired for any practical applications. A good TTS system will find many applications like virtual assistant, virtual news reader, story telling in audio books, screen reader, telephone services, speech to speech translation and so on.

With the increase in contribution of technology in day-to-day life, a natural sounding compact TTS

system becomes of paramount significance. As the practical applicability demands a low footprint TTS system with sufficient intelligible synthesized speech, the research in speech synthesis has been moving from concatenation of speech segments (from stored waveforms) towards use of statistical generative models, in the last decade. As the concatenative approach of speech synthesis is able to produce natural speech compared to model based approach, therefore the goal of model based approach is to improve the quality of synthesized speech. Instead of directly storing the waveforms corresponding to training speech corpora, the SPSS preserves the generative models trained using source and spectral features with reference to corresponding labels. During synthesis, these models are used to generate the same features, which are further passed through a vocoder to derive the synthesized speech signal. However, the employed generative models capture the average information from similar segments of speech signal, which may loose prosodical as well as other higher level detailed information that are intact in the natural speech signal. This results in several developments in different modules of SPSS with a common goal to bring the synthesized speech quality to the level of natural speech. The mostly used generative statistical models for speech synthesis are HMMs and deep neural network (DNN) models. In the last few years DNN based SPSS has gained popularity [22]. However, some common issues like over-smoothing of parameter sequences, inefficient voicing decision can be observed in both the DNN and HMM based approaches. Looking into the relatively simple architecture and wide rage of studies made in the literature, we have investigated the HMM based speech synthesis framework in this thesis.

The issues involved with the SPSS framework may arise from divergent facts and distinctive aspects that result in muffled quality of synthesized speech. The three basic facts are simple vocoder architecture, inaccuracy in statistical modeling of parameter sequences and over-smoothing of generated parameter contours. Apart from these three primary reasons the other facts include poor representation of speech signal in parametric form, inadequate parameter generation algorithm, error in voicing decision algorithm, simple excitation source generation method and lack of adequate prosodic information. The scopes for improving the quality of synthesized speech obtained from SPSS include improvement in the vocoder architecture, improved source generation using different additional parameters, spectral representation, adequate modeling techniques, parameter generation algorithms and methods for alleviating over-smoothing effect. In this thesis, our contributions focus more on signal processing based solutions instead of datadriven statistical approaches. We mainly focus on

three aspects related to SPSS, namely:

- Poor representation of speech signal in parametric form.
- Over-smoothing of generated parameter sequence.
- Inaccurate voicing decision algorithm in generation of excitation source.

Despite the fact that, the first aspect mentioned above cannot act as a standalone issue or solution, there is need to extract additional information apart from conventional source and spectral features to solve the other two issues. The additional information in terms of features may be further employed in different modules to improve the synthesized speech quality. In the basic HMM based SPSS approach the source parameters used are log-fundamental frequency (log $F_0$) and spectral parameters are MGC, and their dynamic (delta and delta-delta) counterpart. The refinements made in the literature have brought several other acoustic features to be extracted from the speech signal, that act as additional information for improvements in other aspects related to synthesized speech. The extra information aided by the feature may lead to improvement in the vocoder, to generate a better excitation source or in the PF approach. These additional features are also employed to build HMMs along with the traditional features. In this regard, different parameters reported in the literature and their impact on the synthesized speech are also reviewed in this chapter.

Although incorporation of improved statistical model tends to estimate the parameters accurately, the over-smoothing effect cannot be eliminated to a great extent. Improving the dynamic range of parameter sequence seems to affect quality of synthesized speech to a great extent [23]. Therefore, this chapter focuses on detailed description of different techniques employed in the literature to overcome the over-smoothing imposed by statistical models. Further, it reviews the efforts made in the literature to incorporate correct voicing decision required for excitation source generation. The improved voicing decision may lead to increase the accuracy of generated excitation source that impact on naturalness of synthesized speech. Based on these objectives, state-of-the-art techniques used for TTS, followed by a short explanation on advantages of SPSS over USS are described in Section 2.2 of this chapter. In Section 2.3, different features integrated to improve different components of the SPSS framework are reviewed in detail. Section 2.4 establishes the effect of over-smoothing on synthesized speech followed by a review of existing techniques to overcome the same. In the same way effect of voicing decision on synthesized speech and different approaches followed in the literature to correctly detect voiced/unvoiced frames are described in Section 2.5.

## 2.2 Approaches for text-to-speech synthesis

The two state-of-the-art approaches to develop TTS system are concatenative speech synthesis using unit selection algorithm and statistical generative model based approach. The concatenative speech synthesis joins speech segments corresponding to different units based on the text labels. In contrast to that, the SPSS approach generates source and spectral parameters from the generative models corresponding to text labels developed in the training stage. The parameters are further used in a vocoder to restore the speech signal. The developments in concatenative speech synthesis aims to get smooth boundaries at the joining points and less footprint, while that of SPSS target to accomplish more naturalness and flexibility.

### 2.2.1 Concatenative speech synthesis

The concatenative synthesis has progressed from diphone based fixed inventory TTS [24] to USS, where appropriate units are automatically picked from a large database based on the cost function [25]. Since the natural speech segments are concatenated, the prosodic variation is intact in the synthesized speech that leads to its *natural* quality. Based on the text to be synthesized, the target specifications of the units are obtained from the prosody prediction module. The distance of target specifications with candidate unit is measured by *join cost* and *target cost* [25]. As shown in Figure 2.1, the target cost $C^{(t)}(t_i, u_i)$ between $i^{\text{th}}$ candidate unit $(u_i)$ and target unit $(t_i)$, is defined as,

$$C^{(t)}(t_i, u_i) = \sum_{j=1}^{p} w_j^{(t)} C_j^{(t)}(t_i, u_i), \tag{2.1}$$

where, $j = 1, 2, ..., p$ denote indexes for different features used in calculating target cost, which include phonetic and prosodic contexts and $w_j^{(t)}$ are the weights assigned to each of these features. $C_j^{(t)}(t_i, u_i)$ is the target sub-cost for $j^{\text{th}}$ feature between $t_i$ and $u_i$. The concatenation cost $C^{(c)}(u_{i-1}, u_i)$ between two subsequent candidate units $u_{i-1}$ and $u_i$ is

$$C^{(c)}(u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^{(c)} C_k^{(c)}(u_{i-1}, u_i), \tag{2.2}$$

here, $k = 1, 2, ..., q$ denotes indexes for spectral and other acoustic features used in deriving concatenation cost and $w_k^{(c)}$ are the weights associated with each of them. $C_k^{(c)}(u_{i-1}, u_i)$ is the concatenation sub-cost for $k^{\text{th}}$ feature between $u_{i-1}$ and $u_i$. The concatenation cost calculation framework is shown

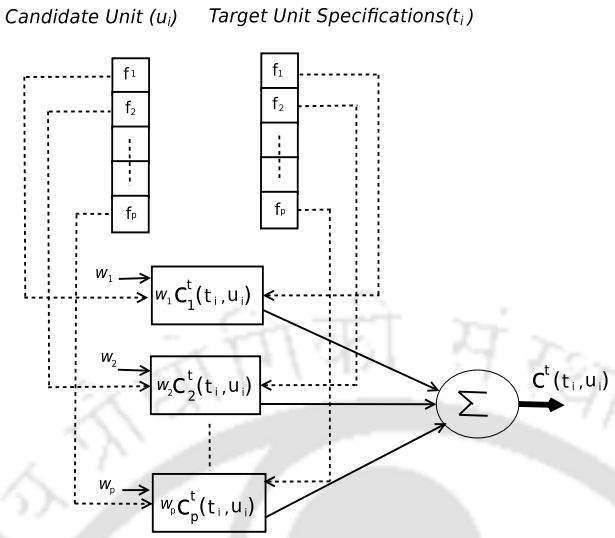


**Figure 2.1:** Target cost calculation.

in Figure 2.2. The overall cost function ($C(t_{1:n}, u_{1:n})$) is the sum of both the cost functions.

$$C(t_{1:n}, u_{1:n}) = \sum_{i=1}^{n} C^{(t)}(t_i, u_i) + \sum_{i=2}^{n} C^{(c)}(u_{i-1}, u_i), \tag{2.3}$$

where, $n$ is the number of units to be selected for synthesizing a particular utterance. The optimal string of units $\hat{u}_{1:n} = u_1, ..., u_n$ is selected from the database by minimizing the overall cost function.

$$\hat{u}_{1:n} = arg \min_{u_{1:n}} C(t_{1:n}, u_{1:n}), \tag{2.4}$$

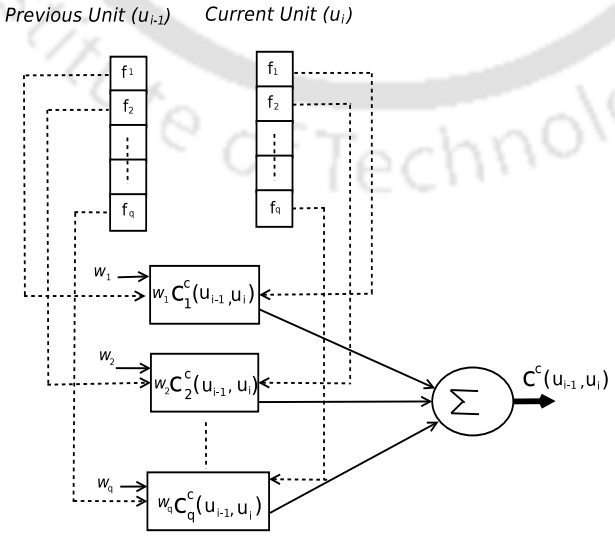


**Figure 2.2:** Concatenation cost calculation.

Generally, clustering based method is used to make the calculation of target cost easier. Each

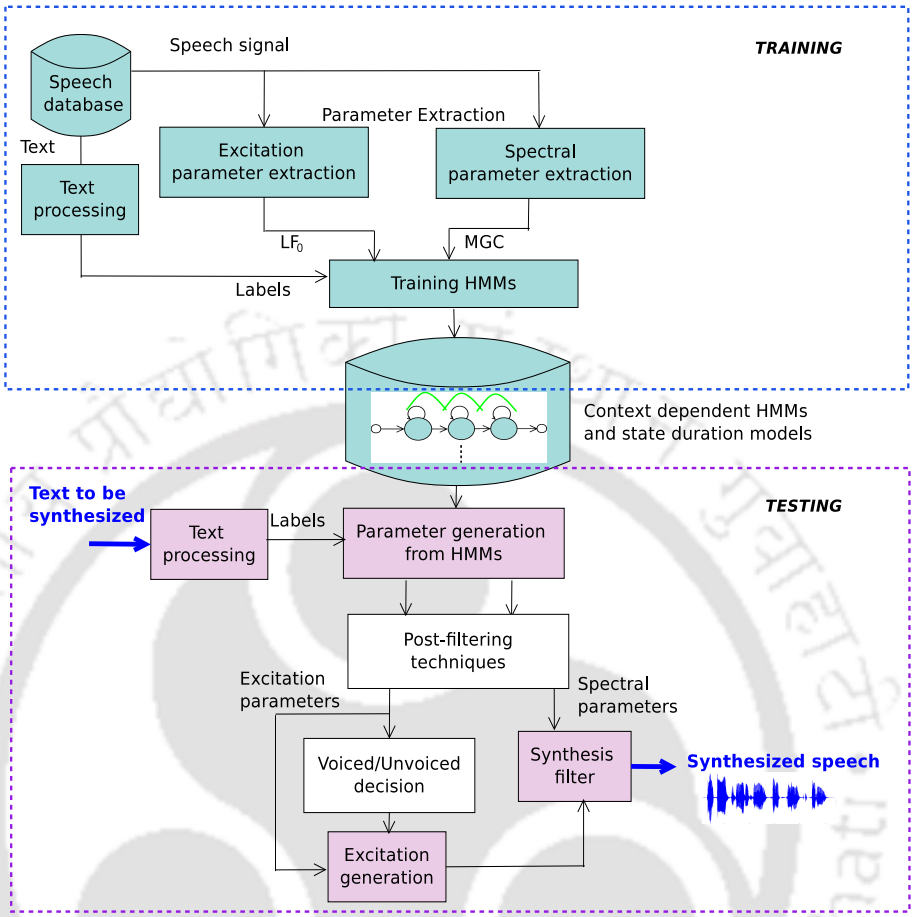


**Figure 2.3:** Overview of general SPSS scheme.

decision tree corresponds to a unique unit (phone/syllable), where each node is split based on questions about phonetic, prosodic and contextual features [26]. Calculation of target cost and concatenation cost functions based on statistical methods are also proposed [27–30]. Another aspect of concern in concatenative speech synthesis is the optimal size or type of the basic unit. It may be frame sized [30,31], HMM-state sized [32,33], half phones [34], diphones [26] and larger units [35,36]. In [37], it is reported that smaller units can have more optimal joining points and less footprint size.

Another limitation of USS is the lack of flexibility to convert the output speech to desired style, emotion or change the voice of speaker. This can only be incorporated in the USS based TTS by designing specific database according to the desired speaker, speaking style and emotion. This is a difficult and cost intensive task [38, 39]. Owing to the above mentioned difficulties in USS based concatenative speech synthesis, SPSS has grown in popularity over the last decade [5, 40].

### 2.2.2 Statistical parametric speech synthesis

Instead of storing actual instances of units in the database, SPSS uses statistical generative models (usually HMM) to synthesize speech signal. The models are developed by using parameters related to source and spectral information that can be further used in synthesis. The basic framework of HMM based SPSS approach is shown in Figure 2.3, with two primary modules, *training* and *testing*. During training spectral parameters representing vocal-tract information and excitation parameters are extracted from the given speech files with respect to corresponding labels derived from text analysis block. Although different representations of VTS can be used, typically MGC coefficients and their dynamic features are employed for this. The excitation parameters include log $F_0$ and its dynamic features. Dynamic features are first and second order derivatives of speech parameters used to include dependency of features of one frame over its nearby frames [41]. If $\boldsymbol{c}_t$ represents MGC coefficients vector at time instant $t$, then its dynamic feature vector $\Delta\boldsymbol{c}_t$ is represented as

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L}^{L} \boldsymbol{w}(\tau)\boldsymbol{c}_{t+\tau}, \tag{2.5}$$

where, $\{\boldsymbol{w}(\tau)\}_{\tau=-L}^{L}$ are window coefficients and $L$ is the window length.

A maximum likelihood (ML) estimation by using expectation maximization (EM) algorithm [42] is used to estimate the model parameters as follows.

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} p(\boldsymbol{O}|\mathcal{W}, \boldsymbol{\lambda}), \tag{2.6}$$

where, $\boldsymbol{\lambda}$ is the set of model parameters and $\hat{\boldsymbol{\lambda}}$ is the set of estimated models, $\boldsymbol{O}$ is a set of training data corresponding to the word sequence $\mathcal{W}$. Linguistic and prosodic contexts are also considered along with the phonetic ones. Using these features and the time-aligned phonetic transcriptions, context-independent monophone HMMs are trained. All HMMs are left to right topology with each state modeled using single Gaussian distribution with diagonal covariance. The basic sub-word unit considered for HMM synthesis system is the context-dependent quinphones. These context-dependent models are built starting with a set of context-independent monophone HMMs. In this process, acoustically similar states are tied in order to reduce the total number of parameters without degrading the performance of the models. Here tree-based clustering is used for state-tying. The quinphone models are then re-estimated using Baum-Welch (BW) algorithm. The feature vectors (log $F_0$ and MGC coefficients) can be modeled simultaneously using separate stream and each phoneme HMM

has its state duration densities [40]. Simultaneous modeling is used to avoid discrepant segmentation related to different features. The distributions for MGC coefficient, log $F_0$ and state duration are clustered independently using a decision tree based context clustering technique. In this clustering the factors that affect spectrum, pitch and duration parameters are used. The MGC coefficients are modeled using continuous density HMMs, whereas log $F_0$ along with voicing decision is modeled using MSD-HMM. In this case, the observation sequence of pitch pattern consists of one-dimensional continuous pitch values and discrete symbols representing voicing decision [43]. The state duration densities are calculated on the trellis obtained in the embedded training stage and modeled by the Gaussian distributions. The context dependent duration models are clustered by using the decision-tree based context clustering technique [44]. In this case, the state duration models are explicitly used as they are not re-estimated in the EM algorithm that may result in inconsistency leading to unnatural synthesized speech. Therefore in [45], hidden semi-Markov model (HSMM) is introduced that allows the incorporation of state duration models explicitly, not only during the synthesis but also in the training phase of the system.

In the synthesis module, input text to be synthesized is converted into a context-dependent label sequence. By concatenating the context dependent HMMs, a sentence HMM is constructed. State durations of the sentence HMM are determined so as to maximize the likelihood of state duration densities [44]. For a given speech length $T$, the state sequence $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$ can be obtained by maximizing the following term in (2.7), under the constraint given in (2.8).

$$\log p(\boldsymbol{q}|\boldsymbol{\lambda}, T) = \sum_{k=1}^{K} \log p_k(d_k), \tag{2.7}$$

$$T = \sum_{k=1}^{K} d_k, \tag{2.8}$$

where, $p_k(d_k)$ is the probability of duration $d_k$ in state $k$ modeled by Gaussian distributions, and $K$ is the number of states in HMM $\boldsymbol{\lambda}$. According to obtained state durations the sequence of spectral parameters (MGC coefficients) and source parameter (log $F_0$) along with voicing information are generated using the parameter generation algorithm. Although there are several variants of parameter generation algorithms, usually it is followed from [6] as follows:

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}}\{p(\boldsymbol{o}|W,\hat{\boldsymbol{\lambda}})\}, \tag{2.9}$$

$$= \arg\max_{\boldsymbol{o}}\left\{\sum_{\boldsymbol{q}} p(\boldsymbol{o},\boldsymbol{q}|W,\hat{\boldsymbol{\lambda}})\right\}, \tag{2.10}$$

$$\approx \arg\max_{\boldsymbol{o}}\max_{\boldsymbol{q}}\left\{p(\boldsymbol{o},\boldsymbol{q}|W,\hat{\boldsymbol{\lambda}})\right\}, \tag{2.11}$$

$$= \arg\max_{\boldsymbol{o}}\max_{\boldsymbol{q}}\left\{p(\boldsymbol{q}|W,\hat{\boldsymbol{\lambda}}).p(\boldsymbol{o}|W,\hat{\boldsymbol{\lambda}})\right\}, \tag{2.12}$$

$$\approx \arg\max_{\boldsymbol{o}} p(\boldsymbol{o}|\hat{\boldsymbol{q}},\hat{\boldsymbol{\lambda}}), \tag{2.13}$$

$$= \arg\max_{\boldsymbol{o}}\left\{\mathcal{N}(\boldsymbol{o};\boldsymbol{\mu}_{\hat{\boldsymbol{q}}},\boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}})\right\}, \tag{2.14}$$

where, $\boldsymbol{o} = [\boldsymbol{o}_1^{\mathrm{T}},...,\boldsymbol{o}_T^{\mathrm{T}}]^T$ is the state output vector sequence to be generated, $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$ is the optimal state sequence and $\boldsymbol{\mu_q} = [\boldsymbol{\mu}_{q_1}^T,...,\boldsymbol{\mu}_{q_T}^T]^T$ and $\boldsymbol{\Sigma_q} = \mathrm{diag}[\Sigma_{q_1},...,\Sigma_{q_T}]$ are the corresponding mean vector and covariance matrix. $T$ is the total number of frames in $\boldsymbol{o}$. In this case the generated parameter contours $\hat{\boldsymbol{o}}$ become piece-wise constant. Therefore the parameter generation algorithm uses the relationship between static and dynamic parameters as constraint given in (2.14) as follows.

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}}\left\{\mathcal{N}\left(\boldsymbol{W}\boldsymbol{c};\boldsymbol{\mu}_{\hat{\boldsymbol{q}}},\boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}\right)\right\}, \tag{2.15}$$

where, $\boldsymbol{c} = [\boldsymbol{c}_1^T,...,\boldsymbol{c}_T^T]^T$ is a static feature vector sequence, $\boldsymbol{W}$ is the matrix that appends dynamic features to $\boldsymbol{c}$.

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}, \tag{2.16}$$

The generated spectral and excitation parameter sequence using (2.15) are then passed through Mel log spectrum approximation (MLSA) filter to derive the synthesized speech [46].

### 2.2.3 Comparison of USS and SPSS

Both USS and SPSS carry their own advantages and disadvantages based on the context in which they are used. The best synthesized speech obtained from USS is always better than that of SPSS and it sounds very natural.. However, when the particular context is not available in the inventory, the quality of USS based synthesized speech severely degrades that limits its application in domain unspecific scenario and reduces its flexibility. Figure 2.4 shows the synthesized speech and corresponding spectrogram for the same utterance obtained from both USS and SPSS. The USS system

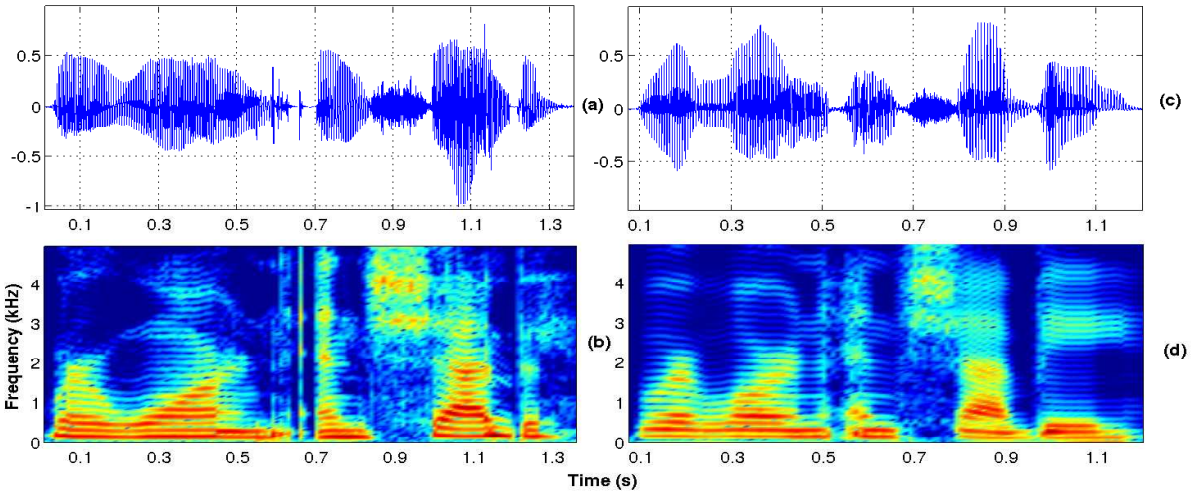


**Figure 2.4:** (a) Synthesized speech obtained from USS, (b) corresponding spectrogram, (c) synthesized speech obtained from SPSS, (d) corresponding spectrogram in SLT speakers voice for the utterance "were worn and shabby".

is developed using Festival [47], that provides all necessary modules required for building voice in a new language. The SPSS voice is developed using the traditional HMM-based speech synthesis system (HTS) toolkit [8]. Figure 2.4(a) and Figure 2.4(b) show synthesized speech signal and corresponding spectrogram in case of USS. Figure 2.4(c) and Figure 2.4(d) show synthesized speech signal with spectrogram for SPSS. Both USS and SPSS systems are developed using first 1000 utterances from CMU arctic database of SLT speaker [48]. In this example, the context for some phones are not available in the training database of both the systems. Therefore, we can observe sudden glitches and discontinuities at the joining point if we observe the region around 0.7 *s* from Figure 2.4(a) and Figure 2.4(b) obtained from USS based system. However SPSS generates smooth variation of the speech signal in the boundary between phones as observed from Figure 2.4(c) and Figure 2.4(d).

Figure 2.5 (a)-(c) show linear prediction (LP) residual that is a representation of excitation source corresponding to speech synthesized using USS, speech synthesized using SPSS and natural speech. It can be observed that the behavior of LP residual in Figure 2.5(a) corresponding to USS synthesized speech is similar to Figure 2.5(c), that corresponds to natural speech. Whereas, the LP residual obtained from synthesized speech of SPSS lacks the high frequency variations present in the excitation source of USS speech, that reduces its naturalness. In order to elucidate the VTS characteristics, the normalized log magnitude LP spectrum corresponding to the same phone of synthesized (using USS and SPSS) and natural speech are shown in Figure 2.6. It can be observed from Figure 2.6 that

the spectral peaks are more prominent in case of natural and synthesized speech obtained from USS, whereas the same is smoothed out in case of the synthesized speech obtained from SPSS. These facts limit quality synthesized speech of SPSS despite of its enormous adaptability towards applications.

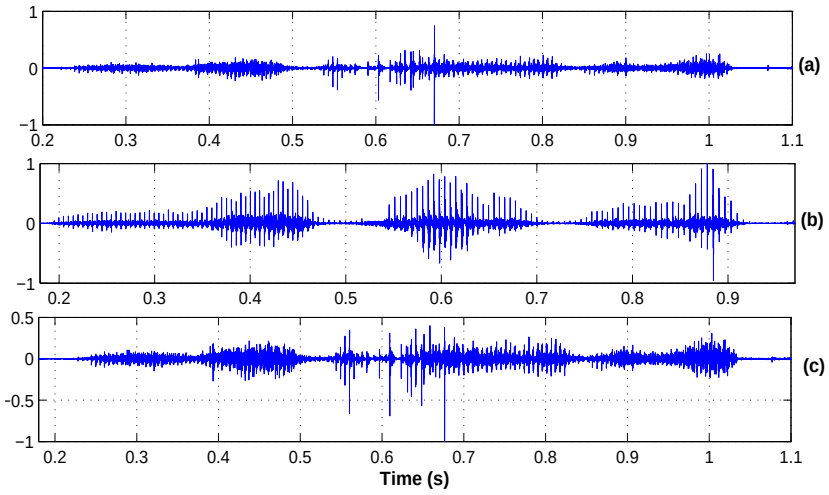


**Figure 2.5:** Linear prediction residual obtained from (a) USS, (b) SPSS synthesized speech, (c) Natural speech for the utterance "you have all" for SLT speaker.

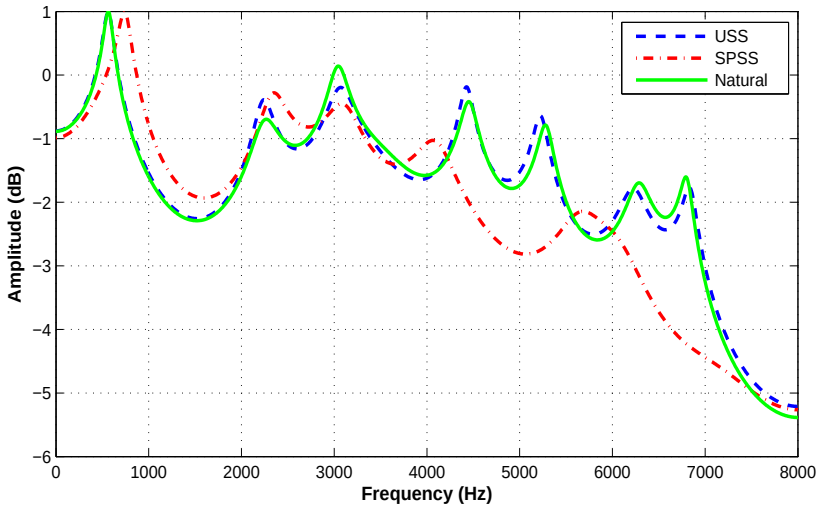


**Figure 2.6:** Spectrum extracted from natural speech signal and generated from HMM for the utterance "But she had become an automaton" for SLT speaker.

The main advantage of SPSS is its flexibility to change voice characteristics, speaking style and emotions by transforming the statistical behavior of model characteristics, which remain an issue in case of USS. For this, the adaptation techniques developed for speech recognition [49, 50] can be effectively employed using small amount of target speaker data [1, 51]. Two major techniques

used for adaptation are maximum a posteriori (MAP) [50] estimation and maximum likelihood linear regression (MLLR) [49]. Moreover, for low resource languages where collection of more data is a serious issue, SPSS systems can be developed with acceptable quality provided well designed phonetically balanced data is present. These advantages of SPSS can be helpful in a multilingual scenario or extending an existing system to new languages. The adaptation techniques are also employed to derive particular target speakers voice using small amount of training data in SPSS framework [1, 51]. Another advantage of SPSS is its ability to perform interpolation or mixing of voices, that enables the production of voices with different speaking styles and emotions that are not present in training data [2, 52]. Apart from these, the basic advantage of SPSS lies in the less amount of training data compared to that of USS, as well as its low footprint that make it easily embeddable with practical handheld devices. Moreover, in USS approach the number of units are always finite although with more training data we can increase it. The statistical approach can produce more diverse number of units with respect to different context information that results in maintaining appropriate characteristics of synthesized speech units based on different context. Due to its numerous advantages several efforts are made in the literature to bring the naturalness of synthesized speech obtained from SPSS to a level that is a competitive as USS. These advancement are described in the following sections.

## 2.3 Exploration of different features for SPSS

There are different approaches in the literature towards improving naturalness of SPSS synthesized speech by employing modification in various components of the framework shown in Figure 2.3. As the speech signal is represented in some finite dimension parametric form, all the important aspects may not be captured using only few conventional parameters or features. As mentioned earlier the conventional source and spectral features are log $F_0$, MGC coefficients and their dynamic features, that may not have enough information to reconstruct back the speech signal with all characteristics intact. If we view from this perspective, we can find several other features used along with the conventional features to adequately represent the speech signal in parametric form. For example, Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) has been found to yield high quality synthesized speech with advanced excitation source and spectral representations [53]. In the conventional HMM based SPSS system the training features include, log $F_0$ and its dynamic features (3-dimensional), MGC coefficients and its dynamic features (105-dimensional) resulting in a

total of 108 dimensional feature vector to be modeled in HMM (for speech signal with 48 kHz sampling frequency). If we use a high quality vocoder like STRAIGHT, the additional parameters or features need to be added with increased computational cost. STRAIGHT is a very high dimensional representation of speech that includes spectrum representation using 50-order mel cepstral coefficients (for speech signal with 48kHz sampling frequency) along with dynamic counterpart (150-dimensional), STRAIGHT static and dynamic log $F_0$ (3-dimensional) and mean band aperiodicity components with dynamic counterpart (78-dimensional). These parameters are modeled in HMM and during synthesis, the static parameters are generated from parameter generation algorithm as explained in [6]. The generated parameters are used to reconstruct the synthesized speech using STRAIGHT vocoder, that achieves higher quality compared to conventional MLSA vocoder. The additional parameters included and the complexity involved in STRAIGHT analysis-synthesis framework in reconstruction contribute to this improvement.

In [54], mixed excitation for HMM based speech synthesis is proposed instead of using conventional pulse train and noise based excitation source. In this framework, pitch, bandpass voicing strength and Fourier magnitudes are extracted from speech signal as representation of excitation source. During synthesis the same excitation source parameters along with spectral features are generated from the parameter generation algorithm. From generated bandpass voicing strength, the bandpass filters for pulse train and white noise are determined, from which the mixed excitation source is derived. The pulse excitation is calculated from Fourier magnitudes using an inverse discrete Fourier transform (DFT) of one pitch period in length. The periodic/aperiodic decision is obtained from the bandpass voicing strength which is used for adjusting the pitch. The same MLSA filter is used for synthesizing speech from the generated mixed excitation source and Mel cepstral coefficients. In this case, the additional features included are bandpass voicing strength and Fourier magnitudes that is further employed in excitation generation for reducing buzziness in synthesized speech.

In [55], multi-band excitation model is used, where the frame bandwidth is divided into a number of sub-bands and each band is marked as either voiced or unvoiced. Moreover, instead of MGCs a fixed number of spectral envelop samples are used to represent the vocal-tract information. These parameters are modeled using HMM instead of conventional parameters to improve the quality of synthesized speech. In the synthesis phase, the voiced part is generated based on the sinusoidal harmonic model [56], and the unvoiced part is generated as a filtered white noise. The voiced and

unvoiced speech parts are mixed according to generated bands voicing parameters to derive synthesized speech [56]. In this case due to the use of more number of parameters the model size increases along with improvement in synthesized speech quality. Authors of [57] integrated harmonic plus noise model (HNM) into HMM based speech synthesis, where speech is represented using HNM parameters that are linear predictive cepstral coefficients (LPCC) and $F_0$. These parameters are trained using HMM in the training stage. During testing, LPCC and $F_0$ are used to compute pitch synchronous HNM parameters which can be further used for HNM synthesis algorithm. In [58] Liljencrants-Fant (LF) model is used to represent glottal source signal. Glottal inverse filtering is utilized in [59] to represent the speech signal in parametric form. The parameters used in this case are $F_0$, energy of each frame, spectral energy of five bands, voice source spectrum, linear prediction coefficient (LPC) model corresponding to voiced frames and LPC model corresponding to unvoiced frames. As the number of parameters increases the complexity also grows.

Most of the methods used in the literature for improvement in naturalness requires additional parameters to represent detailed information intact in the speech signal. Among the above discussed additional parameters majority are used in excitation source generation, representation of spectral information that in turn add advantage in the vocoder. However, limited parameters are utilized to alleviate the over-smoothing issue.

## 2.4 Techniques to alleviate over-smoothing

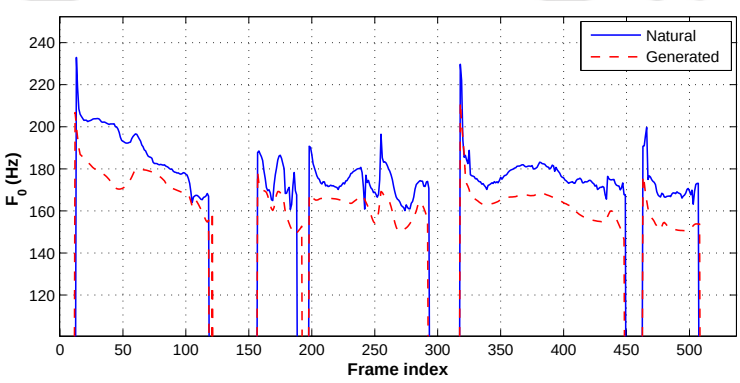


**Figure 2.7:** $F_0$ contours extracted from natural speech signal and generated from HMM for the utterance "But she had become an automaton" for SLT speaker.

The synthesized speech obtained from SPSS is of muffled quality which can be attributed as the result of over-smoothing of generated source and spectral parameter sequences. The parameter

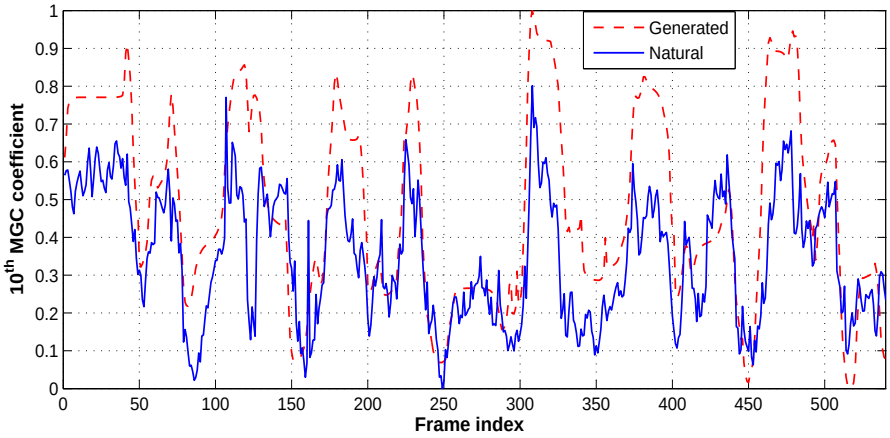


**Figure 2.8:** Contours of $10^{\text{th}}$ MGC sequences extracted from natural speech signal and generated from HMM for the utterance "We must give ourselves and not our money alone" of SLT speaker

generation algorithm followed in [6] generates source and spectral parameters from HMMs to maximize their output probabilities under the constraints of static and dynamic features as given in (2.15). The dynamic variation in temporal domain as well as the sharp formant structure intact in natural speech are absent in the synthesized speech. Figure 2.7 shows $F_0$ contours corresponding to generated from HMMs and extracted from natural speech. It can be easily observed that the natural $F_0$ contour comprises of much more variation in $F_0$ values compared to that of generated from HMMs. Similarly to show the temporal smoothing in the spectral parameter, the contour of $10^{\text{th}}$ MGC coefficients for all the frames over an utterance corresponding to natural and synthesized speech are shown in Figure 2.8, that clearly depicts the smoothing effect in spectral domain due to statistical averaging. Figure 2.9 shows the spectrum for single frame of vowel, semivowel and nasal corresponding to natural and synthesized. In each case, it is observable that most of the formant peaks in the synthesized case are deemphasized compared to that of the natural. The over-smoothing effect in the temporal domain over frames of an utterance reduces the naturalness, while within one frame the smoothed formant peaks affect the intelligibility of synthesized speech. Although the advanced modeling techniques may improve the over-smoothing to some extent, but application of an external algorithms over the generated parameter sequences is much appreciated [5]. Therefore different techniques have come out to overcome the over-smoothing introduced by statistical models. The algorithms proposed in this direction can be classified into three broad categories: (a) using real speech data, (b) using multiple level statistics, (c) using PF techniques. These are discussed in detail as follows.

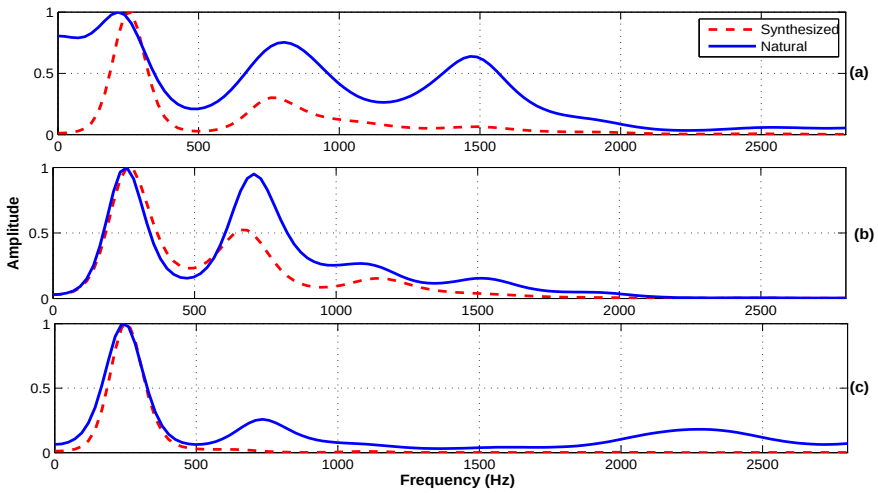


**Figure 2.9:** Synthesized and natural spectrum for (a) vowel (b) semivowel (c) nasal.

### 2.4.1 Using real speech data

This type of approach is followed in [60] that uses knowledge from training data in generation of parameters from HMMs. This conditional parameter generation algorithm generates speech parameters to maximize their output probabilities under some additional constraint as follows:

$$\hat{\boldsymbol{c}} = \arg \max_{\boldsymbol{c}} \{\mathcal{N}\ (\boldsymbol{W}\boldsymbol{c}\ ; \boldsymbol{\mu}_{\hat{\boldsymbol{q}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}})\}_{c_{t_1}=\tilde{c}_{t_1},\ldots,c_{t_N}=\tilde{c}_{t_N}}, \tag{2.17}$$

where, $\tilde{c}_{t_1}, ..., \tilde{c}_{t_N}$ are the fixed frames that can be copied from the training data and above equation can be solved using Lagrange multiplier method. An important aspect is how to select the fixed frames from the training data as discussed in [5]. In [60], the central frame of each state is fixed and the training sample is selected based on the maximum state output probability of the corresponding state. This results in selection of frames from training without spectral details which lead to limited improvement. This method is similar to hybrid approaches as it requires storing of training data along with the statistical models.

### 2.4.2 Using multiple level statistics

The other methods followed to alleviate over-smoothing is to use multiple level statistical models to generate speech parameter trajectories. Some of these methods include boosting-style additive trees [61], discrete cosine transform (DCT) based $F_0$ models [62, 63]. In [62] N-order DCT along with additional parameters like the gradient of the syllable average pitch are used for parameterization of log $F_0$ contour of syllables. A statistical model of the syllable pitch contour is then created by

clustering the parameterized vectors with a decision tree. Similar statistical models are also created for other linguistic levels other than the syllable. Use of these models during synthesis results in natural $F_0$ contour. Similarly, in [64] both phone layer and syllable layer $F_0$ models are incorporated in conventional SPSS framework. Combined multiple-level duration models [65, 66] and improved intra-phoneme dynamics models [67] are also used in the literature.

Among these category of methods the most successful one is global variance (GV) based parameter generation method that aims to improve the dynamic range of generated parameter by imposing additional constraints during parameter generation [23, 68]. It attempts to maintain the variance of generated parameter sequence similar to that of the variance of the natural. The GV, $\boldsymbol{v(c)}$ is defined as follows.

$$\boldsymbol{v(c)} = [v(1), ..., v(d), ..., v(D)]^T, \tag{2.18}$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(c_t(d) - \frac{1}{T}\sum_{\tau=1}^{T}c_\tau(d)\right)^2, \tag{2.19}$$

where, $D$ represents the feature dimension for which variance is calculated and $T$ is the number of frames of an utterance. The GVs for all training utterances are calculated and modeled using multivariate Gaussian distributions.

$$p(\boldsymbol{v(c)}|\boldsymbol{\lambda}_{GV}) = \mathcal{N}(\boldsymbol{v(c)}; \boldsymbol{\mu}_{GV}, \boldsymbol{\Sigma}_{GV}), \tag{2.20}$$

where, $\boldsymbol{\mu}_{GV}$ is the mean vector and $\boldsymbol{\Sigma}_{GV}$ is covariance matrix of GVs. The speech parameter generation algorithm maximizes the following objective function.

$$\mathcal{F}_{GV}(\boldsymbol{c}; \boldsymbol{\lambda}, \boldsymbol{\lambda}_{GV}) = P(\boldsymbol{Wc}|\boldsymbol{\lambda})P(\boldsymbol{v(c)}|\boldsymbol{\lambda}_{GV})^w, \tag{2.21}$$

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \mathcal{F}_{GV}(\boldsymbol{c}; \boldsymbol{\lambda}, \boldsymbol{\lambda}_{GV}), \tag{2.22}$$

where, $\boldsymbol{\lambda}_{GV}$ is the parameter set of GV, $w$ is the weight of GV likelihood. The second term given in (2.21) is included as penalty term to maintain the variance of generated parameter sequence. As an improvement over this, [69] proposes a two-class model for GV that discriminates between consonants and vowels. Authors in [70] changed the GV Gaussian probability density function (PDF) from a single global distribution to context dependent, where decision tree based clustering was applied to the context dependent GV PDFs to tie their parameters. Moreover, in this case GV vector was calculated only from the speech signal by excluding silence and pause regions. Another modification

of GV algorithm is proposed in [71], where the parameter generation error is re-defined by introducing an additional generation error component which measures the distortion between the generated global/local variance and original global/local variance. The parameters of HMMs are optimized so as to minimize the new generation error function. This is named as minimum generation error (MGE) criterion. Authors of [72] integrated GV criterion into a trajectory training method to provide a consistent optimization criterion as well as a closed form solution to parameter generation algorithm. The GV method does not compensate the smoothing of spectral characteristics. Therefore, an extended algorithm of GV is applied in spectral domain by considering GV of log power spectrum obtained from line spectral pairs (LSPs) into MGE model training [73]. In GV, the variance is considered over the entire utterance irrespective of the sound units present. These methods do not consider the fact that, the characteristics of the speech parameters may extensively vary based on broad categories of the sound units. Representation of these parameters using single mean and variance may not be able to reflect the large variance and fine structure, with respect to different categories sound units.

### 2.4.3 Using post-filtering techniques

The PF methods are applied to source and spectral parameters derived from the parameter generation algorithm, before passing these to the vocoder to render the speech waveform. The GV based parameter generation method described above can be considered as statistical PF method. These methods aim to improve the naturalness and intelligibility of synthesized speech. One of the well known approaches for this is enhancement of spectral peaks. Due to the over-smoothing in statistical parameter generation methods, the formants in the spectrum of synthesized speech are smoothed out. This de-emphasizing of formants affects on the intelligibility of the synthesized speech. In [74], formants are emphasized by applying post-filter to MFCC, which was originally proposed in [75] for applications in speech coding. In this case, a PF factor $\beta$ is used, $\beta = 0$ indicates no enhancement and as $\beta$ increases from 0 to 1, the formants are more emphasized. In [76] line spectral pair (LSP) based formant enhancement is performed. The difference between successive order of LSPs is weighted by some factor to enhance the formants, where less difference between LSPs makes the spectral peaks sharper and more difference makes the spectral peaks wider. For this, LSPs are derived from the HMM generated spectral parameters. The LSPs for one frame is represented by $\boldsymbol{l}_i$, $i = 1, 2, ..., D$, where $D$ is the order of LSP. The enhanced LSPs from $i = 2$ to $i = D-1$ can be calculated recursively

as follows:

$$\boldsymbol{l}'_i = \boldsymbol{l}_{i-1} + \boldsymbol{d}_{i-1} + \frac{\boldsymbol{d}^2_{i-1}}{\boldsymbol{d}^2_{i-1} + \boldsymbol{d}^2_i}[(\boldsymbol{l}_{i+1} - \boldsymbol{l}_{i-1}) - (\boldsymbol{d}_i + \boldsymbol{d}_{i-1})], \tag{2.23}$$

$$\boldsymbol{d}_i = \alpha.(\boldsymbol{l}_{i+1} - \boldsymbol{l}_i), \alpha < 1, i = 2, ...D - 1, \tag{2.24}$$

where, $\alpha$ controls the degree of enhancement. More $\alpha$ represents less enhancement. As LSPs are shifted in this method, it may result in shifting of formant peaks which is not desirable.

In [77], LPC based formant enhancement is proposed where the spectral dips in the power spectrum obtained from generated spectral parameters are modified or suppressed by weighing the nearby spectral regions by a small real valued coefficient. This enhancement applies two parameters, the width ($\delta$) of unmodified area within a spectral peak and the weighting factor ($\gamma$) for spectral valleys. The modified power spectrum is inverse Fourier transformed to derive a new autocorrelation function, that is used in the Yule Walker equations to compute a new LPC filter. The new LPC model shows sharper spectral peaks. In these methods, due to increase in prominence of spectral peaks and valleys, the enhanced synthesized speech are reported to be more intelligible. However, in these cases, all the spectral peaks are enhanced by a constant factor irrespective of their position in the spectrum. These methods increase the formant sharpness of each frame in the synthesized speech, which can be referred as a spectral enhancement over each frame. The degree of variation of the generated parameters over the temporal axis is not improved, which limits the improvement in naturalness.

Histogram equalization method is used to construct the emphasis rule for spectral parameters to reduce over-smoothing [78]. Cumulative distribution functions (CDFs) are calculated for LSP parameters corresponding to natural speech and generated from HMMs. An emphasis rule is constructed using histogram equalization method to obtain similar CDFs for natural and generated parameters. The advantage of this method over the previous methods is in terms of adjustment of the enhancement parameters. Separate emphasis rules can be obtained for different dimensions of LSPs. The frequency-dependent temporal modulations of the parameter trajectories can be explicitly enhanced with the modulation spectrum (MS) based PF method. Here, MS represents the power spectrum of the parameter trajectory. This post-filter ensures that the generated power spectrum of the parameter trajectory exhibits similar statistical behavior as that of its natural counterpart, in terms of

distribution [79, 80]. The MS $\boldsymbol{s}(\boldsymbol{c})$ of parameter sequence $\boldsymbol{c}$ of dimension $D$ is calculated as:

$$\boldsymbol{s}(\boldsymbol{c}) = [\boldsymbol{s}(1)^T, ..., \boldsymbol{s}(d)^T, .., \boldsymbol{s}(D)^T]^T, \tag{2.25}$$

$$\boldsymbol{s}(d) = [s_d(0), ..., s_d(m), ..., s_d(M)], \tag{2.26}$$

where, $s_d(m)$ is the $m^{\text{th}}$ MS of the $d^{\text{th}}$ dimension of parameter sequence $[c_1(d), ..., c_t(d), .., c_T(d)]$, $m$ is the modulation frequency index and $M$ is the half of DFT length. The MS statistics are assumed to be normally distributed as

$$P(\boldsymbol{s}(\boldsymbol{c})|\boldsymbol{\lambda}_s) = \mathcal{N}(\boldsymbol{s}(\boldsymbol{c}); \boldsymbol{\mu}^{(N)}, \boldsymbol{\Sigma}^{(N)}), \tag{2.27}$$

where, $\mathcal{N}(.; \boldsymbol{\mu}^{(N)}, \boldsymbol{\Sigma}^{(N)})$ is normal distribution with mean vector $\boldsymbol{\mu}^{(N)}$ and covariance matrix $\boldsymbol{\Sigma}^{(N)}$ of $s_d(m)$. $\boldsymbol{\lambda}_s$ is the parameter set of MS. Here, $\boldsymbol{s}(\boldsymbol{c})$ represents the MS of parameter sequence extracted from natural speech. In the same way, probability distribution function $\mathcal{N}(.; \boldsymbol{\mu}^{(G)}, \boldsymbol{\Sigma}^{(G)})$ is assumed for the generated parameter sequence. This is included in the training process. During synthesis, the post-filter is designed as follows

$$s_d^{'}(m) = (1-k)s_d(m) + k[\frac{\sigma_{d,m}^{(N)}}{\sigma_{d,m}^{(G)}}(s_d(m) - \mu_{d,m}^{(G)}) + \mu_{d,m}^{(N)}], \tag{2.28}$$

where, $k$ is post-filter emphasis coefficient. $\mu_{d,m}^{(N)}$, $\sigma_{d,m}^{(N)}$ and $\mu_{d,m}^{(G)}$, $\sigma_{d,m}^{(G)}$ represent means and standard deviations of MS of natural and generated parameter sequence. $s_d^{'}(m)$ represents the enhanced $m^{\text{th}}$ MS of $d^{\text{th}}$ dimension. This method alleviates the difference between natural and generated parameter trajectory to some extent. Similar post-filter is also applied in case of $F_0$. Recently, DNN based PF is proposed in [81], which models the conditional probability of the spectrum of a natural speech given that of synthetic speech. This data-driven post-filter helps to retain information in a higher-dimensional spectral domain. However, this method does not employ any detailed investigation of the acoustic properties of generated speech parameter sequences. Two band radial cepstral PF method proposed in [82] enables the application of different PF factors to low and high frequencies. These PF factors and cut-frequencies (to separate lower and higher frequency regions) are tunable parameters as per the requirement. As different frequency band may be affected by over-smoothing to different extent, this method helps to compensate this by applying enhancement by different factors. It also ensures the smooth transition between the bands. In [83], spectral texture in short term Fourier transform (STFT) spectrograms are reconstructed using a generative adversarial network-based post-

filter. As the network cannot be easily trained for very high-dimensional data, the spectrogram is divided into multiple frequency bands with overlap. Different post-filters are applied to individual bands and again overlapped to form a single reconstructed spectrogram. This post-filter is applied to a DNN based TTS system.

Some of the above mentioned methods attempt to emphasize the spectral peaks to retain the spectral dynamics, while the other methods adapt the temporal variation from natural speech. These approaches are able to deal with the over-smoothing issue to a large extent. At this point, we may focus on what other parameters can be modified in case of synthesized speech to further compensate the gap between natural and synthesized counterpart.

Another parameter that may be attempted for modification to enhance the intelligibility of synthesized speech is spectral slope or tilt. There are several studies in the literature of Lombard speech relating to spectral slope, which infer that average spectral tilt in Lombard speech is lesser than that of speech produced in a quiet environment [84, 85]. Several methods can be employed to flatten the spectrum, one of which is shifting of line spectral pairs towards the higher frequencies [86]. Statistical mapping of spectral tilt from Lombard speech to normal speech is done in [85], which also reduces spectral tilt. In [84], to achieve spectral tilt flattening, speech is passed through an infinite impulse response filter, where magnitude response of the filter is designed to meet the required tilt enhancement. However, methods employed for changing the tilt and their impact on different categories of sounds are not studied extensively. Moreover, the spectral tilt difference between synthesized and natural speech and its modification yet needs extensive study. As spectral tilt is only one aspect of the spectrum, therefore modifying only spectral tilt limits the improvement. In other spectral PF methods discussed above spectral peaks and valleys are considered as candidate parameters to be modified. However, dynamic range of SPSS generated parameters are still less compared to natural counterpart in both temporal and spectral domain. The statistical approach to add further constraints in parameter generation algorithm, new modeling techniques to preserve the variabilities of features or signal processing methods efficiently used to alter synthesized speech characteristics will be appreciated to solve these issues.

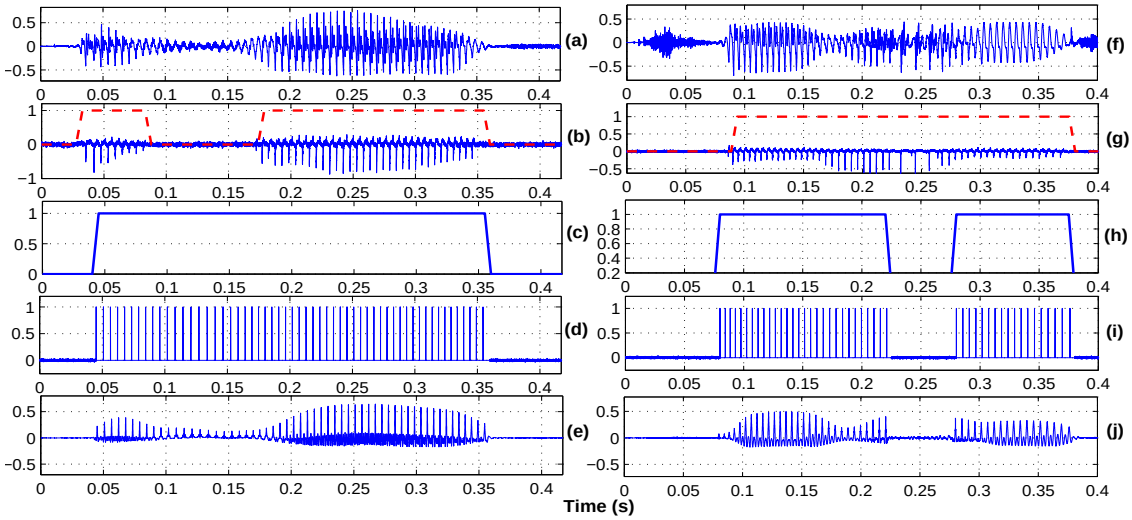


**Figure 2.10:** (a), (f) Natural speech signal; (b), (g) corresponding DEGG with reference voiced/unvoiced marking; (c), (h) voiced/unvoiced decision obtained from RAPT; (d), (i) generated excitation from voicing decision in (c); (e), (j) synthesized speech signal with the excitation shown in (d) and (i) respectively, for SLT speaker.

## 2.5 Techniques for voicing decision

Another aspect having predominant role in naturalness of synthesized speech is generation of excitation source from the HMM generated source parameter sequence. As shown in Figure 2.3, the generated excitation source parameter $F_0$ is used for deriving the voiced/unvoiced information. Based on the voicing decision excitation source is generated from the parameters using different source generation algorithms. The errors in voicing decision may lead to poor synthesized speech with reduced naturalness. In the simplistic excitation source generation algorithm, impulse sequence with pitch period and random noise are generated as excitation source component corresponding to voiced and unvoiced frames respectively. Therefore, for the frames that are misclassified as unvoiced frames random noise will be generated instead of impulses leading to harsh synthesized voiced. Similarly, for the frames with false voice labels, impulse sequence is used as excitation instead of random noise resulting in buzziness in synthesized speech. If such misclassifications are repeating for several frames in the same utterance the naturalness of the synthesized speech severely degrades.

Figure 2.10 shows the effect of wrong voicing decision on synthesized speech. Figure 2.10(a) and Figure 2.10(b) demonstrate speech signal and corresponding differential electro-glotto-graph (DEGG) with reference voiced/unvoiced marking respectively. Figure 2.10(c) shows corresponding voicing decision obtained from robust algorithm for pitch tracking (RAPT) method [87], where some false voiced

frames (corresponding to /hh/) can be observed. Corresponding excitation source and synthesized speech is shown in Figure 2.10(d) and Figure 2.10(e) respectively. Similarly, in Figure 2.10(h) falsely detected unvoiced frames can be seen corresponding to the sound /ae/ from the speech signal in Figure 2.10(f). This error is propagated to the generated excitation source and synthesized speech as shown in Figure 2.10(i), (j) respectively.

The above discussion justifies the impact of voicing decision on synthesized speech quality. There are several time and frequency domain approaches in the literature towards voiced/unvoiced detection [88]. Approaches that use zero crossing rate, autocorrelation coefficient at the first lag, normalized autocorrelation peak strength, normalized LP error, normalized low frequency energy, cepstral peak strength are employed in voiced/unvoiced classification that require some heuristic thresholding [89–91]. In [88] epochs are extracted using zero-frequency filtering (ZFF) method. By applying small amount of additive noise to the speech signal, the drift in detected epoch location compared to that of clean speech is found. Based on the robustness of the ZFF method to extract epochs in voiced regions, the voiced and unvoiced frames are classified. Statistical methods that employ neural network models, Gaussian mixure model (GMM), HMMs are also integrated with different evidence to derive better performance [89, 92]. However these methods are not integrated with the SPSS framework to derive voicing decision.

In conventional HMM based SPSS framework, $F_0$ along with the voicing information is extracted for every frame of speech during training. The $F_0$ patterns are modeled by HMMs. In order to simultaneously model discrete voiced/unvoiced decision and continuous $F_0$ patterns, MSD-HMM is used [7]. In MSD, each space has its weight and continuous probability density function whose dimensionality depends on the space. Therefore MSD-HMM includes both discrete and continuous mixture HMMs as special case. In the voiced regions $F_0$ is modeled as continuous Gaussian distribution and discrete symbols in unvoiced regions as given below.

$$b_s(o) = \begin{cases} p_v \mathcal{N}(o; \mu_s, \sigma_s), & o \ \epsilon \text{ voiced region}, \\ p_{uv}, & o \ \epsilon \text{ unvoiced region}, \end{cases} \tag{2.29}$$

where, $o$ is the observation at state $s$, $b_s(o)$ represents continuous probability density in voiced regions and discrete probability in unvoiced regions. $p_v$ and $p_{uv}$ are the probabilities of voiced and unvoiced regions, $\mu_s$ and $\sigma_s$ are mean and standard deviation corresponding to Gaussian distribution of $F_0$ in

voiced regions. During synthesis each HMM state is classified as voiced (if $p_v > 0.5$) or unvoiced (if $p_{uv} > 0.5$). MSD-HMM is widely accepted to be used in SPSS for its efficiency and performance. However, it also includes some limitations. Calculation of dynamic feature is not straightforward due to the discontinuity at the boundary between voiced and unvoiced regions. This is solved by using separate streams for modeling static and dynamic $F_0$ features. As mentioned by the authors of [93], this may result in redundant voicing probability parameters that weaken the correlation modeling between static and dynamic parameters leading to inaccurate $F_0$ estimation. Furthermore, in this case the voicing decision may also go wrong due to $F_0$ and modeling error. In MSD-HMM each stream has independent voicing decision, therefore the model becomes inconsistent in providing voiced/unvoiced information for each frame that is propagated to the excitation source generation algorithm.

To avoid such issues, continuous $F_0$ modeling approach is introduced in [93], that improves in accuracy of $F_0$ estimation and voicing decision as well. In this case, continuous $F_0$ observation is present in both voiced and unvoiced regions that can be modeled by regular HMM. Based on the $F_0$ observation, voicing decision can be obtained as opposed to MSD-HMM. There are two variants of continuous $F_0$ modeling based on implicit or explicit voicing decision. In case of implicit voicing condition modeling, the $F_0$ values extracted from voiced frames are used as observation, while for unvoiced frames other method for $F_0$ computation is used. The explicit voicing condition modeling assumes that voicing labels are observable and hence they can be modeled independently in a separate stream [93]. This makes the voicing decision independent of $F_0$. In globally tied distribution method [94], $F_0$ values for unvoiced frames are extended from neighboring voiced frames by interpolation and smoothing, that makes static and dynamic $F_0$ to be modeled in the same stream. Moreover, each state is represented by two Gaussian mixtures, one for $F_0$ values corresponding to voiced and other for unvoiced. Finally, based on voicing mixture weight, voicing decision is made. As the voicing decision can be derived independent of $F_0$, it shows significant improvement in case of voiced/unvoiced decision and therefore excitation source generation.

In [95], GMM based voiced/unvoiced change time model is used to improve the voicing decision. This imposes the constraint in voiced/unvoiced decision that can not be changed within a unit such as vowel. For voiced/unvoiced decision, [96] uses voicing strength estimation based on multilayer perceptron. DNN-based hierarchical $F_0$ modeling is implemented in [97]. In this case long-term suprasegmental property of $F_0$ features have been exploited using deep neural models to derive improved voicing

decision. In [98], instead of conventional $F_0$ extraction method, ZFF based $F_0$ estimation is used to derive the pitch contour during training. The voicing decision is obtained from SoE. The derived $F_0$ and voicing information is used in modeling of MSD-HMMs. The addition of improved $F_0$ extraction method and $F_0$ independent voicing decision add improvement to synthesized speech quality. But the drawbacks of MSD-HMM remains intact in this method. There are several other algorithms for extraction of $F_0$ and voicing decision in the literature like STRAIGHT, RAPT, robust epoch and pitch estimator (REAPER), summation of residual harmonics (SRH). STRAIGHT is a high quality speech analysis synthesis algorithm that uses wavelet based instantaneous frequency analysis for $F_0$ extraction. Here, initially instantaneous frequency based on estimate of $F_0$ is generated. Then based on maximum carrier to noise ratio the best $F_0$ estimate is selected. Based on the voicing strength of different bands voicing decision is made [99]. In RAPT and REAPER, voicing decision is made based on the autocorrelation analysis of speech and using dynamic programming for decision making. In these two methods, the voiced frames where aperiodicity components are present to little extent, gets selected as unvoiced and these methods found to have more number of voiced frames classified as unvoiced [87, 100]. In SRH, the pitch tracking is performed based on residual harmonicity [101]. As the SRH values are found to be lower in the unvoiced regions, therefore thresholding is used to classify voiced/unvoiced frames. However, SRH method is sensitive to threshold.

As discussed above the sonority information is correlated with basic production pattern of sound units. Moreover, degree of sonority is associated with spectral sharpness, SoE and periodicity. These three aspects greatly effects the perception of speech signal. Therefore, the representation of sonority can be incorporated as an additional information in the SPSS framework to improve perceptual quality of synthesized speech.

## 2.6 Organzation of the work

In the above sections different approaches employed in the literature for improving naturalness of synthesized speech obtained from SPSS is explained in detail. This section discusses the drawbacks of these methods along with the scope of further improvement. Based on these issues, the new approaches using sonority information for modification of perceptual quality of synthesized speech is proposed.

The scope of practically feasible TTS synthesis system demands naturalness, intelligibility, low footprint and flexibility to adapt the required characteristics in synthesized speech. These require-

ments aid to prefer SPSS approach of speech synthesis over other methods and lead progressive research to overcome the drawbacks of SPSS. The USS synthesized speech resembles more with natural speech in both temporal and spectral domain. The aim of all the methods to improve SPSS is to adapt the characteristics of natural speech to synthesized by overcoming the issues associated with different modules. The three basic shortcomings of SPSS discussed in this report are: inadequate parametric representation of speech signal, over-smoothing imposed by statistical modeling in the generated parameter sequences and voicing error during excitation source generation.

The first issue deals with representation of speech signal using different attributes. Use of more number of substantial features to capture detailed information in the speech signal may result in efficient reconstruction of speech signal from those parameters. However, as we increase number of parameters, the model size grows and computational cost increases. Also, the parameters used to represent the speech signal should also be exploited in the synthesis module. There are different parameters used in the literature to represent basically source and spectral information and those are mostly used in excitation source generation module as discussed in Section 2.3. Some of these parameters are of high dimension and donot have much impact on synthesis. Exploiting the usefulness of additional features in improvement of different modules may be more useful instead of using it in a single module. Moreover, the additional feature should carry complementary information to that of conventional features. In this regard, the feature representing sonority information is proposed in Chapter 3. Along with the feature extraction its application in different speech processing task is also demonstrated. As the sonority feature has correlation with speech perception, therefore it is incorporated in the SPSS framework, along with conventional features which is further used for enhancement of synthesized speech quality.

The over-smoothing effect is reflected in the less variance of generated parameters compared to their natural counterpart. The methods followed to solve this issue can be sub-categorized into: using real speech data, multiple level of statistics and PF techniques. The widely used GV based parameter generation method to alleviate over-smoothing falls under the second category. While most of the other successful methods use PF techniques for the same. A class of PF methods attempt to enhance the formants, that are smoothed out in the generated spectrum. These methods donot improve the dynamic range of parameters. The MS based method aim to preserve the behavior of natural parameters in synthesized speech, resulting in alleviating the over-smoothing by improving

the dynamic range. Both MS and GV methods consider single mean and variance of parameters for all the sound units. However, the nature of the parameters may extensively vary based on different categories of sound units. To overcome this issue a dynamic PF approach is proposed in Chapter 4. The proposed PF method introduces different amount modifications to the source and spectral parameters corresponding to different sonorant sound categories. The sonority feature extracted in the previous chapter is modeled using HMM and generated sonority feature is used for classification of frames into sonorant categories. This improves the dynamic range of the parameters and adapts the characteristics of natural speech parameter sequences. Chapter 4 also presents a method for modification of spectral tilt in synthesized speech.

The third aspect of SPSS discussed in the current chapter is methods for voicing decision. As the excitation source component plays a great role in naturalness of synthesized speech, it is always pivotal to correctly classify voiced/unvoiced frames before excitation source generation. Misclassification of voicing information may lead to wrong excitation source generation. This aspect also has significant contribution in the synthesis quality. As discussed in Section 2.5, in MSD-HMM method for $F_0$ modeling both continuous values of $F_0$ and voicing decision are modeled together. In this case, if we employ a new voiced/unvoiced classification method, the additional features are not required to be modeled, instead the resultant voicing decision will be modeled. On the contrary, in continuous $F_0$ modeling, the $F_0$ values in both voiced and unvoiced frames will be modeled. Along with that some additional features to be used for voicing decision can also be modeled. In the synthesis stage based on the classification by using generated additional feature the voicing decision can be derived. From this point of view continuous $F_0$ modeling approach can be considered as more flexible. For this purpose the usefulness of the sonority feature explained in Chapter 3 is explored. Chapter 5 explains the voicing decision method using sonority information during excitation source generation.

The work presented in this thesis first explores a potent representation of sonority information extracted from speech signal and its application in different speech processing tasks. The sonority feature is integrated in the SPSS framework along with the conventional source and spectral features. During parameter generation, the generated sonority feature is further used in dynamic PF method to alleviate the over-smoothing associated with source and spectral parameters. The same feature is further employed for voicing decision during the excitation source generation.

# 3

# Sonority Measurement using System, Source and Suprasegmental Information

Contents

### Objective

*Sonorant sounds are characterized by regions with prominent formant structure, high energy and high degree of periodicity. In this chapter, the vocal-tract system, excitation source and suprasegmental features derived from the speech signal are analyzed to measure the sonority information present in each of them. Vocal-tract system information is extracted from the Hilbert envelope of numerator of group delay (HNGD) function. It is derived from zero time windowed speech signal that provides better resolution of the formants. A five-dimensional feature set is computed from the estimated formants to measure the prominence of the spectral peaks. A feature representing strength of excitation is derived from the Hilbert envelope (HE) of LP residual, which represents the excitation source information. Correlation of speech over ten consecutive pitch periods is used as the suprasegmental feature representing the periodicity information. The combination of evidence from the three different aspects of speech signal provides better discrimination among different sonorant classes, compared to the baseline MFCC features. The application of the proposed sonority feature in phoneme recognition and VOP detection brings significant improvement in performance.*

## 3.1 Introduction

Sonority refers to relative loudness of speech sounds [9]. The most sonorant sounds, vowels, are produced with less constricted vocal-tract configuration through manipulation of the vocal-tract between glottis and lips. Position and configuration of different articulators has effect on the spectrum of generated speech signal. Narrowing the cross sectional area in the front part of vocal-tract and widening towards the back results in the decrease of first formant frequency ($F_1$). As a consequence of variation in position and length of constriction, second formant frequency ($F_2$) changes for different category of sonorants. The bandwidth of formant is associated with loss in the vocal-tract. Thus with the increase in sonority, the vocal-tract constriction decreases that results in increase in $F_1$, $F_2$ and decrease in formant bandwidth.

Compared to the obstruents, sonorants have sufficient opening of the vocal-tract to produce voicing and well defined prominent formant structure [102]. Studying the formant structure of Sonora TTS is likely to enable accurate estimation of the VTS with change in the degree of sonority. Studying the formant structure of sonorants in TTS is likely to enable accurate estimation of the VTS with change in the degree of sonority. Due to the glottal open and closed phase, the formant structure does not

show a constant behavior during one pitch period [103, 104]. The characteristics of the vocal-tract system in the open phase varies due to the coupling with vocal-fold and trachea. Whereas, during the closed phase, the speech signal is mainly due to free resonances since there is no coupling with trachea and vocal-folds [105]. Therefore, extraction of VTS from speech signal corresponding to the closed phase of each pitch period may give accurate formant estimation along with its associated measures. But, in voiced region, the glottal closing is fast and the duration of the closed phase is smaller than that of the open phase. For extracting the VTS, processes like LP analysis and STFT involve block processing and are dependent on the size and position of window. Also, these methods mask the changing shape of the vocal-tract and give an average spectrum [105].

In this chapter, HNGD spectrum derived from speech signal around the glottal closure instant (GCI) is used to estimate the VTS [106]. The GCI locations are estimated using the zero frequency filtered (ZFF) signal [107], as it is found to be more robust compared to other state-of-the-art techniques [108]. A highly tapering window is used to emphasize the speech samples around each GCI that correspond to the glottal closed phase. The sonority information present in the VTS is extracted using knowledge from the first three formants of the HNGD spectrum.

With change in the vocal-tract constriction, there is also an effect on the amplitude and spectrum of the source. Due to the change in constriction, there is fluctuation in supra-glottal pressure which has an impact on the pressure inside the glottis during the open phase of glottal vibration. This changes mechanical motion of the vocal-folds. The net effect is reduction in the amplitude of glottal source, which is reflected in the HE of LP residual as strong peaks. These peaks have correlation with an acoustic feature called SoE as discussed in [109]. With the increase in degree of sonority, SoE also increases. Hence, it can be hypothesized that, deriving an adequate representation of SoE may add some advantage in deriving sonority information from the speech signal.

Along with the change in behavior of the vocal-tract system and the excitation source with degree of sonority, temporal variation in the speech signal also takes place. This can be observed over several pitch periods. One such measure is periodicity, which corresponds to the tendency of the signal to repeat similar structure over several pitch periods. This occurs, since human speech production system changes in a continuous manner. During the production of sonorant sounds, the vocal-tract shape changes slowly and hence maintains periodicity over longer duration compared to other sounds [110]. This suprasegmental behavior of sonorants is not taken into account while analyzing vocal-tract system

and excitation source perspectives. Hence, examining the regularity in the signal structure or correlation over several small segments of the speech signal may be helpful to obtain feature representing this aspect of sonority.

### 3.1.1 Usefulness of sonority feature

The sonority feature can be applied to detection of syllable nucleus, VOP detection, phoneme classification, study of syllable structure and syllabification in different languages. Sounds with higher degree of sonority correspond to the syllable nucleus. It gives information about number of syllables present in the speech signal. Number of syllables divided by duration of the signal defines syllable rate/speaking rate. There are several approaches in the literature towards this direction. In [111], syllable nucleus is detected by loudness estimation. Energy peaks in the frequency range from 250 - 2500 Hz have good correlation with syllable nuclei. Many other methods use vowel recognizer to find syllable nucleus as given in [112].

Correlation between prominent subbands is used to capture well defined formant structure in the syllable nuclei in [113]. Before applying cross-correlation between subband energy vectors, frames are weighted by Gaussian window and then temporal correlation is estimated in order to retain inter-syllable discontinuity in case of fast speech. Then, thresholding and pitch validation of subband correlation envelope is performed to enhance the detection of syllable nucleus. In the same work, experiments are also performed to find syllable nuclei which include sonorant sounds other than vowels. The mean error calculated is more in this case. This proves that the feature cannot detect all sonorant sounds. In [114], perceptually significant evidence such as excitation source peaks in LP residual and formant peaks which contribute to the loudness are used to find the most sonorous region within syllable. All these efforts are aimed to detect basically the most sonorous sound, the vowels. There are many confusions reported within the sonorants (vowels, glides, liquids, nasals) while detecting the vowels.

Segmentation of speech into sonorant regions with high accuracy is essential for applications like automatic speech recognition (ASR) to detect the regions with high signal to noise ratio (SNR) in the speech signal [115]. In the literature, sonorant segmentation is performed using MFCCs, knowledge based acoustic features or a combination of both [10, 116]. Recently in [115], features based on both spectral and source information are proposed and a hierarchical algorithm is developed to detect sonorant and non-sonorant regions in continuous speech. However, the feature may not have potential

to further divide the sonorant regions based on the degree of sonority associated with the sound units. In order to improve the performance of sonorant detection, it is important to first quantify the degree of sonority associated with different sound units in a given speech segment, without having knowledge of phone sequence. In this work, an evidence is obtained which represents instantaneous sonority i.e. continuous change in sonority with time in the speech signal. In traditional methods, sonority is derived from the phone identity knowledge.

Looking into these studies present in the literature, it can be considered important to derive some feature which represents degree of sonority from speech signal. In this work, three different aspects of speech signal, namely vocal-tract system, excitation source and suprasegmental are analyzed to extract prospective features to discriminate among different classes of sonorants. The three attributes are analyzed individually and effectively combined to derive a multi-dimensional feature which can represent sonority. The obtained sonority feature is used in phoneme recognition, VOP detection and results show improvement. In the analysis of all the features, focus is on classifying within the sonorants according to the sonority hierarchy.

Rest of the chapter is organized as follows. Features of vocal-tract system for sonority detection are proposed in Section 3.2. Excitation source and suprasegmental features are presented in Section 3.3 and Section 3.4, respectively. Section 3.5 describes the combination of proposed evidence to represent sonority measure. Section 3.6 shows the experiments performed to demonstrate the usefulness of sonority evidence. The application of the proposed sonority feature is demonstrated in different speech processing tasks such as phoneme classifier and VOP detection in Section 3.7. In Section 3.8, summary, conclusions and future direction are mentioned.

## 3.2 Features of vocal-tract system for sonority detection

The categorical formant structure in the VTS of sonorant sounds can be interpreted by measures associated with amplitudes of spectral peaks and valleys, formant bandwidths and formant slope. Bandwidth of the spectral peak decreases, while the spectral peak value increases with increase in degree of sonority. The peak-to-valley ratio (PVR) of spectral peak is a direct representation of spectral prominence, that is inversely proportional to the corresponding bandwidth. Spectral prominence refers to spectral peaks with more sharpness and higher energy, which increases with degree of sonority. This depends on PVR, slope, bandwidth and amplitude associated with the spectral peaks. Narrow

constriction results in relatively low values of formant frequencies and spectral peaks. High-vowels are produced by raising the tongue body thus forming narrow constriction in the front part of vocal-tract. This results in decrease in $F_1$ and increase in bandwidth, primarily due to acoustic losses in the vocal-tract walls and glottis. As explained in [20], due to less spacing between $F_1$ and $F_0$, the response of low frequency auditory nerve fibers are dominated in low frequency region by $F_1$, resulting in production of relatively stable response in auditory system. In contrast to high-vowels, low-vowels are produced by narrowing the posterior part and widening towards lips, resulting in increase in $F_1$ and higher difference between $F_1$ and $F_0$. Due to this difference, the auditory nerve fibers near $F_0$ are not dominated by $F_1$. As a consequence, there is a fall in the spectrum below $F_1$ [20]. Due to the intermediate position of tongue body during production of mid-vowels, $F_1$ also lies in between that of high-vowels and low-vowels. In this case, the auditory nerve fibers are in synchrony with either $F_1$ or $F_0$. Fluctuation of second and third formant frequencies, $F_2$ and $F_3$ depends on the constriction length and position in the vocal-tract.

During the production of nasals, the vocal-tract is completely closed, while the velopharyngeal part is open and there is no pressure increase behind the constriction. In this case, during the time of closure of vocal-tract, if the vocal-folds are in a position of voicing, the same will continue after the closure [20]. Nasals have the first formant at a very low frequency and with less energy. The higher formants are also of weak amplitudes. Glides are produced by forming narrow constriction to an extent, so that there is no significant pressure drop across the constriction. This results in vibration of vocal-folds and lower $F_1$ with wider bandwidth. As an influence of the narrow constriction, the glottal source also gets modified. The liquids are also produced with narrow vocal-tract constriction, but the length of the constriction is shorter than that of the glides. As a consequence, $F_1$ of liquids is higher than that of glides. During production of liquids, the tongue is shaped in such a way that there is a split in the vocal-tract, which cannot be compared with an uniform tube [20].

With the increase in vocal-tract constriction, $F_1$ decreases and bandwidth of first formant increases gradually along the sequence of following sounds: low-vowels, mid-vowels, high-vowels, liquids, glides and nasals. With decrease in $F_1$, there is significant reduction in the overall spectrum amplitude. Amplitude of $F_2$ is dependent on $F_1$ and on the point of constriction along the vocal-tract. Since sonority associated with a sound unit depends on the vocal-tract constriction, the process for extraction of VTS should be appropriate.

### 3.2.1 HNGD Spectrum

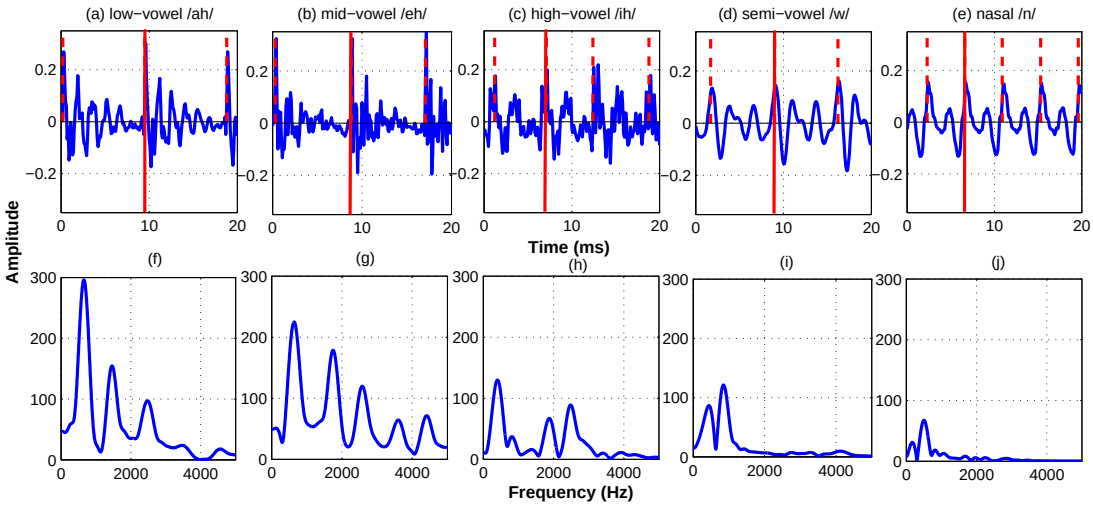


**Figure 3.1:** HNGD spectra for different classes of sounds showing apparent discrepancy in the spectrum shape. First row depicts 20 ms segment of (a) low-vowel /ah/, (b) mid-vowel /eh/, (c) high-vowel /ih/, (d) semi-vowel /w/, (e) nasal /n/ from TIMIT test database with dashed vertical lines representing epoch locations. Second row (f), (g), (h), (i), (j) show corresponding HNGD spectra, respectively, for 5 ms segment around the epoch location represented by solid line.

HNGD is found to have potential in deriving VTS for a very short segment of speech signal around GCI that mostly corresponds to the glottal closed phase as reported in [106]. It is employed in this work to analyze different characteristics of VTS for sonorant sounds. The same process of deriving HNGD spectrum around each GCI in the speech signal, as in [106] is used here:

- The frequency response of ZFF as proposed in [107] can be represented by (3.1). The analogous time domain window function shown in (3.2) is used to emphasize the speech samples closest to each GCI location. This windowing method is referred as zero time windowing (ZTW) [106].

$$\begin{aligned} |\boldsymbol{H}(w)| = |1/(1-z^{-1})^2|_{z=e^{jw}} &= 1/2(1-cosw), \\ &= 1/4sin^2(w/2), \end{aligned} \tag{3.1}$$

$$\boldsymbol{w}[n] = \begin{cases} 0 & n = 0; \\ 1/(4sin^2(\pi n/(2N))), & n = 1, 2, .....N-1, \end{cases} \tag{3.2}$$

  where, $N$ is the length of the window.

- Let $\boldsymbol{s}(n)$ be the speech signal and corresponding epoch locations are extracted using ZFF signal as explained in [107]. This can be represented by a train of impulses as shown in (3.3), where

$M$ is total number of epochs and $i_k$ is the estimated epoch location.

$$\sum_{k=1}^{M} \delta(n - i_k), \tag{3.3}$$

- Let $\boldsymbol{x}_k(n)$ be the windowed signal derived by placing the window at each epoch location.

$$\boldsymbol{x}_k(p) = \boldsymbol{s}(p) \times \boldsymbol{w}(n), \tag{3.4}$$

  where, $p = i_k, i_k + 1, ...i_k + N - 1$ and $N$ is length of window function ($\boldsymbol{w}(n)$).
- As the window function decays sharply, it is likely to mask the formant information. This effect of peak merging or smoothing can be avoided using Fourier transform phase spectra i.e. group delay (GD) spectra instead of usual magnitude spectra [117]. The numerator of the GD function (NGD) ($\boldsymbol{g(w)}$) of $\boldsymbol{x_k}(n)$ is computed as in [106].

$$\boldsymbol{g}(w) = \boldsymbol{X}_R(w)\boldsymbol{Y}_R(w) + \boldsymbol{X}_I(w)\boldsymbol{Y}_I(w), \tag{3.5}$$

  where, $\boldsymbol{X}(w) = \boldsymbol{X}_R(w) + j\boldsymbol{X}_I(w)$ is the discrete time Fourier transform (DTFT) of $\boldsymbol{x}_k(n)$ and $\boldsymbol{Y}(w) = \boldsymbol{Y}_R(w) + j\boldsymbol{Y}_I(w)$ is the DTFT of $\boldsymbol{y}_k(n) = n\boldsymbol{x}_k(n)$. The subscripts R and I denote real and imaginary parts, respectively.
- The spectral resolution is enhanced by successively differentiating NGD two times (DNGD), which shows sharp peaks at each formant location.
- In order to highlight these peaks further, HE of the DNGD is computed which is defined as HNGD spectrum, which can precisely estimate the VTS.

For different categories of sound units, HNGD is found to have the potential to detect formant characteristics with accuracy for short window, as reported in [106]. This motivates us to exploit usefulness of the HNGD spectrum in characterizing VTS to derive the sonority feature.

### 3.2.2 Effectiveness of HNGD spectrum for sonority detection

In order to substantiate the variation in formant structure of the HNGD spectrum with respect to degree of sonority, the same is shown in Figure. 3.1 for different classes of sounds. Figures 3.1 (a) - (e) show 20 ms segments of low-vowel /ah/, mid-vowel /eh/, high-vowel /ih/, semi-vowel /w/, nasal /n/, respectively. The epoch locations marked with dashed vertical lines are derived using ZFF method as described in [107]. Figures 3.1 (f) - (j) show HNGD spectra around the epochs represented by solid lines in Figure 3.1 (a) - (e), respectively. For the spectrum of low-vowel /ah/, first three spectral

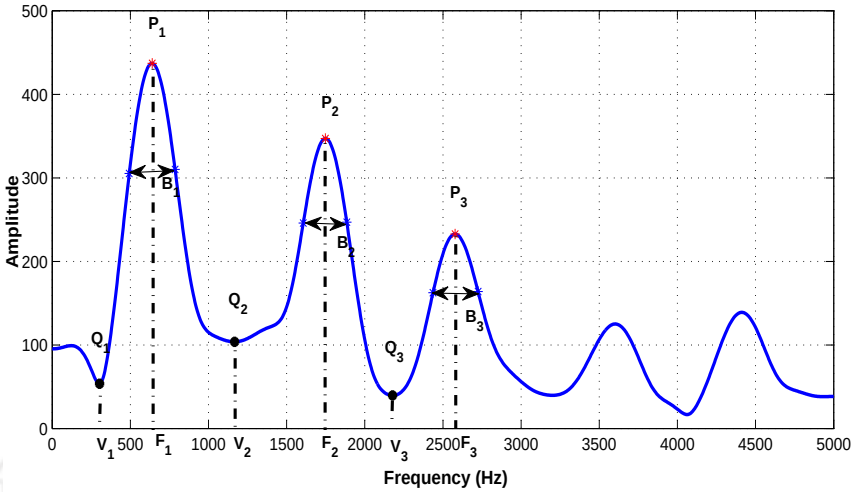


**Figure 3.2:** VTS represented by HNGD spectrum corresponding to /eh/ showing different measurements i.e. first three formant frequency values (in Hz), amplitude of spectral peaks, frequency at spectral valleys (in Hz), amplitude of spectral valleys and bandwidth.

peaks have higher amplitudes, higher slopes and lower bandwidths. The slope represents rate of decay of the spectrum amplitude from spectral peaks $(P_1, P_2, P_3)$ to corresponding preceding spectral valleys $(Q_1, Q_2, Q_3)$ as in Figure 3.2. On the other hand, mid-vowel /eh/ has lower $F_1$ and hence lower spectral prominence than that of the low-vowel. For high-vowel, $F_1$ decreases further. With the decrease in $F_1$ value, reduction in overall spectrum amplitude can also be observed. For semi-vowel and nasal sounds, differences between different attributes of spectra are depicted in Fig 3.1(i) and (j). Influenced by these observations, sonority feature is represented using different statistics from the VTS represented by the HNGD spectrum.

### 3.2.3 Proposed features of vocal-tract system to find degree of sonority

In order to find the degree of sonority associated with a sound unit, different attributes of VTS are derived from the HNGD spectrum, obtained around each epoch location. Different classes of sonorant sounds from TIMIT database used in this study are *nasals* ([m], [n], [ng]), *liquids* ([r], [l]), *glides* ([w], [y]), *high-vowels* ([ih], [iy], [uh], [uy]), *mid-vowels* ([eh], [ey], [oy], [ow]) and *low-vowels* ([aa], [ah], [ae]). These categories of sound units are segmented according to the information in TIMIT label files, succeeded by normalization with respect to the maximum value of each sound unit, for which epoch locations are derived. HNGD spectrum of energy normalized speech segment after each epoch location, is obtained as described in Section 3.2.1, which has potential to correctly characterize the

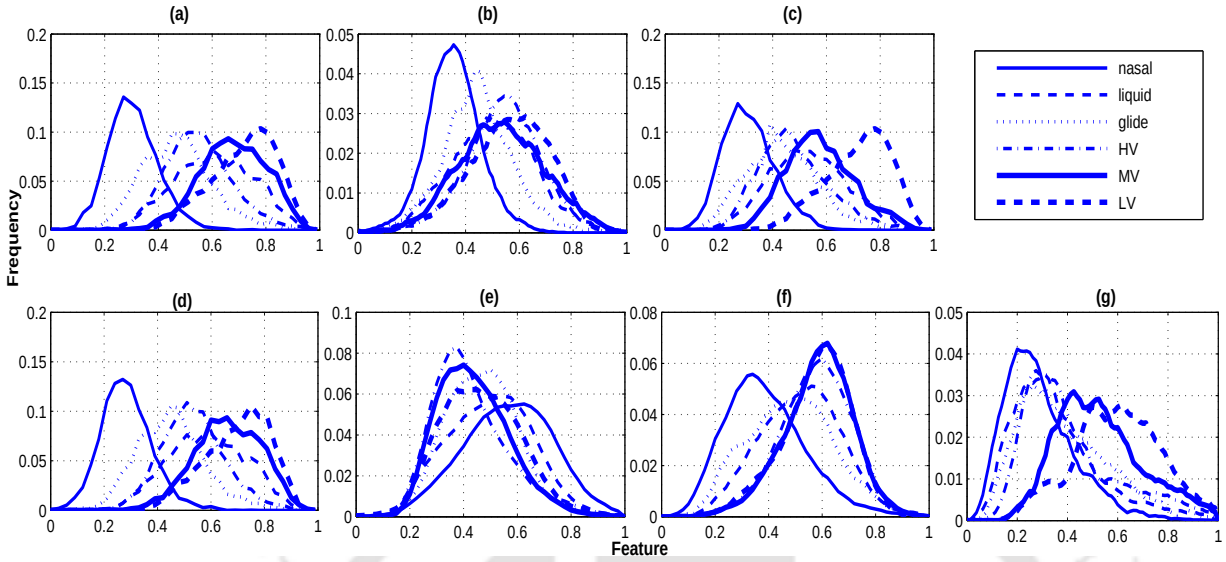


**Figure 3.3:** Distributions of the proposed sonority features for different sonorant sound units. Distribution for feature (a) $f_1$, (b) $f_2$, (c) $f_3$, (d) $f_4$, (e) $f_5$, (f) feature of excitation source ($f_6$) and (g) suprasegmental feature ($f_7$).

VTS [106]. The first three formant frequencies and associated measures are of crucial importance in many speech processing studies. Therefore, the same obtained from the HNGD spectra are employed for the task of extraction of features having capability to represent sonority. The effectiveness of each of the proposed features can be justified from the distribution curves obtained for the entire TIMIT test database for different classes as shown in Figure 3.3. Following measures are extracted from the estimated VTS for measuring sonority.

#### 3.2.3.1 Formant peak values

The first three formant frequency values (in Hz) obtained from HNGD spectrum are $F_1$, $F_2$, $F_3$ and the corresponding amplitude of spectral peaks are represented by $P_1$, $P_2$, $P_3$ as shown in Figure 3.2. With the increase in degree of sonority, $F_1$ (in Hz) increases. This is also reflected in the amplitude of spectral peaks, as increase in $F_1$ results in overall increase in the spectrum amplitude. The mean amplitude of first three spectral peaks is calculated, which is represented as $f_1$, where, $f_1 = \frac{1}{3}\sum_{i=1}^{3} P_i$. The estimated distribution of normalized value of $f_1$ for different classes of sonorant sounds is shown in Figure 3.3(a). It can be observed from Figure 3.3(a) that $f_1$ may not discriminate well between different sonorant classes, but it does provide some evidence along the lines of sonority hierarchy.

#### 3.2.3.2 Formant peak deviation

When two or more formant frequencies come close together, there is an increase in spectrum value in the vicinity of these formant frequencies. The next measure for sonority measurement from VTS is the mean of relative deviation between amplitude of first three spectral peaks. Here $D_1$ and $D_2$ are differences between amplitudes of first and second spectral peaks, and second and third spectral peaks, respectively. The mean of these differences is represented as $f_2 = \frac{1}{2}\sum_{i=1}^{2} D_i$. The distribution corresponding to normalized value of $f_2$ for different sonorant classes derived from whole TIMIT test database is shown in Figure 3.3(b). $f_2$ may provide some information along the sonority hierarchy.

#### 3.2.3.3 Spectral valleys preceding the first three formant peaks

Along with spectral peaks, spectral valleys are also of importance for overall study of the spectrum shape. Spectral valleys $(V_1, V_2, V_3)$ preceding to the first three formant frequencies $(F_1, F_2, F_3)$ are detected and the mean value of corresponding spectral amplitudes $Q_1, Q_2, Q_3$ is calculated. It is represented as $f_3 = \frac{1}{3}\sum_{i=1}^{3} Q_i$. The distribution of normalized $f_3$ derived from segments of different sonorant classes from entire TIMIT test database is shown in Figure 3.3(c).

#### 3.2.3.4 Slope associated with each formant peak

In order to detect spectral prominence, slope associated with each spectral peak is also measured. To measure the slope, first three spectral peaks $(P_1, P_2, P_3)$ corresponding to formant frequency values $F_1, F_2, F_3$ are detected. Similarly, preceding amplitude of spectral valleys $(Q_1, Q_2, Q_3)$ corresponding to frequency values $V_1, V_2, V_3$ are determined as shown in Figure 3.2. Then, slope associated with each of the first three spectral peaks is calculated as follows:

$$SP_1 = \frac{P_1 - Q_1}{F_1 - V_1}; SP_2 = \frac{P_2 - Q_2}{F_2 - V_2}; SP_3 = \frac{P_3 - Q_3}{F_3 - V_3} \tag{3.6}$$

To represent this feature, average value of $SP_1, SP_2$ and $SP_3$ is calculated as, $f_4 = \frac{1}{3}\sum_{i=1}^{3} SP_i$. The distributions are obtained for normalized value of $f_4$ for different sonorant classes in the TIMIT test database as shown in Figure 3.3(d).

#### 3.2.3.5 Formant bandwidth

Formant bandwidth is directly proportional to the loss associated with vocal-tract. This may arise from different sources such as vocal-tract walls, viscosity, heat conduction and radiation. Hence,

**Table 3.1:** Canonical correlation analysis (CCA) between different features of vocal-tract system.

| Features | Correlation value |
|---|---|
| $f_1$ and $f_2$ | 0.89 |
| $f_1$ and $f_3$ | 0.88 |
| $f_1$ and $f_4$ | 0.63 |
| $f_1$ and $f_5$ | 0.40 |
| $f_2$ and $f_3$ | 0.89 |
| $f_2$ and $f_4$ | 0.52 |
| $f_2$ and $f_5$ | 0.38 |
| $f_3$ and $f_4$ | 0.59 |
| $f_3$ and $f_5$ | 0.39 |
| $f_4$ and $f_5$ | 0.33 |

with more constricted vocal-tract configuration, bandwidth associated with peaks also increases. This results in decrease in degree of sonority. In this case we have not used the power spectrum to calculate the bandwidth. For each of the first three spectral peaks $(P_1, P_2, P_3)$, the difference between the frequencies in both side of the peaks, where amplitude of the HNGD spectrum is 0.707 of maximum value of corresponding peak is calculated. This difference is termed as bandwidth associated of the first three formant peaks $(B_1, B_2, B_3)$ in the rest of the thesis. The average bandwidth is calculated as $f_5 = \frac{1}{3}\sum_{i=1}^{3} B_i$. The distributions corresponding to normalized bandwidth is shown in Figure 3.3(e), which decreases with the increase in sonority.

The values of each of the features $f_1, f_2, f_3, f_4, f_5$ obtained from all the frames across all instances of the six types of sounds are normalized as follows:

$$f_i = \frac{f_i - \min(f_i)}{\max(f_i) - \min(f_i)}, \tag{3.7}$$

where, $i$ ranges from 1 to 5. $\min(f_i)$ and $\max(f_i)$ represent minimum and maximum values of $f_i$ extracted over all classes of sonorant sounds for entire TIMIT test database.

### 3.2.4 Combined vocal-tract feature to find degree of sonority

Each of the features $f_1, f_2, f_3, f_4$ and $f_5$ are normalized and approximated by Gaussian probability density function as shown in Figure 3.3 (a), (b), (c), (d), (e), respectively. The distributions do not provide clear discrimination among different classes of sonorants. However, still the increasing trend of the features $f_1, f_2, f_3$ and $f_4$ from nasals to low-vowels can be observed, while $f_5$ exhibits a decreasing

trend for the same. Also, some disparity in terms of overlap of distributions among different classes of sounds for each of the features of VTS can be interpreted from Figure 3.3 (a)-(e). For example, in the distribution of $f_2$, a distinct overlap between the low-vowel, mid-vowel and high-vowel can be observed. $f_1$ shows less overlap between the three vowel categories along the line of sonority hierarchy. $f_2$ has lower amount of overlap between the distributions of glides and nasals.

It can be inferred from Figure 3.3(c) that, $f_3$ possess better adequacy to bring out the differences between low-vowel and mid-vowel compared to other features. In each of $f_1$, $f_3$ and $f_4$, the liquids have higher values than that of glides, whereas according to the sonority hierarchy, glides are more sonorous than the liquids. In Figure 3.3(e), $f_5$ shows a correct reverse trend of feature values with respect to the sonority hierarchy. However, the extent of overlap between different classes is more compared to other features. Based on this interpretation, it can be inferred that the five derived features of vocal-tract system may carry different information.

The redundancy among the five attributes derived from the VTS is elucidated using canonical correlation analysis (CCA) [118,119]. The correlation values derived from CCA among different pairs of features are shown in Table 3.1. Although correlation exists between the five features of vocal-tract system, there is some extra information captured by each feature, as the correlation value is less than 1 in each case. Based on these observations, a five-dimensional feature vector of vocal-tract system is proposed in this work, which has the ability to quantify the sonority hierarchy.

## 3.3 Excitation source information for sonority detection

The SoE is related to the abruptness of the glottal closure, which is maximum for an ideal impulse and corresponds to strength of differenced electro-glotto-graph (DEGG) signal at GCIs. In order to visualize how SoE changes with degree of sonority, an effective representation of SoE derived from excitation source needs to be explored. Given the speech segment of particular sound unit (vowels, semi-vowels or nasals), LP analysis can be performed to derive the LP coefficients. The residual signal is obtained by inverse filtering the speech signal using LP coefficients. The inverse filtering suppresses the vocal-tract characteristics from the speech signal and mostly contains information about the excitation source. The residual signal shows noise like characteristics in unvoiced regions and large discontinuity in voiced regions of the speech signal. This is a good approximation of excitation source signal when LP order is properly chosen [120]. In this work, the LP residual is derived by performing

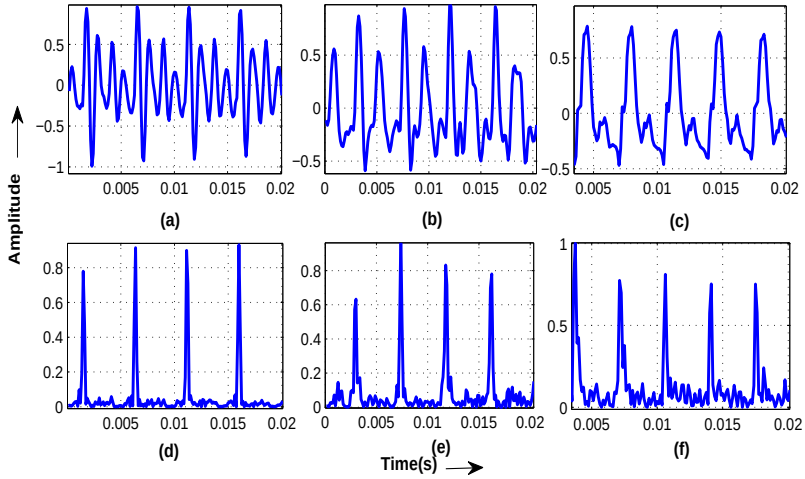


**Figure 3.4:** Illustration of difference in nature of excitation source in vowels, semi-vowels and nasals. (a)-(c) show 20 ms speech segment of vowels, semi-vowels and nasals. (d)-(f) show corresponding HE of LP residual, respectively.

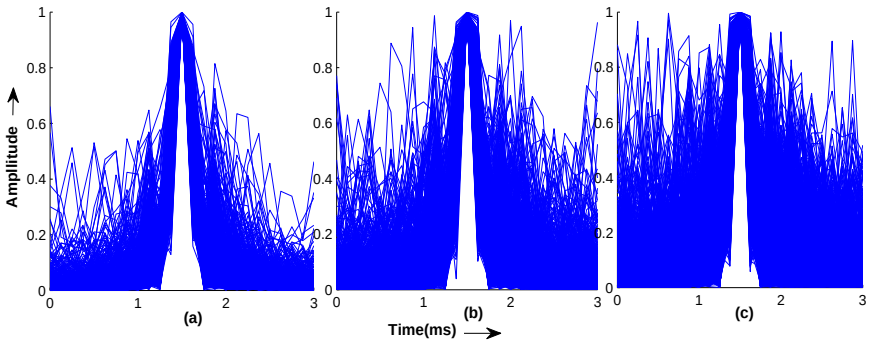


**Figure 3.5:** 3 ms duration of superimposed segments of HE of LP residual in the vicinity of impulse-like excitations for (a) vowels, (b) semi-vowels, (c) nasals.

LP analysis on overlapped segments of speech signal (size of frame =25 ms, frame shift = 5 ms, LP order = 10 and sampling frequency = 8 kHz). The GCIs are manifested as large amplitude fluctuations, either in positive or negative polarity in the LP residual. This difficulty can be overcome using the HE of LP residual [121]. The HE $\boldsymbol{h_e}(n)$ of LP residual $\boldsymbol{e}(n)$ is defined as

$$\boldsymbol{h_e}(n) = \sqrt{\boldsymbol{e}^2(n) + \boldsymbol{e}_{\boldsymbol{h}}^2(n)}, \tag{3.8}$$

where, $\boldsymbol{e_h}(n)$ is Hilbert transform of $\boldsymbol{e}(n)$ and in given by

$$\boldsymbol{e_h}(n) = IDFT[\boldsymbol{E_h}(k)], \tag{3.9}$$

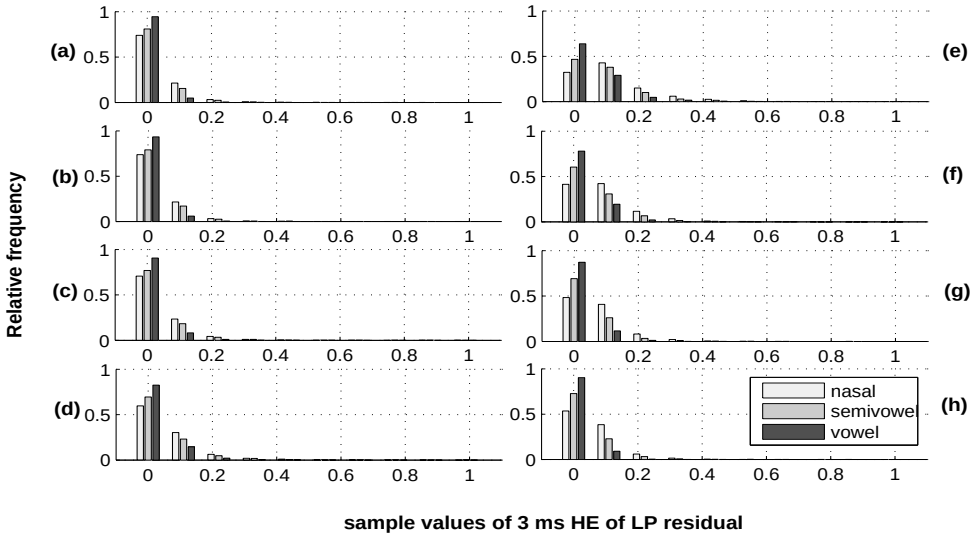


**Figure 3.6:** Histogram plot of sample values of 3 ms HE of LP residual. 3 ms segment is divided into 0.25 ms frames. (a), (b), (c), (d) correspond to 0 to 1 ms and (e), (f), (g), (h) corresponds to 2 to 3 ms of the 3 ms segment.

where,

$$\boldsymbol{E_h}[k] = \begin{cases} -j\boldsymbol{E}(k) & k = 0, 1, ...(\frac{N}{2}) - 1; \\ j\boldsymbol{E}(k) & k = (\frac{N}{2}), (\frac{N}{2}) + 1, .....(N-1), \end{cases} \tag{3.10}$$

IDFT denotes inverse discrete Fourier transform and $\boldsymbol{E}(k)$ is DFT of $\boldsymbol{e}(n)$ and $N$ is the number of points for computing DFT.

Speech segments of 20 ms and corresponding HE for vowel, semi-vowel and nasal are shown in Figure 3.4 (a) - (c) and (d) - (f), respectively. It can be observed that, the pattern of side-lobes of each peak in HE (corresponding to GCI) is different for nasals, semi-vowels and vowels. The side-lobes have higher values with respect to peak values in case of nasals than semi-vowels. In case of vowels, the amplitude of side-lobes are further reduced than that of semi-vowels.

For the entire TIMIT test database, HE of LP residual of vowels, semi-vowels and nasals are obtained. The GCIs are derived from the ZFF signal and then by searching for the nearest peaks in the HE of LP residual [107]. For each GCI, 1.5 ms segment towards right and 1.5 ms segment towards left is selected from the HE of LP residual of speech signal. These 3 ms segments are normalized (each sample is divided by maximum value among the 3 ms samples) and superimposed for each class (vowels, semi-vowels and nasals). The number of such superimposed frames used is equal for each

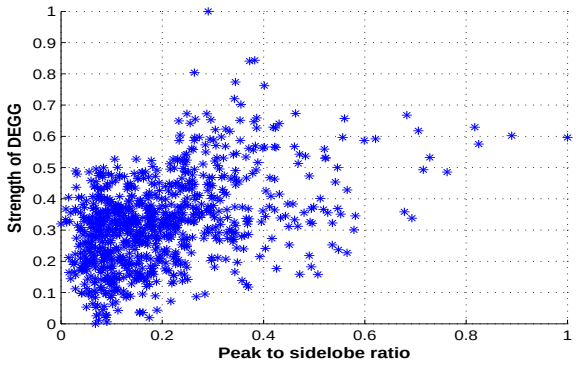


**Figure 3.7:** Scatter plot of DEGG versus peak to side-lobe ratio of short segment of HE pf LP residual in the vicinity of GCIs.

class. The resulting plot is shown in Figure 3.5. It can be clearly observed that the distribution of side-lobes around the center peak is different for the three classes of speech sounds.

To investigate the difference among the three, the 3 ms segment is divided further into frames of 0.25 ms. The distribution of values for each 0.25 ms frame is plotted using a discrete histogram as shown in Figure 3.6, where, (a), (b), (c), (d) correspond to first 0 to 1 ms (4 frames each of 0.25 ms) and (e), (f), (g), (h) correspond to 2 to 3 ms of 3 ms of HE segment. It can be observed from Figure 3.6 that (e), (f), (g), (h) show more discrimination between the classes (vowels, semi-vowels and nasals) than first 1 ms frames i.e. (a), (b), (c), (d). For example: the bins corresponding to vowels, semivowels and nasals are more separated in (f) compared to that in (b). Based on this analysis, we considered only the region from 2 to 3 ms of the 3 ms HE segment to quantify the source evidence. Since the distribution of values of HE of LP residual in glottal closure region is different for broad classes of sonorant sounds (vowels, semi-vowels and nasals), it may be appropriate to analyze the same to quantify the sonority hierarchy.

The source feature for sonority is defined as $f_6 = \frac{P}{\mu}$, where, $P$ is the value of central peak at the GCI location and $\mu$ is the mean of sample values from 2 to 3 ms duration in the 3 ms HE segment. This can be referred as *peak to side-lobe ratio* around the epoch locations which can represent SoE. As shown in Figure 3.7, the SoE derived from HE of LP residual (peak to side-lobe ratio) has approximately linear correspondence with strength of DEGG signal. The distributions of peak to side-lobe ratio representing SoE for different classes of sound show an increasing trend with the increase in sonority as observed from Figure 3.3(f). The feature of excitation source shows a significant overlap within the vowel categories, whereas it has potential to correctly discriminate source aspect of nasals and

vowels. Semi-vowels (glides and liquids) also seem to have overlapped distributions. However, the distributions of $f_6$ for each class show less variance compared to that of features of vocal-tract system.

## 3.4 Suprasegmental evidence for sonority measurement

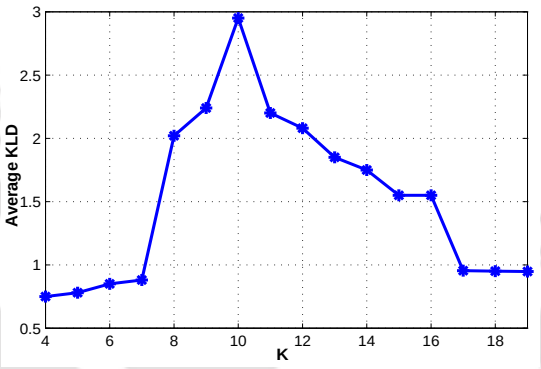


**Figure 3.8:** Change in average KLD between Gaussian distributions derived from suprasegmental feature of six classes of sonorant sound with respect to the value of K.

Sonorant sounds are prolonged with higher periodicity, where similar signal structure repeats for longer duration due to the slow change in vocal-tract configuration during production. This behavior of the sonorant sounds can be captured by measuring similarity of speech signal samples over several pitch periods rather than just one pitch period. In this work, a suprasegmental feature is derived by computing the correlation of the speech signal over $K$ pitch periods, as a manifestation of regularity in the structure of speech signal. If there are $M$ number of epochs in the given speech signal, $\boldsymbol{x_1}, \boldsymbol{x_2}, .....\boldsymbol{x_{M-1}}$ are the segments corresponding to $M-1$ number of cycles starting from one epoch to the next. The similarity over $K$ number of cycles (pitch periods) is measured as follows:

$$f_7(i) = \frac{1}{K} \sum_{j=i+1}^{i+K} \frac{< \boldsymbol{x_i}, \boldsymbol{x_j} >}{\sum^{N_i} \boldsymbol{x_i^2} \sum^{N_j} \boldsymbol{x_j^2}}; i = 1, 2, ....M-1-K, \tag{3.11}$$

where, $f_7(i)$ is the correlation coefficient that represents the suprasegmental evidence of sonority. $< \boldsymbol{x_i}, \boldsymbol{x_j} >$ represents the inner product between samples corresponding to $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, which are $i^{\text{th}}$ and $j^{\text{th}}$ pitch cycles in the speech segment. Zero padding is performed to match the dimension of $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. $N_i$ and $N_j$ are the number of samples present in $i^{\text{th}}$ and $j^{\text{th}}$ cycles. $M$ is the total number of GCIs in the given speech segment and $K$ is the number of cycles over which the feature is calculated.

For finding appropriate value of $K$, the suprasegmental feature is derived by varying $K$ value from 4 to 19. For each value of $K$, Gaussian distributions of the six classes are obtained and average

**Table 3.2:** Means and standard deviations (std) of different features of vocal-tract system ($f_1$, $f_2$, $f_3$, $f_4$, $f_5$), feature of excitation source ($f_6$) and suprasegmental feature ($f_7$) for different classes of sonorants (low-vowels, mid-vowels, high-vowels, liquids, glides and nasals).

| Evidence | Low-vowels | | Mid-vowels | | High-vowels | | Glides | | Liquids | | Nasals | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mean | std | mean | std | mean | std | mean | std | mean | std | mean | std |
| Formant Peak Values ($f_1$) | 0.73 | 0.11 | 0.69 | 0.12 | 0.56 | 0.12 | 0.48 | 0.13 | 0.62 | 0.14 | 0.32 | 0.09 |
| Formant peak deviation ($f_2$) | 0.60 | 0.14 | 0.56 | 0.14 | 0.54 | 0.12 | 0.46 | 0.11 | 0.53 | 0.14 | 0.38 | 0.08 |
| Spectral valleys ($f_3$) | 0.62 | 0.12 | 0.59 | 0.12 | 0.49 | 0.13 | 0.45 | 0.13 | 0.55 | 0.14 | 0.33 | 0.09 |
| Slope ($f_4$) | 0.71 | 0.12 | 0.67 | 0.12 | 0.54 | 0.11 | 0.46 | 0.12 | 0.60 | 0.14 | 0.29 | 0.09 |
| Formant Bandwidth ($f_5$) | 0.55 | 0.05 | 0.58 | 0.05 | 0.57 | 0.05 | 0.59 | 0.05 | 0.61 | 0.06 | 0.63 | 0.06 |
| Source($f_6$) | 0.29 | 0.06 | 0.29 | 0.06 | 0.29 | 0.06 | 0.24 | 0.08 | 0.27 | 0.08 | 0.20 | 0.08 |
| Suprasegmental($f_7$) | 0.49 | 0.14 | 0.44 | 0.15 | 0.34 | 0.16 | 0.32 | 0.15 | 0.29 | 0.14 | 0.24 | 0.11 |

Kullback Leibler divergence (KLD) measure among the six classes is calculated. The $K$ value which gives maximum KLD distance between the distribution of six sonorant classes is selected. Figure 3.8 shows that for $K = 10$, the KLD measure has the highest value. If the length of the speech segment is less than 10 pitch periods, the $K$ value is changed to two less than the number of pitch periods in the signal. For $M$ number of GCIs in the speech signal, suprasegmental feature $f_7$ will have $M - 1 - K$ number of values. This corresponds to first $M - 1 - K$ number of epochs. For last $K + 1$ number of epochs, the last value of the feature is repeated to match the suprasegmental feature dimension with that of vocal-tract system and excitation source features. The derived correlation feature is obtained for different categories of sonorants from TIMIT test database and the corresponding distributions are depicted in Figure 3.3(g). As hypothesized, proposed suprasegmental aspect of speech signal has the adequacy to delineate the sonority hierarchy. Regardless of the significant overlap between distributions of liquids, glides and high-vowels in Figure 3.3(g), it shows an increase in feature value as one moves from nasals (least sonorous) to low-vowels (most sonorous).

## 3.5 Combination of source, system and suprasegmental evidence

The means and standard deviations of each of the derived features are shown in Table 3.2. As elaborated in Section 3.2.4, the means and standard deviations of five different features of vocal-tract system carry contrasting information regarding the degree of sonority associated with. As observed from Table 3.2, from low-vowels to nasals, the mean values of $f_1, f_2, f_3$ and $f_4$ decrease sequentially with a disparity in case of glides and liquids. The latter having higher mean value than the former in

case of all the four features. It can be observed that the mean values of $f_5$ increase from low-vowels to nasals. The deviation in mean values of $f_5$ among different classes is less. Also, the standard deviation values of $f_5$ are low compared to other features of vocal tract system.

From production point of view, the difference between glides and liquids is that, in case of liquids the constriction is shorter than that of the glides. This results in higher $F_1$ for liquids than the glides. Moreover, the acoustic path in the oral cavity for liquids contains side branch or parallel paths unlike glides. This introduces extra poles and zeros in the spectrum of liquids, which lead to higher values of features of vocal-tract system for liquids than glides. The pattern of mean values of the suprasegmental feature is found to have good correlation with the degree of sonority. All the evidence derived from three different perspectives of sonorant sounds demonstrate unique trend with the change in degree of sonority. To obtain a faithful feature representation of sonority, the combination of vocal-tract system, of excitation source and suprasegmental features may be helpful. All the seven evidence have single value at each epoch location.

For each of the seven features, six Gaussian distributions can be derived representing six classes of sonorant sounds. The distance between each pair of Gaussian probability density function can be measured by KLD [122] as given by (3.12).

$$D_{KL}(A,B) = \frac{1}{2}\left\{\frac{\sigma_A^2}{\sigma_B^2} + \frac{\sigma_B^2}{\sigma_A^2}\right\} - 1 + \frac{1}{2}\{\mu_A - \mu_B\}^2\left\{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}\right\}, \tag{3.12}$$

where, $A$ and $B$ are two univariate Gaussian distributions with means $\mu_A$, $\mu_B$ and standard deviations $\sigma_A$, $\sigma_B$, respectively. $A$ and $B$ represent samples of one feature for two classes of sonorant sounds. As there are 6 classes of sonorant sounds, each feature will have 6 Gaussian distributions i.e. 15 pairs of distributions as shown in Figure 3.3. The average KLD measure is calculated for each of the 7 features over these 15 pairs of distributions as in (3.13).

$$\{D_{KL}(A,B)\}_{avg} = \frac{1}{15}\sum_{i=1}^{15} D_{KL}(A,B)_i, \tag{3.13}$$

The average KLD for each feature is tabulated in Table 3.3. The seven features shown in Table 3.3 have difference in terms of their ability to differentiate between the classes of sonorant sounds. Higher value of KLD represents greater ability of the feature to discriminate different classes of sonorants, and hence more weight should be assigned to that particular feature dimension. Based on the average KLD between different classes of sound, weights corresponding to each of the seven features ($w_i$) are

**Table 3.3:** Average KLD between Gaussian distributions of six classes of sonorant sounds and corresponding weights assigned for different features of vocal-tract system, excitation source and suprasegmental feature.

| Features | Average KLD | Weights |
|---|---|---|
| Formant Peak Values ($f_1$) | 1.14 | 0.1049 |
| Formant peak deviation ($f_2$) | 0.95 | 0.0874 |
| Spectral valleys ($f_3$) | 1.10 | 0.1012 |
| Slope ($f_4$) | 1.09 | 0.1003 |
| Formant Bandwidth ($f_5$) | 1.62 | 0.1490 |
| Source ($f_6$) | 2.02 | 0.1858 |
| Suprasegmental ($f_7$) | 2.95 | 0.2714 |

derived such that the sum of all weights are equal to unity.

$$\sum_{i=1}^{7} w_i = 1, \tag{3.14}$$

where,

$$w_i = \frac{[\{D_{KL}(A,B)\}_{avg}]_{f_i}}{\sum_{i=1}^{7}[\{D_{KL}(A,B)\}_{avg}]_{f_i}}, \tag{3.15}$$

The weights assigned to each of the seven features according to their potential to classify different sonorant sounds are also shown in Table 3.3. Thus a competent representation of degree of sonority associated with a sound unit is derived in this work.

The overall block diagram of the proposed work is depicted in Figure 3.9. Three different features are derived using the knowledge of vocal-tract system, excitation source and suprasegmental aspects of sonorants. To derive the feature of vocal-tract system, ZTW is applied around each epoch location of the speech signal. For the windowed segments of 5 ms of speech signal around each epoch, HNGD spectrum is derived. The excitation source feature is derived from the HE of LP residual of speech signal, which is peak to sidelobe ratio around each peak of the HE . In contrast to these two evidence, the suprasegmental feature is derived from correlation of the speech signal over ten pitch periods. The three evidence are weighted and fused together to derive the seven-dimensional sonority evidence (vocal-tract system (five-dimension), excitation source (one-dimension) and suprasegmental feature (one-dimension)). The evidence is further utilized in the task of sonorant/non-sonorant classification and multiclass sonorant classification to verify the efficacy of the proposed feature.

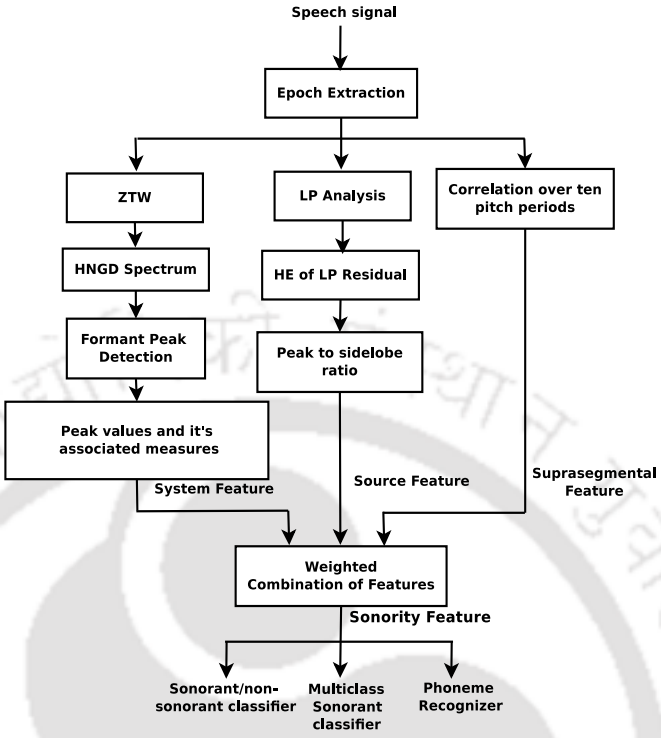


**Figure 3.9:** Overall block diagram of the proposed sonority feature extraction from speech signal, where vocal-tract system, excitation source and suprasegmental features are derived from HNGD spectrum, HE of LP residual and speech signal, respectively. These features are combined to derive the sonority feature.

## 3.6 Experimental evaluation

The distributions of the proposed sonority evidence correlate well with the sonority hierarchy as can be observed from Figure 3.3 and Table 3.2. To establish the efficacy of the proposed seven-dimensional sonority feature vector in representing sonority associated with a sound unit, the following classification experiments are performed.

### 3.6.1 Sonorant/non-sonorant classification

The first level of classification that exploits the usefulness of prospective features representing sonority is sonorant/non-sonorant classification. In [115], it has been demonstrated that the attributes derived from speech signal like zero frequency resonator (ZFR) signal energy, slope of ZFR signal around epoch locations and dominant resonance frequency (DRF), can be used for the task of sonorant/non-sonorant segmentation, both at frame and epoch levels. An hierarchical algorithm is used for the classification task. To compare the effectiveness of the proposed feature with the features used in [115], a sonorant/non-sonorant classifier using support vector machine (SVM) (with radial

**Table 3.4:** Comparison of performance of proposed feature (using SVM) and existing feature using hierarchical algorithm (within braces) in sonorant/non-sonorant segmentation on utterances from TIMIT database in both clean speech and noisy speech across different SNR levels.

| SNR | Proposed feature (Existing Feature) | | | | | |
|---|---|---|---|---|---|---|
| | Epoch based results | | | Frame based results | | |
| | Acc(%) | TPR(%) | FAR(%) | Acc(%) | TPR(%) | FAR(%) |
| clean | 96.3 (93.9) | 98.5 (94.5) | 5.5 (7.5) | 95.0 (92.8) | 97.3 (93.6) | 6.8 (8.0) |
| 30 dB | 96.0 (93.9) | 98.4 (94.5) | 5.7 (7.6) | 94.8 (92.8) | 96.7 (93.6) | 7.6 (8.0) |
| 20 dB | 95.4 (94.39) | 96.6 (94.4) | 6.2 (7.7) | 93.5 (92.7) | 95.8 (93.5) | 8.0 (8.1) |
| 10 dB | 95.0 (93.4) | 94.3 (94.0) | 6.8 (8.5) | 93.4 (92.1) | 95.3 (92.9) | 8.4 (9.0) |
| 5 dB | 93.4 (92.4) | 93.6 (93.1) | 8.3 (8.9) | 93.0 (91.0) | 92.8 (91.9) | 9.3 (9.5) |
| 0 dB | 90.5 (90.7) | 91.4 (91.0) | 9.5 (9.9) | 90.0 (89.6) | 90.7 (89.9) | 10.3 (10.6) |

basis function (RBF) kernel, $c = 16$, $\gamma = 4$) is developed using the proposed sonority feature vector. The training and testing feature vectors are derived from all SI and SX utterances of TIMIT train and test database, respectively. This is followed by feature normalization to make the feature values within 0 to 1 range. Similar normalization is performed in training and testing of clean and noisy speech. The same SVM classifier trained using clean speech is employed in the testing of sentences mixed with white noise across various SNR levels.

To demonstrate the robustness of the features for classification, the performance evaluation parameters used are: number of epochs/frames correctly detected in the sonorant regions (true positive rate (TPR)), number of spurious epochs/frames hypothesized in the non-sonorant regions (false alarm rate (FAR)) and total number of correctly detected epochs/frames in both the sonorant and non-sonorant regions (accuracy (Acc)). As shown in Table 3.4, the proposed feature can segment sonorant regions with more accuracy compared to the existing method (within braces). Therefore, the proposed feature has better ability to classify sonorant/non-sonorant segments from the given speech signal.

### 3.6.2 Classification of sonorant sounds into different classes

The primary motivation of this work is to derive feature to characterize the degree of sonority associated with a sound unit. The straightforward way to validate this would be to develop a multi-class sonorant classifier, where each class represents different sonorant sounds (low-vowels, mid-vowels, high-vowels, liquids, glides and nasals). As described in Section 3.5, the proposed seven-dimensional sonority feature is derived for each class of sonorant sound for the entire TIMIT test database. This

is followed by normalization to make the feature value within the range of 0 to 1. Individual feature dimension consists of a single value at each epoch location. A six-class SVM classifier (with RBF kernel, $c = 256$, $\gamma = 16$) has been developed using the normalized sonority feature vector. Values of parameters, $c$ and $\gamma$ are set using train-test 5-fold cross validation for the entire TIMIT test database. For the optimized value of $c$ and $\gamma$, the six-class SVM model is trained using randomly chosen 80% of TIMIT-test data. The rest 20% data is used for testing.

The classification accuracy of each class and confusion among different classes are reported in Table 3.5. The average accuracy achieved is 66.55%. The accuracy is observed to be the lowest for the liquids and highest for the nasals. It can be interpreted from Table 3.5 that, 14.41% of low-vowels are misclassified as mid-vowels. This is due to the fact that the properties of low-vowels and mid-vowels are close to each other. Moreover, as observed from Figure 3.3, formant bandwidth and feature of excitation source exhibit overlap between the two classes. As the height of the tongue body for mid-vowels is intermediate between that of the high and low-vowels, it affects the constriction size and length. This in-turn alters the VTS evidence.

Although the vocal-tract constriction in case of the liquids is narrower than the glides, resulting in wider $F_1$ bandwidth for liquids, the length of constriction is shorter in case of the liquids. This increases $F_1$ for liquids and introduces confusion between glides and liquids. Thus there is possibility of confusion of liquids with low-vowels and mid-vowels. This is evident from $1^{st}$, $2^{nd}$ and $5^{th}$ rows of Table 3.5. The common attribute of the liquids with the vowels is that, in both cases air flows through the constriction without pressure drop. As a result, the vocal-folds continue to vibrate in the period of constriction. In the distribution of feature of excitation source in Figure 3.3(f), confusion between glides and liquids can be apparently observed. As reported in Table 3.5, majority of misclassification of high-vowels is due to the confusion with mid-vowels and glides. The configuration of vocal-tract for glides may also change based on the preceding vowels. A glide adjacent to high-vowel is produced with more constricted structure compared to the one preceded or followed by a low-vowel. Therefore, when a glide is contiguous with low-vowel or mid-vowel, due to less constriction, $F_1$ may increase. The bandwidth may decrease compared to the glide that is adjacent with a high-vowel.

The proposed features are analogous to formant based measures and do not use the temporal information of nearby sounds. Therefore, there is a possibility of misclassification of each category to its adjacent category of sound in the sonority hierarchy. It is notable from Figure 3.3 that, compared

**Table 3.5:** Classification accuracy (epoch level) of different sonorant sounds from TIMIT test database using SVM ($c = 256, \gamma = 16$) obtained by employing the proposed seven-dimensional sonority feature.

| Category | % Accuracy | | | | | |
|---|---|---|---|---|---|---|
| | **Low-vowels** | **Mid-vowels** | **High-vowels** | **Glides** | **Liquids** | **Nasals** |
| **Low-vowels** | **68.0** | 14.4 | 4.1 | 2.8 | 9.2 | 1.5 |
| **Mid-vowels** | 9.8 | **63.9** | 9.2 | 4.5 | 10.9 | 1.7 |
| **High-vowels** | 1.7 | 10.3 | **67.3** | 11.7 | 4.6 | 4.4 |
| **Glides** | 1.4 | 6.4 | 12.7 | **59.4** | 6.7 | 13.4 |
| **Liquids** | 7.2 | 13.3 | 9.9 | 8.5 | **55.9** | 5.2 |
| **Nasals** | 0.5 | 1.9 | 3.4 | 1.5 | 7.9 | **84.8** |

to other categories of sonorants, the distribution corresponding to nasals has less overlap with other distributions. Only in case of suprasegmental feature in Figure 3.3(g), some confusion of the nasals with other categories is observable. This correlates with the highest accuracy of the nasals as reported in Table 3.5. As the front part of vocal-tract is completely closed during nasal murmur, the first formant frequency and its prominence eventually decreases with a weak second formant followed by an extended valley in the VTS. This is more contrasting with other sonorants. However, the common acoustic behavior of nasals and glides is that, the vocal-fold does not change the vibration pattern before and after the constriction happens. Based on this discussion and the classification accuracy of sonorants presented in Table 3.5, it can be inferred that the proposed features have ability to quantify sonority level associated with a sound unit. Although, some aspects of the speech signal corresponding to a specific category of sound unit may vary based on the adjacent sound units present.

To further demonstrate the ability of the proposed features for discriminating different sonorant classes, in addition to MFCC, two SVM classifiers (one using sonority feature and the other using MFCC feature) are fused at score level [123]. For this thirteen-dimensional MFCC feature is used to develop another six class SVM classifier (with RBF kernel, $c = 2, \gamma = 4$), where $c$ and $\gamma$ values are set using train-test 5-fold cross validation for entire TIMIT test database. For the optimized values of $c$ and $\gamma$, the six-class SVM model is trained. The randomly chosen 80% of TIMIT-test data is used for training and rest 20% is used for testing. The average accuracy of the MFCC based classifier is found to be 80.41%. The detailed performance for each class can be seen in Table 3.6 (within braces). As there are 6 classes, each of the classifiers using MFCC and sonority feature will produce 6 posterior probabilities for each feature vector.

**Table 3.6:** Classification accuracy of different sonorant segments (frame level) from TIMIT database using combined sonority and MFCC feature based SVM classifier. Classification accuracy obtained using only MFCC feature vector is shown within braces ($c = 2, \gamma = 4$).

| Category | % Accuracy | | | | | |
|---|---|---|---|---|---|---|
| | **Low-vowels** | **Mid-vowels** | **High-vowels** | **Glides** | **Liquids** | **Nasals** |
| **Low-vowels** | **86.3 (78.7)** | 6.5 (13.6) | 3.2 (3.4) | 0.2 (0.4) | 3.8 (2.7) | 0.0 (0.9) |
| **Mid-vowels** | 10.8 (10.4) | **75.4 (68.8)** | 5.3 (11.4) | 0.7 (1.6) | 7.8 (7.3) | 0.0 (0.5) |
| **High-vowels** | 0.5 (0.4) | 7.3 (9.6) | **85.2 (80.8)** | 5.8 (5.1) | 0.9 (3.1) | 0.3 (0.4) |
| **Glides** | 0.1 (0.3) | 2.0 (1.1) | 6.8 (8.8) | **83.5 (80.7)** | 5.4 (5.3) | 2.2 (3.8) |
| **Liquids** | 3.6 (4.1) | 6.5 (5.3) | 1.5 (2.8) | 5.8 (4.5) | **80.7 (78.8)** | 1.9 (4.5) |
| **Nasals** | 0.2 (0.2) | 0.8 (0.5) | 0.8 (1.8) | 0.9 (1.2) | 1.3 (1.7) | **96.0 (94.7)** |

For the sonority based classifier, the posterior probability scores corresponding to epochs within one frame are averaged to derive single probability score corresponding to each class for each frame. The mean value of probabilities of the two classifiers for each class corresponding to each frame is calculated to derive the fused probability score. The class with maximum average probability score is considered as final output of the combined classifier. The resultant accuracy of the combined classifier is found to be 84.51%, which is 80.41% when only MFCC feature is used. The classification accuracy for each class using the combined classifier and only MFCC based classifier is shown in Table 3.6 for comparison. By comparing both % accuracy values in Table 3.6, an absolute improvement of 4.1% can be observed when the two classifiers are fused. For each of the classes, along with improvement in classification, reduction in confusion among different sonorant classes can also be observed. It is interesting to observe from Table 3.6 that, with increase in correct classification of each class, the percentage of confusion with other classes is reduced for most of the cases.

To study individual performances of sonorant classification for male and female, we have developed two sonorant classifiers using SVM (with RBF kernel, $c = 256$, $\gamma = 16$) for male and female utterances from TIMIT test database. For developing each classifier 80% of male/female data is used for training and rest 20% is used for testing. The average accuracy of the six class sonorant classification is found to be is 68.4% for male and 65.6% for female. The relatively poor performance for the female case may be attributed to the associated high non-stationarity nature.

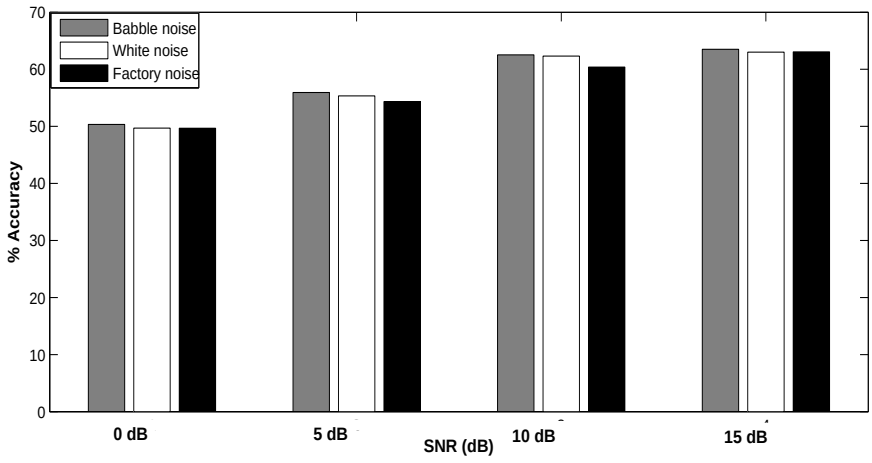


**Figure 3.10:** Bar plot representing average % accuracy for SVM based six-class sonorant segment classification in presence of different types of noise with different SNR levels.

### 3.6.3 Effect of noise on sonority feature

In order to analyze the impact of noise on the proposed features, the classifier trained using features derived from the clean speech is employed for testing of noisy cases. The test features are derived after addition of different kinds of noises (babble noise, factory noise, white noise) to the speech signal at different SNR levels (0 dB, 5 dB, 10 dB, 15 dB). The average accuracy of the classifier for different types and levels of noise is shown as bar plot in Figure 3.10. It can be observed that % accuracy significantly decreases in case of 0 dB and 5 dB SNR levels. Whereas, for 10 dB and 15 dB cases, % accuracy is less effected. Further, to analyze the robustness of each of the system, source and suprasegmental features, three six-class SVM classifiers are developed using individual features derived from clean speech. The test features are derived after adding different levels of babble noise with the speech signal.

Figure 3.11 demonstrates the degradation of % accuracy of the three classifiers with the increased noise level. This depicts that the suprasegmental feature is more affected due to noise compared to the features of vocal-tract system and excitation source. This may be due to the reason that, suprasegmental feature is directly derived from the speech signal by measuring correlation over successive pitch periods. Furthermore, it is not derived in synchrony with the glottal closed phase, which may be less susceptible to degradation due to noise. The features of vocal-tract system are derived from the HNGD spectrum which is reported to be less affected by different types of noise [106]. This hap-

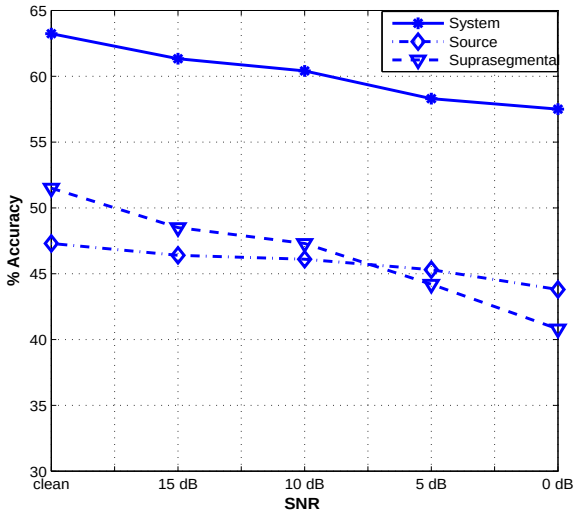


**Figure 3.11:** Average % accuracy of six-class sonorant classifier using each of the system, source and suprasegmental features in with respect to different levels of noise.

pens due to the short and tapered window used in obtaining the HNGD spectrum. For deriving the feature of excitation source, the samples corresponding to glottal closed phase around epoch locations is accessed. Hence this feature is also found to be less affected by noise.

The above experiments validate the effectiveness of the proposed feature in discriminating the sonorant sounds or characterization of degree of sonority from given the speech signal, without the knowledge of labels. Further, we demonstrate its usefulness in different speech processing applications.

## 3.7 Applications of sonority feature

The proposed sonority feature can be used in different speech processing applications as it has ability to differentiate between the sonorant category to some extent. To establish the efficacy of the feature, its implementation in phoneme recognition and VOP detection is presented in this section.

### 3.7.1 Sonority as a feature for phoneme recognizer

The proposed sonority feature may be helpful to improve the performance of a phoneme recognizer by incorporating additional information to reduce confusion among different sonorants. In this regard, phoneme recognition framework for TIMIT database is developed in Kaldi toolkit [124, 125], where DNN based acoustic modeling is implemented [126]. In addition to traditional MFCC feature, proposed seven-dimensional weighted sonority feature is employed for developing the recognizer. The proposed

**Table 3.7:** Phone error rate (PER) for DNN based phoneme recognizer using MFCC and (MFCC + Sonority) feature.

| Evaluation on | PER(%) | |
|---|---|---|
| | MFCC | MFCC+sonority feature |
| Test set | 22.7 | 21.4 |
| Dev set | 21.2 | 20.3 |

feature is epoch synchronous. In order to match dimension with MFCC at frame level, average value of feature corresponding to epochs within one frame is calculated. It is then appended with the thirteen-dimensional MFCC feature resulting in a twenty-dimension feature vector. A bigram phoneme language model is created from the training set is incorporated in the recognizer.

The 61 phonemes are mapped into 39 phonemes for the training and testing, the acoustic model is an HMM-DNN hybrid model. The training set contains 3,696 sentences from 462 speakers. The development set contains 400 sentences from 50 speakers. Core test set is also used as test set, which contains 192 sentences from 24 speakers. The number of hidden layers used is 2. It is reported in Kaldi documentation that 4 hidden layers are effective when 100 hours of speech data is available. An initial learning rate of 0.015 is selected which is reduced to 0.002 in 20 epochs. Additional 10 epochs are employed after reducing the learning rate to 0.002. Kaldi employs a preconditioned form of stochastic gradient descent. A matrix-valued learning rate is employed instead of using a scalar learning rate in order to reduce the learning rate in dimensions, where the derivatives have a high variance. This is in order to control instability and stop the parameters moving too fast in any one direction.

The overall performance of the baseline phoneme recognizer using MFCC as feature and using additional proposed feature (MFCC + sonority) is shown in Table 3.7 in terms phone error rate (% PER). It is improved while using the proposed features along with the MFCC. Also, the improvement in case of different sonorant phones in terms of accuracy (%) and correct (%) identification is shown in the bar plot of Figure 3.12. The performance increases after using the proposed sonority features. It is observed that with the addition of proposed evidence, insertion and substitution of sonorant phones decreases significantly, whereas the reduction in deletion is comparatively less. However, the confusion among different classes of sonorant phones is analyzed in terms of % substitution. It seems to reduce while employing the proposed feature in addition to MFCC as shown in Table 3.8. Thus, the sonority feature is found be potent for application in the phoneme recognition.

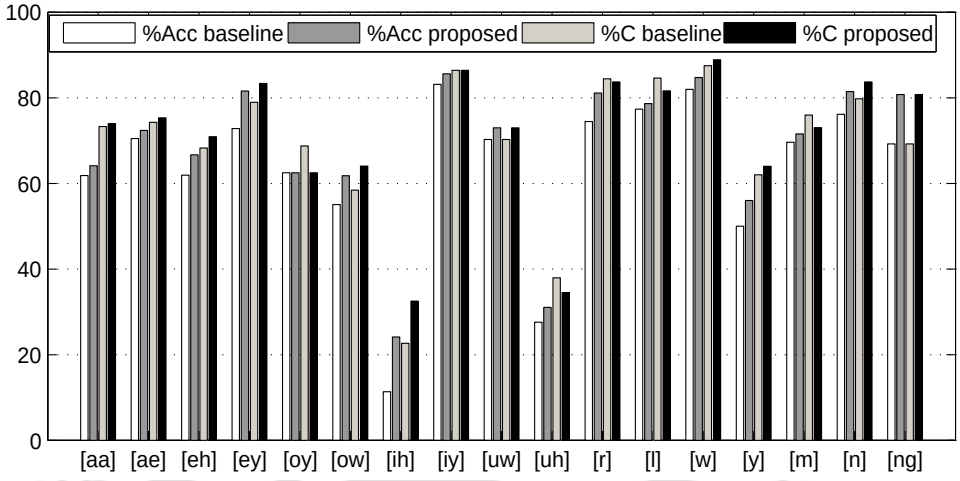


**Figure 3.12:** Correction percentage (%C) and accuracy (%Acc), before and after appending the sonority for various sonorant phones of TIMIT.

**Table 3.8:** % substitution of different sonorant phones before and after appending the proposed sonority evidence for various sonorant phones of TIMIT. Baseline result using MFCC is shown braces.

| Category | % Substitution | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Low-vowels** | **Mid-vowels** | **High-vowels** | **Glides** | **Liquids** | **Nasals** | **Total** |
| Low-vowels | 4.0 (4.1) | 5.5 (6.1) | 6.8 (7.0) | 0.0 (0.1) | 0.5 (0.5) | 0.3 (0.4) | **17.1 (18.2)** |
| Mid-vowels | 3.5 (4.9) | 1.3 (1.3) | 4.8 (4.9) | 0.0 (0.0) | 0.8 (1.2) | 0.1 (0.1) | **10.5 (12.4)** |
| High-vowels | 4.3 (4.7) | 1.9 (2.1) | 4.7 (5.3) | 0.4 (0.9) | 0.2 (0.2) | 0.2 (0.7) | **11.7 (13.9)** |
| Glides | 0.0 (0.0) | 0.0 (0.0) | 1.0 (1.8) | 0.3 (0.3) | 0.3 (0.5) | 0.5 (0.8) | **2.1 (3.4)** |
| Liquids | 0.6 (1.2) | 0.8 (0.9) | 0.3 (0.2) | 0.1 (0.3) | 0.1 (0.1) | 0.4 (0.7) | **2.3 (3.4)** |
| Nasals | 0.5 (0.5) | 0.2 (0.3) | 0.5 (0.5) | 0.1 (0.2) | 0.2 (0.3) | 3.6 (3.9) | **5.1 (5.7)** |

### 3.7.2 Sonority in vowel onset point detection

The VOP refers to the starting event of a vowel, that may be reflected in different aspects of speech signal. Vowels are the most sonorant sounds followed by semivowels, nasals, voiced fricatives, voiced stops. Detection of VOP is challenging in case of continuous speech and recent methods for detecting VOPs have high errors when the vowel is preceded by other sonorant sounds [127], as they are not capable to discriminate well between them. As the sonority feature has the capability to discriminate among different sonorant sound units, it can be used to reduce the confusion among onset of vowels and that of other sonorant sound units.

Figure 3.13 shows the steps involved in VOP detection using the sonority evidence and compares with existing VOP detection features in [127]. The features $f_1$, $f_2$, $f_3$, $f_4$ are directly proportional to

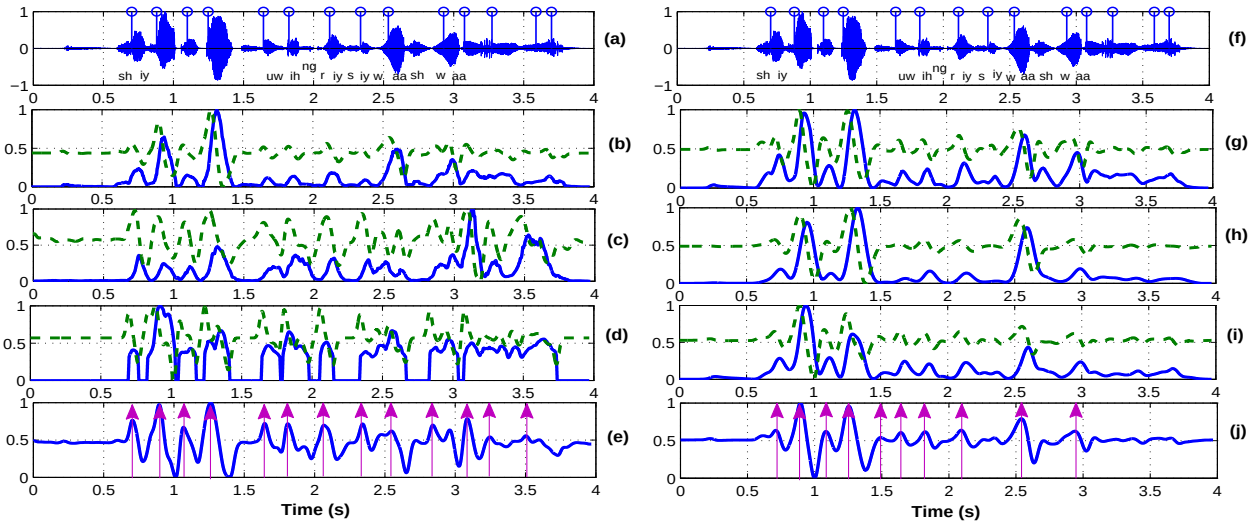


**Figure 3.13:** Steps involved in VOP detection using sonority evidence for the utterance "she had your dark suit in greasy wash water all year" taken from TIMIT database, using sonority feature and existing feature. (a) Speech signal with reference VOPs; VOP evidence from (b) combined feature of vocal-tract system, (c) feature of exciation source, (d) suprasegmental feature (the dotted contour in (a), (b), (c) are corresponding FOGD convolved features); (e) combination of FOGD convolved signals in (a), (b), (c); (f) speech signal with reference VOPs; VOP evidence from (g) energy of spectral peaks, (h) modulation spectrum energy, (i) smoothed Hilbert envelope (the dotted contour in (g), (h), (i) are corresponding FOGD convolved features); (j) VOP evidence from combination of FOGD convolved signals in (g), (h), (i).

the sonority associated with a sound unit, whereas $f_5$ is inversely proportional to the same. The first four dimensions are normalized and summed up; the resultant feature is added with normalized inverse of $f_5$ to derive the combined vocal-tract evidence representing sonority which is shown in Figure 3.13(b) with solid line. The normalized SoE derived from HE of LP residual is illustrated in Figure 3.13(c) with solid line. The derived suprasegmental feature seems to have more temporal variation which effects in VOP detection. Hence in the normalized suprasegmental feature, the values less than 0.2 are made zero and resultant signal is smoothed over an window of 50 ms which is more than one pitch period. The contour of normalized post-processed suprasegmental feature for an utterance from TIMIT database is shown in Figure 3.13(d), which carries significantly different information compared to features of vocal-tract system and excitation source.

In all the three features demonstrated in Figure 3.13(b), (c), (d), variation in the feature value can be observed with change of sound unit along the temporal axis. To track these changes each feature is convolved with a first order Gaussian differentiator (FOGD) of length 100 ms (800 samples for 8 kHz) and standard deviation as one sixth of the window length (134 for 8 kHz) [127]. The convolved normalized signals corresponding to each feature are shown with dotted line in Figure 3.13(b), (c),

(d). Each of the convolved signals show peaks at the VOPs, which are marked by solid lines in Figure 3.13(a) over the speech signal.

The convolved signals corresponding to each of VTS, excitation source and suprasegmental features are added to derive combined evidence of sonority feature as shown in Figure 3.13(e). The peaks in the combined evidence in Figure 3.13(e) represent the VOPs, which are detected by finding the maximum value between two successive positive to negative zero crossings with some threshold to eliminate the spurious peaks. It can be observed from Figure 3.13(a) and (e) that, the sonority feature has potentiality to correctly characterize VOP in continuous speech. After 3 $s$ in Figure 3.13(a) although the energy of the speech signal seems to be very low at vowel regions, the combined sonority evidence is able detect the VOPs. Moreover, it is significant to observe from Figure 3.13(b), (c), (d) that each evidence carries discriminative information along the utterance irrespective of the wideband energy associated with a particular instant of the speech signal. For example, around the instant 3.5 $s$ in the speech signal shown in Figure 3.13(a), the feature of VTS exhibits lower values with minute variation. On the other hand, both excitation source and suprasegmental features show higher values and prominent variation around 3.5 $s$. Due to the combined effect of the features with discriminative information, the VOP at around 3.5 $s$ is correctly detected.

The VOP detection evidence used in [127] are SHE, MSE and spectral peaks energy, which are shown in Figure 3.13(g), (h), (i), respectively for the utterance in Figure 3.13(f). It infers that all three features show less variation with the change in speech sound in the continuous utterance. Furthermore, for low energy regions the variation in feature value is less. The corresponding FOGD convolved signal is depicted in dotted line over each feature which shows peaks with less strength compared to that of sonority features in Figure 3.13(b), (c), (d). Around 0.5 $s$ in the speech signal, [sh] is followed by a high-vowel [iy] as shown in Figure 3.13(a). At transition point from [sh] to [iy], the sonority features seem to have a sharp transition compared to features shown in Figure 3.13(g), (h), (i). Similar observation can be made for the speech segments of around 2 $s$, where [ih], [ng], [r], [iy] sounds are continuously uttered which are sonorants. The variation of each sonority evidence can be distinctly observed compared to that of existing evidence. Although the VOPs seem to be detected in this case using existing features in Figure 3.13(j), the detected VOP locations are apart from actual VOPs and there is a chance of missing the VOPs. Due to the very low variation in the existing features, authors of [127] have enhanced peaks in the features and then convolved with the FOGD. Although this post

**Table 3.9:** Performance of sonority evidence in VOP detection for 593 sentences comprising of 6818 VOPs. Baseline result is shown within braces.

| **Evaluation Parameter** | **Proposed Method (Baseline Method) (VOPs within ms)** | | | |
|---|---|---|---|---|
| | ±10 | ±20 | ±30 | ±40 |
| Detection Rate (%) | 73.0 (60.2) | 77.5 (62.82) | 82.4 (68.8) | 92.4 (85.7) |
| Spurious Rate (%) | 32.8 (40.3) | 25.7 (32.9) | 23.6 (28.5) | 13.8 (21.1) |

**Table 3.10:** Performance of sonority evidence in VOP detection for different CV units with different tolerance levels. Baseline result is shown within braces. Among 6818 VOPs, 1916 semi-vowels, 1475 fricatives, 803 nasals, 80 affricates and 2544 stops are present.

| **Type of CV unit** | **Detection Rate (%) of proposed Method (Baseline Method) (VOPs within ms)** | | | |
|---|---|---|---|---|
| | ±10 | ±20 | ±30 | ±40 |
| Semi-vowels | **35.77 (22.55)** | **42.27 (25.96)** | **46.99 (28.86)** | **82.58 (66.84)** |
| Fricatives | 84.84 (71.97) | 85.67 (74.79) | 91.40 (78.38) | 96.95 (92.01) |
| Nasals | 73.31 (70.54) | 82.83 (73.04) | 86.44 (78.21) | 89.17 (84.73) |
| Affricatives | 83.20 (62.72) | 87.59 (64.59) | 94.50 (79.67) | 98.30 (92.59) |
| Stops | 88.08 (73.02) | 90.48 (75.73) | 92.50 (79.29) | 95.22 (92.53) |

processing adds to detection of some more VOPs, it may introduce many spurious VOPs. Due to these facts, use of combined sonority feature which has potential to discriminate between different sonorant sounds, found to give better performance in case of VOP detection from continuous speech.

For comparison of the sonority feature in VOP detection task with the existing methods, a set of 593 sentences from TIMIT test database (358 male voiced and 235 female voiced) having 6818 number of VOPs is used. These VOPs are manually adjusted by observing spectrograms and waveforms. The features used in [127] for VOP detection are distinctive from the sonority evidence. Detection rate and spurious rate for each of proposed and baseline methods (within braces) are demonstrated in Table 3.9 for tolerance of ±10 ms, ±20 ms, ±30 ms and ±40 ms around the true VOPs, which shows significant improvement while using sonority feature. It can be observed that, improvement is more significant in ±30 ms tolerance with 13.6% increase in detection rate. Moreover, when tolerance level changes from ±30 ms to ±40 ms, there is highest change in detection rate compared to other tolerance levels.

Different types of consonants are present in the CV units of 6818 VOPs which can be classified into five categories according to the consonant type. Among 6818 VOPs, 1916 semi-vowels, 1475 fricatives,

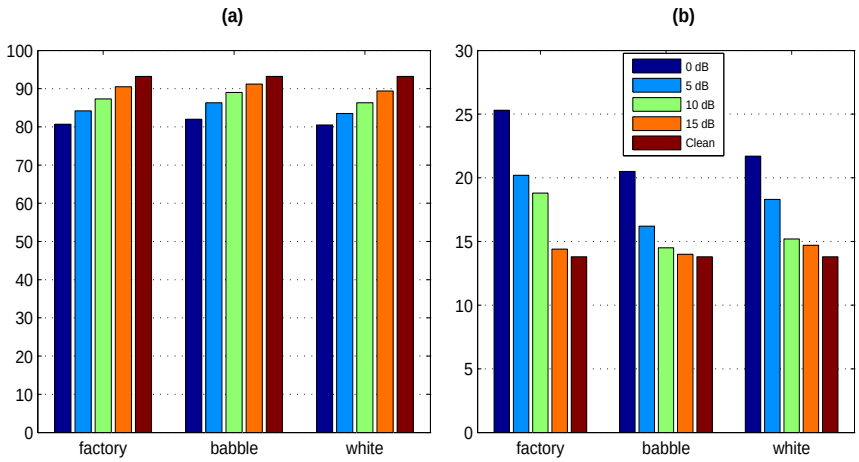


**Figure 3.14:** Bar plot representing (a) detection rate (%), (b) spurious rate (%) of VOP detection within ±40 ms tolerance for different types and levels of noise.

803 nasals, 80 affricates and 2544 stops are present. The detection rate for all the five different categories for both baseline and proposed methods are reported in Table 3.10. It can be inferred that most of the improvement using proposed method is obtained in terms of vowels which have preceding semi-vowels, although for this category detection rate is lowest. The average improvement in case of semivowels over all tolerance levels is 15.9%. Among all the CV units, most of the VOPs preceded by affricates and fricatives are correctly detected. The least detection accuracy is still in VOPs preceded by semivowels and nasals although the use of sonority evidence has increased the same to some extent. It is important to notice that, along with the increase in detection rate for VOPs preceding semivowels and nasals, it also increases in case of fricatives, affricates and stops while using sonority feature. In [127], energies of first 10 largest peaks of VTS are considered as a feature, but for some high energy fricatives, high amplitude peaks may be present at higher frequency region in the VTS which will yield high values of the feature in case of fricatives also, leading to miss detection of the following VOP in that CV unit. Whereas, in the sonority feature the statistics of first three formant peaks are considered which gives low values in obstruents. Moreover, in fricatives irregular sequence of peaks in HE of LP residual may be present, which will be manifested as higher values in the feature SHE used in [127]. In excitation source information of sonority feature, relation among peak of HE of LP residual at GCI and nearby peaks is considered which will give low values at regions with irregular peaks. These facts may lead to improved detection rate in case of vowels preceded by fricatives, affricates and stops. Along with the VOP detection accuracy, another necessary requirement

is robustness of the detection method in noisy scenario. To demonstrate noise robustness of sonority feature in VOP detection, speech signal corresponding to same 593 sentences are added with factory noise, babble noise and white noise, each at different levels of SNR (0dB, 5dB, 10dB, 15dB) and same VOP detection algorithm is applied. Performance of VOP detection by sonority feature in degraded condition is shown in Figure 3.14 with some acceptable level of reduction in performance. Further analysis shows that the noise robustness in the sonority feature is due to the fact that, both VTS and excitation source features are extracted from glottal closed phase, which is less affected by noise. Moreover, the HNGD spectrum is robust to noise as in [106]. As in suprasegmental feature, the correlation is computed over samples of speech signal, it may have some effect of noise.

## 3.8 Summary

In this chapter, an effort is made to extract a feature from speech signal, which can represent the degree of sonority associated with a sound unit. For this task, different characteristics of the sonorant sounds reflected in the speech signal are analyzed. Consequently, features based on the vocal-tract system, excitation source and suprasegmental aspects are derived. These features correlate with less vocal-tract constriction, glottal vibration and periodicity properties of sonorant sounds. To justify, whether each of the proposed features can represent the level of sonority, distributions for feature values are shown for different sonorant sounds along the sonority hierarchy. Each of the proposed features shows increasing/decreasing trend in feature value with the increase in sonority. The proposed seven-dimensional sonority feature is used in sonorant/nonsonorant classification, different sonorant sounds classification and is found to be potential for the same. It is also shown to be useful for the phoneme recognition and VOP detection applications. In the future, we may focus on exploring evidence, which can reduce the confusion among adjacent classes in the sonority hierarchy.

# 4

# Dynamic Post-filtering using Sonority Information

Contents

### Objective

*The synthesized speech obtained from SPSS lacks naturalness, which is often attributed to the over-smoothing of generated parameter sequences, both in the temporal and spectral domains. The extent of the deviation corresponding to different source and spectral parameters between natural and synthesized speech may vary with the classes of speech sounds. Therefore, introducing different PF factors for various speech sound categories may emphasize the fine structure of the parameters and make it closer to that of natural. The source parameters considered for PF are $F_0$ and SoE, and the spectral parameters are first five spectral peaks and valleys. For each of them, different post-filters are trained to obtain a better spectral resolution, using a background dataset of parallel utterances corresponding to natural and synthesized speech. An SVM classifier is trained using the sonority feature to classify the speech sound to different sonorant categories. During synthesis, based on the class information of each frame obtained from the SVM classifier, the decision is made to use the corresponding post-filter mean and variance. Accordingly, modified values of the source, and spectral parameters are obtained. This dynamic modification of the source and spectral parameters helps to reduce the over-smoothing for improving the quality of synthesized speech. The subjective evaluation shows an increase in mean opinion score from 2.38 to 3.27 after employing the proposed PF method. The proposed method achieves 61% preference compared to the recent DNN based PF method. Apart from the conventional source and spectral parameters, spectral tilt is an important aspect of speech perception. Another PF method that attempts to compensate the deviation in spectral tilt between natural and synthesized speech is also proposed in this chapter. This results in improving naturalness, intelligibility and speaker similarity of synthesized speech.*

## 4.1 Introduction

The SPSS is the state-of-the-art technique for speech synthesis that provides high intelligibility and flexibility. It possesses the ability to change voice characteristics, speaking styles and emotions by transforming model parameters using techniques like adaptation, interpolation, eigenvoice, and multiple regression [1, 2, 128, 129]. The SPSS has become one of the most widely used methods due to advantages like the wider range of available units, better multilingual support, and robustness regarding the recording of training data compared to USS [3–5, 130]. In SPSS, HMMs and DNN based models are widely approaches. The naturalness of speech generated using SPSS is still not up to the

level of USS and over-smoothing of generated parameter sequences is one of the reasons [5]. In HMM based speech synthesis, the speech parameter generation algorithm generates excitation source and spectral parameters from HMMs to maximize their output probabilities, under constraints between static and dynamic features [6]. The introduction of dynamic feature constraint leads to the generation of over-smoothed feature trajectories. This over-smoothing results in loss of dynamic variation and fine structure of the generated parameter sequences leading to *muffled* quality of synthesized speech.

In this chapter, we have also analyzed of divergence in tilt or slope of VTS extracted from natural and synthesized speech. Spectral tilt represents energy distribution of the spectrum at different frequency regions [131]. Suppressing higher harmonics i.e. strongly negative slope may make sounds muffled, while enhancing higher harmonics may result in clear sounds. There are several studies in the literature of Lombard speech relating to spectral slope, which infer that average spectral tilt in Lombard speech is lesser than that of speech produced in quiet [84–86]. However, methods employed for changing the tilt and their impact on different categories of sounds are not studied abundantly. The spectral tilt difference between synthesized and natural speech yet needs extensive study.

There are several attempts in the literature to alleviate the over-smoothing effect from the generated parameter sequences by employing PF methods. These methods are applied to enhance the generated source and spectral parameters, before passing them through the vocoder for synthesis. One of the basic approach to achieve this is the enhancement of spectral peaks and valleys which are smoothed out due to the statistical averaging. The MFCCs generated from SPSS are modified to enhance the spectrum peaks and valleys in [74], which was originally proposed in [75] to implement in the area of speech coding. It helps to alleviate the effect of spectral smoothing within a frame. Other similar types of methods for formant enhancement are carried out by using LSPs [76] and LPCs [77]. Nevertheless, the degree of variation of the generated parameters over the temporal axis is not improved in these cases. Moreover, in these approaches, all the spectral peaks are enhanced by a constant factor, irrespective of their positions in the spectrum.

Another most frequently used method to alleviate over-smoothing of parameter contour is GV parameter generation algorithm [23]. In conventional SPSS, the parameter generation algorithm generates a parameter trajectory of static features by maximizing the likelihood of a sequence of given HSMMs states for static and dynamic features under an explicit constraint. On the other hand in case of GV method, along with this another likelihood term is introduced, that reflects the dynamic

range (variance) of each dimension of the parameter sequence at the utterance level. The inclusion of GV during parameter generation increases the dynamic range of each dimension of the parameter sequence at the utterance level, across frames in the time domain. An extended algorithm of GV applied in spectral domain is proposed in [73]. In case of GV, although the variance over an utterance is maintained, the values and behavior of parameters may not be close to the natural counterpart.

In [78] histogram equalization method is used to construct the emphasis rule, which converts the generated LSPs to have similar behavior as the natural LSPs. The frequency-dependent temporal modulations of the parameter trajectories can be explicitly enhanced with the MS based PF method [79]. Here, MS represents the power spectrum of the parameter trajectory. This post-filter tries to make generated power spectrum of the parameter trajectory similar as that of its natural counterpart. Recently, DNN based PF is proposed in [81], which models the conditional probability of the spectrum of a natural speech given that of synthesized speech. This data-driven post-filter helps to retain information in a higher-dimensional spectral domain. This method is automatic and therefore does not employ any detailed investigation of the acoustic properties of generated and natural speech parameter sequences. Even though DNN helps, handcrafted features assist in a limited data case. Moreover, types of different classes are affecting more with the proposed dynamic PF method, which is not possible in DNN based methods. A radial PF method in the cepstral domain is proposed in [82], which applies different PF factors to low and high frequencies, with adjustable cut-off frequency. This method helps to achieve a higher spectral resolution, although it does not take into account the temporal variation over consecutive frames over an utterance.

### 4.1.1 Motivation for the proposed method

In case of MS based PF method, each dimension of MS of natural and generated parameters are represented by single mean and standard deviation. Similarly, in GV the variance is considered over the entire utterance irrespective of the sound units present. These methods do not consider the fact that, the characteristics of the speech parameters may extensively vary based on broad categories of the sound units. Representation of these parameters using single mean and variance may not be able to reflect the intact variabilities and fine structure, with respect to different categories sound units. As discussed in Chapter 2, even after incorporation of GV and MS based PF there is room for improvement of the generated parameter contours in both temporal and spectral domain. In order to achieve naturalness and intelligibility of synthesized speech close to that of the of natural speech,

it is pivotal to analyze the behavior of source and spectral characteristics in both the cases. The existing methods lack in detail study of deviation in acoustic-phonetic features between natural and synthesized speech, based on production behavior of a different category of sound units.

### 4.1.2 Proposed approach

This work focuses on adaptation of statistical behavior of different source and spectral parameters from the natural speech to generated parameter sequences. The source parameters used are $F_0$, the SoE and spectral parameters are the amplitudes of the first five spectral peaks and valleys. The normal distributions of the parameters extracted from frames of different sound categories are obtained for parallel set of natural and synthesized utterances. The means and variances of these distributions are used in further PF. For the spectrum of frames of different classes, separate means and variances are obtained for each of the first five spectral peaks and valleys to restore the fine spectral structure. It can be hypothesized that design of separate post-filters for each of broad categories of sound units may be helpful for achieving wider range, the fine structure of parameter sequence, as well as less deviation from the natural counterpart. Due to the incorporation of different PF factors in both temporal and spectral domains, the proposed method can be termed as *dynamic PF*.

The categorization of the sound units is performed based on similarity of articulatory movements during production, and its manifestation on the source and spectral characteristics. The change in amplitudes of spectral peaks, valleys and SoE can be seen as in increasing order of the sequence: *nasals, liquids, glides, high-vowels, mid-vowels, low-vowels*. This order is also referred to as sonority sequence [11]. Sonority refers to relative loudness of sound units depending on the different place and amount of constriction made by the vocal-tract, which is manifested in the speech signal. The spectral prominence, the energy associated with excitation source and periodicity of speech signal change with the variation in sonority level. Therefore it can be considered as a useful means to classify the sound units concerning variation in their source and spectral characteristics. In Chapter 3, a set of features is proposed using the knowledge of spectral peaks, valleys, bandwidth, and SoE to classify speech signal into these classes. An SVM classifier is trained using this sonority feature to classify speech sounds into different sonorant categories mentioned above.

At the time of synthesis, based on the class information of each frame obtained from the SVM classifier, a decision is made on which post-filter mean and variance should be used to modify the features corresponding to that frame. The post-filtered source and spectral parameters are further

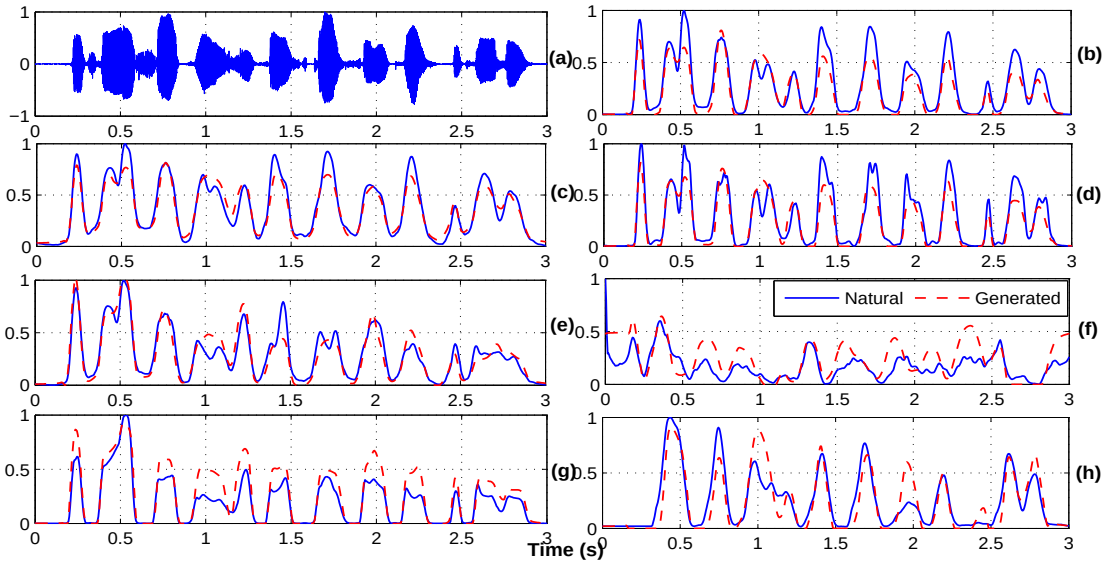


**Figure 4.1:** Extracted (from natural speech) and generated (from HMM) sonority features: (a) natural speech utterance, (b) formant peak values ($f_1$), (c) formant peak deviation ($f_2$), (d) amplitude of spectral valleys ($f_3$), (e) slope associated with formant peaks ($f_4$), (f) formant bandwidth ($f_5$), (g) strength of excitation ($f_6$), (h) suprasegmental feature ($f_7$) for SLT speaker.

used to render the synthesized speech. Further, the derived synthesized is analyzed to observe the deviation in spectral tilt with respect to that of the natural. A spectral tilt modification method is proposed to compensate the same. Some of the major contributions included in this chapter are:

- PF of the source and spectral parameters using varying factors with different sonorant classes.
- Use of different PF factors for enhancement of various spectral peaks and valleys based on their position in the spectrum of a frame.
- Incorporation of dynamic (delta and delta-delta) sonority feature in SPSS.
- Integration of the sonority feature in the SPSS framework.
- Spectral tilt based PF method.

The integration of sonority feature with the SPSS framework and SVM classifier are discussed in Section 4.2. The class-based comparison between natural and generated source and spectral parameters is carried out in Section 4.3 and Section 4.4, respectively. The dynamic post-filter designed to adapt the characteristics of natural speech to that of synthesized speech is explained in Section 4.5. Different experiments performed to establish the efficacy of the dynamic PF are also presented in the same section. The method proposed for modification of spectral tilt is presented in Section 4.6. In Section 4.7, the contributions made in the current chapter is summarized.

## 4.2 Sonority for post-filtering

In this section, the difference in production behavior of different voiced sound categories and its impact on the source and spectral parameters are discussed in brief. During the production of different sound units, the articulatory movements vary widely based on the sound produced. Narrowing the cross-sectional area in the front part of vocal-tract and widening towards the back results in the decrease of $F_1$ and increase in bandwidth of the formant. As $F_1$ increases the amplitude of first formant peak along with the higher formant peaks also increases. Therefore, if we move from low-vowels to high-vowels, $F_1$, $F_2$ decrease and the corresponding formant peak values also decreases. Compared to vowels, both liquids and glides have lower $F_1$ with wider formant bandwidth. Nasals have further lower $F_1$. These effects of vocal-tract constriction are reflected in the entire VTS as well as the excitation source with different categories of sound units [20]. As the vocal-tract constriction decreases, the SoE increases. These sounds with relatively wider vocal-tract shape, higher energy, sharper glottal closure instants are classified as sonorant sounds.

Based on the above discussion, we can hypothesize that depending on the sonority associated with each frame of synthesized speech, the behavior of formant peaks, valleys and SoE may be different. Therefore, the nature of the VTS and excitation source of synthesized and natural speech signal can be analyzed with reference to the sonorant classes. Different PF factors can be derived with respect to the sonority associated with a particular frame.

### 4.2.1 Dynamic sonority feature

The sonority associated with a frame of the sound unit also depends on acoustic characteristics of its nearby frames. The frames of the same sound unit may have different source and spectral attributes based on various left and right context. For example, when a high-vowel is adjacent to a glide, its SoE, and spectral peaks will be more prominent compared to the high-vowel adjacent to a nasal. Similarly, if a low-vowel is adjacent to an obstruent, its source and spectral feature behavior may be different from a low-vowel adjacent to any of the sonorant sounds [20]. To include this fact the dynamic sonority feature is also considered in this chapter, which is not included in the previous chapter. Dynamic sonority feature represents the delta and delta-delta of the derived 7-dimensional sonority feature, calculated as given in (4.1) and (4.2).

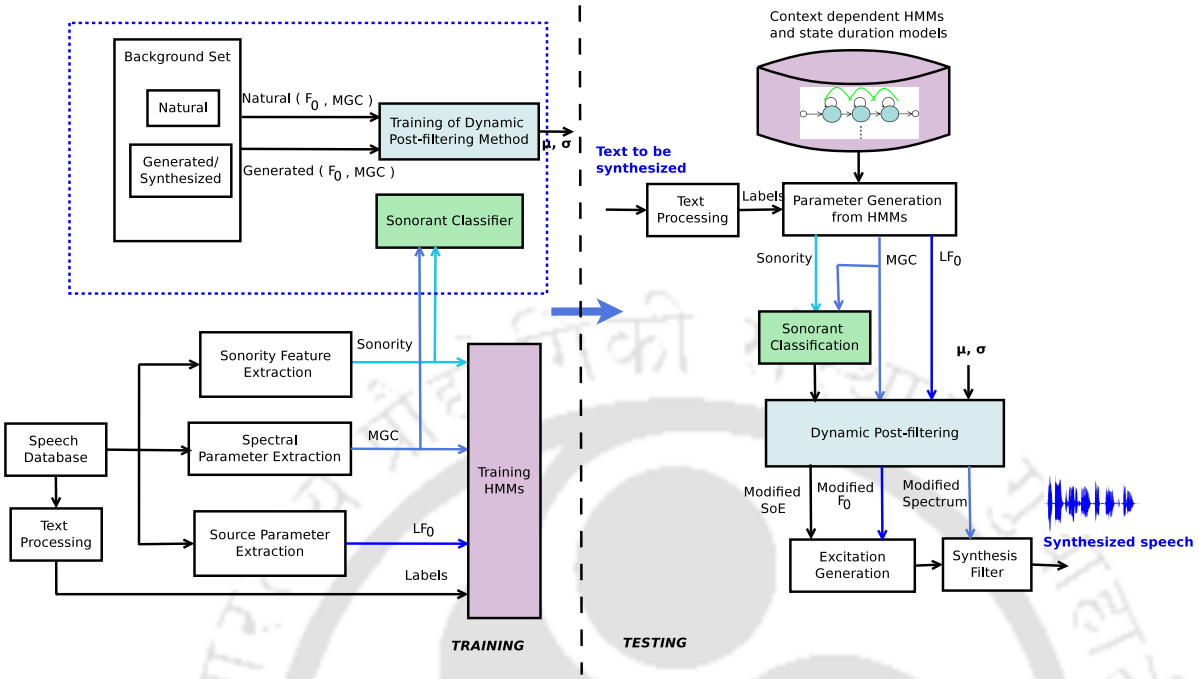


**Figure 4.2:** Block diagram representing the proposed framework.

$$\Delta f_k(i) = \frac{\sum_{n=1}^{N}[f_k(i+n) - f_k(i-n)]}{2\sum_{n=1}^{N} n^2}, \tag{4.1}$$

$$\Delta\Delta f_k(i) = \frac{\sum_{n=1}^{N}[\Delta f_k(i+n) - \Delta f_k(i-n)]}{2\sum_{n=1}^{N} n^2}, \tag{4.2}$$

where, $\Delta f_k(i)$ is the first derivative of the feature $f_k(i)$, $k = 1, 2..7$ and $i$ corresponds to the epoch location. Similarly $\Delta\Delta f_k(i)$ is the second derive of $f_k(i)$. The 7-dimensional sonority feature along with its dynamic features provides better representation of sonority information resulting in a 21-dimensional feature vector. In the later part of the manuscript, this 21-dimensional feature vector is referred as sonority feature.

### 4.2.2 Integration of sonority feature in SPSS

In this part of the manuscript, integration of the sonority feature in SPSS framework is explained. SPSS provides a unified framework to model vocal-tract, excitation, and duration parameters simultaneously in HMMs [8]. We have used the traditional HMM based speech synthesis system (HTS) toolkit to develop the systems [8], which involves training and testing processes. In the training phase, excitation, vocal-tract and duration parameters along with the sonority feature are extracted from the speech signal corresponding to the training database as shown in Figure 4.2. All the phonemes are

modeled with 5 states. In each state, 5 streams are used to model the different parameters extracted for each phoneme. The spectral parameters used are MGC coefficients including the zeroth coefficient (35-dimensional) and their delta and delta-delta coefficients. Total 105-dimensional spectral parameters are modeled using continuous density HMMs in the first stream. The source parameter used are log $F_0$, its delta and delta-delta coefficients (3-dimensional). In this case, $F_0$ values and voicing decisions are obtained from the RAPT [87], which are modeled together MSD-HMM in three independent streams [7]. The sonority feature along with its delta and delta-delta coefficients (21-dimensional) are also modeled using continuous density HMMs in the fifth stream. For each phoneme, the source, spectral and sonority features described above are extracted along with their corresponding labels from the training utterances. In the train phase, the maximum likelihood estimation of each parameter is computed using BW reestimation algorithm. During synthesis, as per the input text, using the maximum likelihood parameter generation algorithm, frame wise MGC coefficients, sonority feature and $F_0$ sequences are computed by maximizing output probability of static and dynamic features [8]. The generated sonority feature is used to classify the test frame into corresponding sonorant category.

### 4.2.3 SVM classifier using sonority feature

Two speakers, SLT (US female) and BDL (US male) from the CMU arctic database are used in this study [48]. All the speech recordings used in this work are at 48 kHz sampling frequency. The sonority feature explained above is extracted from the entire database for both the speakers. For the first 1000 training utterances, the sonority feature is modeled in HMM during the training phase of SPSS and the same feature is generated for the rest 132 utterances during testing. The extracted (from natural speech) and generated (from HMM models) feature contours for the same utterance are shown in Figure 4.1 in bold and dashed lines, respectively. The gross behavior of the generated feature contour is similar to the natural, which indicates that the sonority feature is modeled correctly by preserving all the variations. This generated feature vector is normalized and further employed to develop a six-class SVM (with the RBF kernel) based sonorant classifier, where the six classes are *low-vowels, mid-vowels, high-vowels, glides, liquids, and nasals.* Out of the 132 testing utterances, only 100 are used for developing the SVM classifier. Values of the parameters, $c$ and $\gamma$ are set using train-test 5-fold cross-validation for the 100 utterances. The usefulness of delta and delta-delta sonority feature is checked using SVM classifiers with the same procedure for both the speakers (SLT and BDL) using the following feature vectors.

**Table 4.1:** % Accuracy with corresponding $c$ and $\gamma$ values of SVM-based sonorant classifiers for different features.

| Feature | Dimension | SLT (Female) | | | BDL (Male) | | |
|---|---|---|---|---|---|---|---|
| | | % Acc | c | $\gamma$ | % Acc | c | $\gamma$ |
| Sonority (S) | 7 | 70.72 | 256 | 16 | 73.03 | 128 | 4 |
| S+Δ S+ΔΔ S | 21 | 72.50 | 512 | 16 | 75.02 | 128 | 4 |
| MGC | 35 | 90.96 | 128 | 2 | 92.91 | 64 | 2 |
| MGC+S | 42 | 95.48 | – | – | 96.85 | – | – |
| MGC+S+Δ S+ΔΔ S | 56 | **97.33** | – | – | **98.26** | – | – |

- Static 7-dimensional sonority feature.
- Both static, delta and delta-delta sonority feature.
- Only MGC coefficients feature vector.
- MGC coefficients and static sonority feature vector (score level combination).
- MGC coefficients, static, delta, and delta-delta sonority feature vector (score level combination).

The average classification accuracy obtained for the above sonorant classifiers can be observed in Table 4.1 for both the speakers along with corresponding $c$ and $\gamma$ values. The classification accuracy increases from 70.72% to 95.48% when MGC based classifier and the static sonority based classifier are fused at probability score level. Further, the combination of dynamic sonority feature classifier with MGC classifier improves the performance to 97.33% (for SLT speaker). The inclusion of dynamic sonority feature with MGC provides the best classification accuracy as seen from Table 4.1. Therefore, this sonorant classifier is used further in this work.

## 4.3 Analysis of different aspects of excitation source

While producing different types of sounds, the vibration pattern of glottal source gets modified. Due to the fluctuation of supraglottal pressure with variation in the vocal-tract constriction, the open and closed phase of the glottis changes. This change results in SoE and $F_0$, which are two critical excitation source parameters considered in this study. $F_0$ plays a significant role in speech naturalness that includes its effect on voicing cues, syllable stress, speaker identity, cues of vowel identity and so on. In [132], it is reported that flattening of $F_0$ results in a reduction of naturalness. SoE plays a crucial role in the perception of loudness in the speech signal, which is related to the abrupt closing of vocal folds [109]. During speech production, although the vocal-tract shape governs

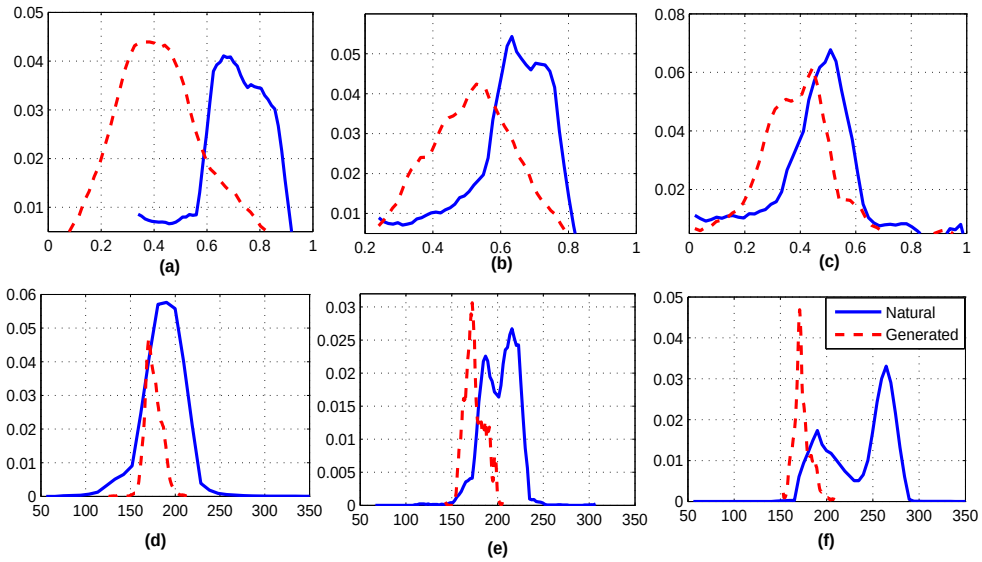


**Figure 4.3:** Distributions of SoE for (a) vowels, (b) semivowels, (c) nasals and $F_0$ for (d) vowels, (e) semivowels, (f) nasals, extracted from the natural speech signal and generated from HTS.

the identity of speech sound, for the same vocal-tract configuration the glottal source characteristics may vary. These characteristics play a greater role in speech naturalness, as well as in the perception of additional information like stress, loudness in voiced sounds. The $6^{\text{th}}$ dimension of the sonority feature ($f_6$) represents the SoE. The procedure described in Chapter 3 is employed to derive the SoE from natural and synthesized speech signals for all the frames corresponding to vowels, semivowels, and nasals. Figure 4.3 (a),(b),(c) show the distributions of normalized SoE for vowels, semivowels, and nasals for both natural and synthesized (after GV+MS based PF) speech. It can be observed that the area of overlapping regions between the pair of distributions is different for vowels, semivowels, and nasals. In the case of nasals, Figure 4.3(c), the distributions of SoE seems to be more overlapping compared to that of vowels in Figure 4.3(c). The fact behind this is that the vowels are produced with sharper glottal closure instants compared to that of the nasals, which gets smoothed out in the statistical modeling framework. Therefore, it can be referred that, these broad classes of sound units required to be enhanced (or post-filtered) by different PF factors, that may add naturalness to the synthesized speech signal.

Similar kind of observations can be made from Figure 4.3(d),(e),(f), which show the distributions corresponding to $F_0$ for natural and synthesized speech of vowels, semivowels, and nasals, respectively. The distance between the distributions of vowels and semivowels is less compared to the distance

**Table 4.2:** Means and standard deviations (std.) corresponding to normal distributions obtained from $F_0$ and SoE of natural and synthesized speech for different categories of sound units.

| Category | $F_0$ | | | | SOE | | | |
|---|---|---|---|---|---|---|---|---|
| | Natural | | Synthesized | | Natural | | Synthesized | |
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| Vowels | 192.51 | 28.20 | 176.22 | 9.10 | 0.70 | 0.20 | 0.44 | 0.17 |
| Semivowels | 203.64 | 18.53 | 175.31 | 10.61 | 0.62 | 0.16 | 0.52 | 0.14 |
| Nasals | 238.28 | 35.67 | 174.22 | 9.42 | 0.47 | 0.18 | 0.41 | 0.17 |

between the distributions of nasals corresponding natural and generated $F_0$ sequence. Therefore, nasals require a greater amount of modification concerning $F_0$ compared to that of vowels and semivowels. For detailed analysis, Table 4.2 depicts the means and standard deviations of $F_0$ and SoE for these classes in case of natural and synthesized speech. In the case of $F_0$, it can be observed that the mean value of the synthesized speech does not deviate much from that of natural. The standard deviation of $F_0$ for synthesized speech is lesser than that of natural in most of the cases due to the over-smoothing effect. The SoE parameter has less mean and standard deviation values for synthesized speech compared to that of natural. However, from Table 4.2, it is clear that the parameters corresponding to each category of the sound unit are required to be enhanced by a different factors for both SoE and $F_0$.

## 4.4 Analysis of vocal-tract parameters

The source parameters discussed in Section 4.3 are dependent on the airflow and vibration of the vocal-folds. However, they are independent of the shape of the vocal cavity to a large extent. The changing shape of the vocal-tract configuration is reflected in the formant characteristics of the VTS. The first three formants play a significant role in different speech processing tasks. Spectral valleys are also of importance for the overall study of the spectrum shape. In this section, a detailed comparison of these three formant peaks and their preceding valleys is carried out between synthesized and natural speech relative to each sonorant category. For this analysis, the MGC coefficients are extracted from the natural speech signal and generated from HMMs for 100 test utterances, corresponding to all the frames of vowels, semivowels, and nasals. Log magnitude spectrum is derived from these MGC coefficients. In the log magnitude spectrum, the first three peaks and their preceding valleys are detected using a peak detection algorithm with a proper threshold of bandwidth, slope, and amplitude. Then, the three peak values for each sonorant class corresponding to natural and generated spectrum

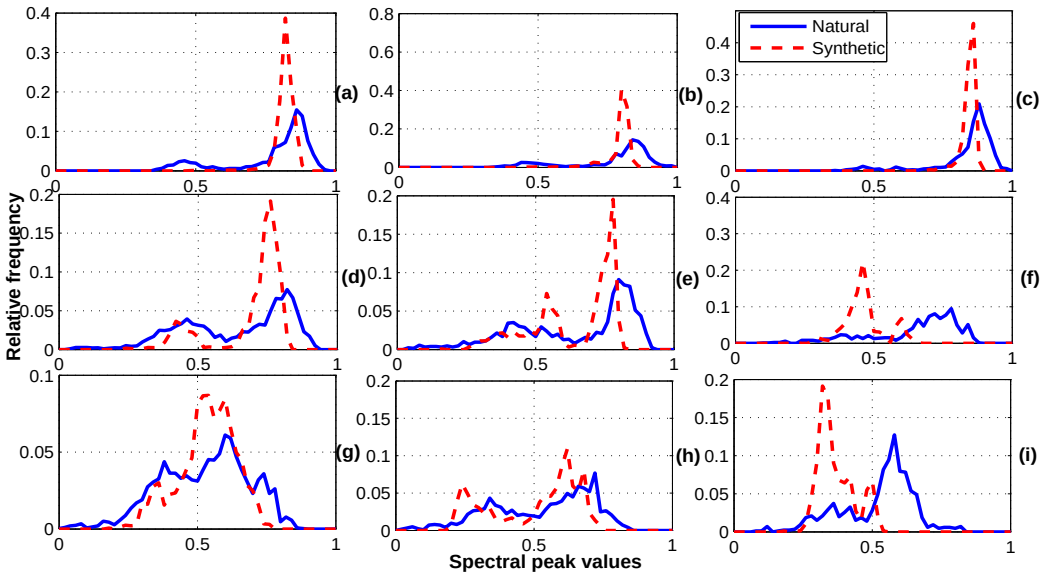


**Figure 4.4:** Distributions of formant peak values for natural and synthesized speech. (a), (b), (c) 1st spectral peaks for vowels, semivowels, nasals; (d), (e), (f) 2nd spectral peaks for vowels, semivowels, nasals; (g), (h), (i) 3rd spectral peaks for vowels, semivowels, nasals.

are accumulated. As three peaks and three classes are considered, 9 clusters are obtained for peaks of natural and synthesized. Similarly, 9 clusters are derived for each of natural and synthesized spectral valleys. Each of the clusters is normalized concerning its maximum value as follows.

$$\mathbf{C}^{(G)}_{\mathbf{P}_{j,i}} = \frac{\mathbf{C}^{(G)}_{\mathbf{P}_{j,i}}}{\max(\mathbf{C}^{(G)}_{\mathbf{P}_{j,i}})}, \tag{4.3}$$

$$\mathbf{C}^{(N)}_{\mathbf{P}_{j,i}} = \frac{\mathbf{C}^{(N)}_{\mathbf{P}_{j,i}}}{\max(\mathbf{C}^{(N)}_{\mathbf{P}_{j,i}})}, \tag{4.4}$$

$$\mathbf{C}^{(G)}_{\mathbf{V}_{j,i}} = \frac{\mathbf{C}^{(G)}_{\mathbf{V}_{j,i}}}{\max(\mathbf{C}^{(G)}_{\mathbf{V}_{j,i}})}, \tag{4.5}$$

$$\mathbf{C}^{(N)}_{\mathbf{V}_{j,i}} = \frac{\mathbf{C}^{(N)}_{\mathbf{V}_{j,i}}}{\max(\mathbf{C}^{(N)}_{\mathbf{V}_{j,i}})}, \tag{4.6}$$

$$i = 1, 2, 3; \quad j = 1, 2, 3;$$

where, $\mathbf{C}^{(G)}_{\mathbf{P}_{j,i}}$ and $\mathbf{C}^{(N)}_{\mathbf{P}_{j,i}}$ represent the clusters corresponding to $i^{\text{th}}$ class and $j^{\text{th}}$ peak of natural and generated spectrum, respectively. Similarly, $\mathbf{C}^{(G)}_{\mathbf{V}_{j,i}}$ and $\mathbf{C}^{(N)}_{\mathbf{V}_{j,i}}$ represent the same for spectral valleys. The distributions corresponding to each of these clusters are shown in Figure 4.4 and Figure 4.5, for peaks and valleys, respectively. Figure 4.4 (a),(b),(c) show distributions of 1st spectral peaks for vowels,

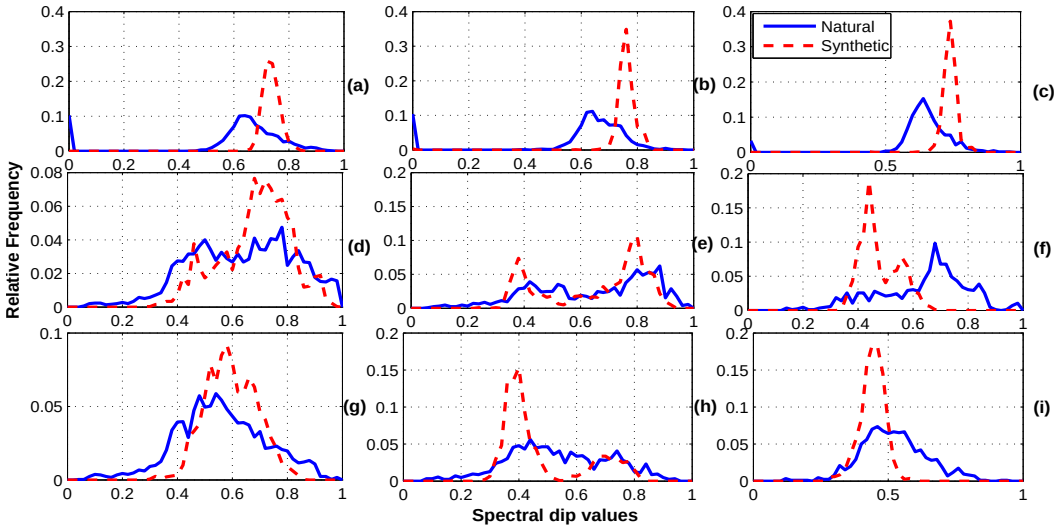


**Figure 4.5:** Distributions corresponding to values of formant valleys for natural and synthesized speech. (a), (b), (c) 1st spectral valleys for vowels, semivowels, nasals; (d), (e), (f) 2nd spectral valleys for vowels, semivowels, nasals; (g), (h), (i) 3rd spectral valleys for vowels, semivowels, nasals.

semivowels and nasals for natural and synthesized speech. Same is shown for 2nd and 3rd spectral peaks in Figure 4.4(d),(e),(f) and Figure 4.4(g),(h),(i), respectively. From Figure 4.4 (a),(b),(c), it can be observed that the amount of overlap between the distributions of 1st spectral peak of natural and synthesized speech is more in the case of vowels compared to semivowels and nasals. The same is more prominent in the case of Figure 4.4(d),(e),(f), where the amount overlap is more for vowels and semivowels. For nasals, the distributions of 2nd spectral peak have more distance. This is also evident from the third row of Figure 4.4. From this observation, we can infer that the deviation of spectral peak amplitude or sharpness between natural and generated spectrum varies with the broad classes of sounds. For improving naturalness, it may be required to introduce more enhancement in the case of nasals compared to that of vowels and semivowels. In Figure 4.4 each column represents vowels, semivowels, and nasals. From the first column if we observe (a), (d), (g) it can be inferred that for vowels the 1st and 2nd spectral peaks has less overlap between natural and generated compared to that of the 3rd spectral peak. Same applies to semivowels if we observe the 2nd column of Figure 4.4. In the case of nasals, the distributions corresponding to both 2nd and 3rd spectral peaks are less overlapped. These deviations can be compensated by applying varying PF for different spectral peaks with respect to various sound categories.

Along with the spectral peaks, spectral valleys also provide important cues regarding the changing shape of the vocal-tract while producing different speech sounds. The distributions of the first three

spectral valleys preceding to the spectral peaks are shown in Figure 4.5. The three spectral valleys amplitudes are normalized individually for each category as given in (4.5) and (4.6). In Figure 4.5, the three rows depict the distributions of 1$^{st}$ ((a),(b),(c)), 2$^{nd}$ (d),(e),(f)) and 3$^{rd}$ ((g),(h),(i)) spectral valleys for vowels, semivowels, and nasals. From the distribution of 1$^{st}$, 2$^{nd}$, and 3$^{rd}$ spectral valleys, it can be observed that the distance between the distributions corresponding to the natural and synthesized cases are increasing as we move from vowels to nasals. In Figure 4.5(d),(e) means and variances of the natural and synthesized cases are almost similar. On the other hand, in Figure 4.5(i) (3$^{rd}$ valley for nasal), the means and variances are quite less in the synthesized case as compared to the natural. For 1$^{st}$ and 2$^{nd}$ spectral valleys, the mean values of the distributions obtained from the natural speech are lesser than that of the synthesized. While, in the distributions of 3$^{rd}$ spectral valleys in Figure 4.5(g),(h),(i) the mean values obtained from natural speech are more than that of the generated spectrum. From this discussion and by observing Figure 4.5, it can be referred that in the generated spectrum obtained from SPSS, the first two spectral valleys have more amplitude compared to that of natural while the 3$^{rd}$ spectral valley has less amplitude compared to that of the natural. Again, it can be inferred from Figure 4.5 that, spectral valley amplitudes in the case of vowels and semivowels in the synthesized case are more accurate compared to that of the nasals.

From the above analysis of first three spectral peaks and valleys of vowels, semivowels, and nasals, it can be hypothesized that enhancement of these spectral peaks and valleys by using separate post-filter for different categories may help to make the synthesized speech spectrum close to that of natural speech. In the analysis of excitation source aspects, it is observed that the vowels and semivowels have more effect of over-smoothing due to statistical averaging compared to that of the nasals. While in spectral based features the distributions of nasals have more distance between synthesized and natural case compared to that of the vowels and semivowels.

## 4.5 Dynamic source and spectral post-filtering

The overall framework of the proposed dynamic PF can be seen from the block diagram in Figure 4.2. During training of the post-filter, parallel utterances of natural and synthesized speech are used to extract the parameters, and corresponding means and standard deviations are retained as shown in Figure 4.2. These are further used in testing phase to modify the corresponding feature contours. The detailed procedure is described below.

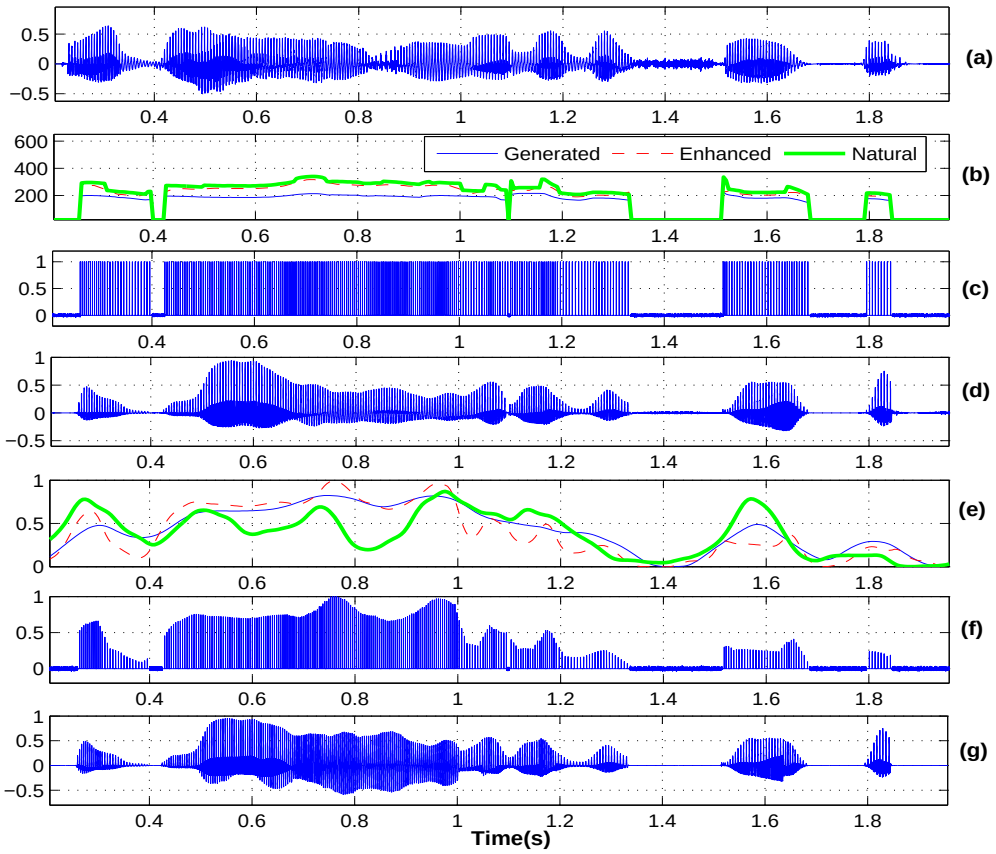


**Figure 4.6:** Enhancement of excitation source; (a) natural speech segment, (b) generated and enhanced $F_0$ contour, (c) generated and enhanced SoE contour, (d) impulse based excitation source generated from the enhanced $F_0$ contour, (e) SoE weighted excitation source (enhanced source), (f) synthesized speech without using source enhancement, (g) synthesized speech using enhanced source, for initial 1.4 seconds of the utterance "But she had become an automaton" for SLT speaker.

### 4.5.1 Source post-filtering

The fundamental frequency, $F_0$ and SoE play an important role in the perception of speech and these two parameters are considered for the proposed source PF. As discussed in Section 4.3, means and standard deviations of the normal distributions of $F_0$ extracted from natural and synthesized speech are calculated. As shown in Section 4.3, the means and standard deviations of $F_0$ may vary with different sound categories. Therefore, these parameters are obtained separately for the classes: low-vowels, mid-vowels, high-vowels, glides, liquids, and nasals during the training stage of the PF method. During testing, given the value of generated $F_0$ for a particular frame, the class to which the frame belongs is determined using the SVM classifier described in Section 4.2.3. If the test frame belongs to $i^{\text{th}}$ class then following post-filter is used to obtained modified $F_0$ for that frame.

$$F_0^{pf}(i,m) = (1-\beta_{F_0})F_0(i,m) + \beta_{F_0}[\frac{\sigma_{F_0,i}^{(N)}}{\sigma_{F_0,i}^{(G)}}(F_0(i,m) - \mu_{F_0,i}^{(G)}) + \mu_{F_0,i}^{(N)}], \tag{4.7}$$

where, $\mu_{F_0,i}^{(G)}$, $\mu_{F_0,i}^{(N)}$, $\sigma_{F_0,i}^{(G)}$, $\sigma_{F_0,i}^{(N)}$ are means and standard deviations of the generated and natural $F_0$ for $i^{\text{th}}$ class, respectively. $F_0(i,m)$ is the fundamental frequency of $m^{\text{th}}$ test frame and $F_0^{\text{pf}}(i,m)$ is the corresponding post-filtered fundamental frequency. $\beta_{F_0}$ is the factor for controlling the PF effect. In this case, we have set $\beta_{F_0} = 1$. The generated and post-filtered $F_0$ contours are shown in Figure 4.6(b), corresponding to the natural speech signal shown in Figure 4.6(a). From the knowledge of $F_0^{\text{pf}}$ contour, impulses of constant amplitude are inserted in voiced regions, and random noise (white noise) is generated in unvoiced regions, which is shown in Figure 4.6(b). As analyzed in Section 4.3, the means and standard deviations of the distributions obtained for SoE values in synthesized speech is lower compared to that of the natural speech. As mentioned in Section 4.2, the excitation information present in sonority feature vector ($6^{\text{th}}$ dimension) represents SoE, which is modeled in the SPSS framework. The generated SoE contour for the utterance is shown in Figure 4.6(e). Similar to $F_0$, post-filters are applied to obtain the modified SoE contour from the generated SoE contour. The impulse like characteristics of excitation source represented by SoE is an important factor affecting speech perception, that is governed by the abruptness of the closing phase of the glottal cycle. The impulse sequence weighted by SoE is found to be effective for speech synthesis in [133]. To introduce the change in SoE, in the excitation source shown in Figure 4.6(c), the impulse sequence in voiced part is weighted by the modified SoE values shown in Figure 4.6(f). The speech signal generated with and without the source PF is shown in Figure 4.6(d) and Figure 4.6(g), respectively.

### 4.5.2 Spectral post-filtering

Due to the difference in vocal-tract constriction and shape, the statistics of peaks and valleys of the estimated VTS also vary with sonorant categories. As described in Section 4.4, the amplitudes of different spectral peaks and valleys for various sonorant categories need to be modified to the required extent. Therefore, in this work separate post-filter parameters are trained for each sonorant class each of the first five spectral peaks and valleys. This kind of modification may give the spectrum of synthesized speech closer to that of the natural. As discussed in Section 4.4, from the logarithmic spectrum, first five peaks and their preceding valleys are detected. The value of these peaks and

valleys are normalized as given in (4.3), (4.4), (4.5), (4.6) for different classes and different positions in the spectrum. The means and standard deviations are calculated from corresponding distributions and retained in the training process.

During the testing phase of PF, given the MGC contour of a particular frame, firstly MGC coefficients are converted to log-magnitude spectrum and normalized with respect to the maximum value of the spectrum. The first five peaks and their preceding valleys are detected from the normalized log spectrum. For instance, the first three peaks detected in the spectrum are shown in Figure 4.7. For the same frame, using the generated sonority feature in the previously trained SVM classifier, the class information is retrieved. Let the $m^{\text{th}}$ test frame (from generated MGC contour) belongs to $i^{\text{th}}$ sound category, then the value of $j^{\text{th}}$ peak for this frame is represented by $P_j(i,m)$. $P_j^{pf}(i,m)$ denotes the corresponding post-filtered peak value, which can be obtained by using the (4.8).

$$P_j^{pf}(i,m) = (1-\beta_p)P_j(i,m) + \beta_p[\frac{\sigma_{P_j,i}^{(N)}}{\sigma_{P_j,i}^{(G)}}(P_j(i,m) - \mu_{P_j,i}^{(G)}) + \mu_{P_j,i}^{(N)}], \tag{4.8}$$

where, $\mu_{P_j,i}^{(N)}$ and $\mu_{P_j,i}^{(G)}$ represent the means of $j^{\text{th}}$ peak values of the spectrum obtained for $i^{\text{th}}$ category of sound units for natural and generated spectrum, respectively. $\sigma_{P_j,i}^{(N)}$ and $\sigma_{P_j,i}^{(G)}$ represent the standard deviations of $j^{\text{th}}$ peak values of the spectrum obtained for $i^{\text{th}}$ category of sound units for natural and generated spectrum, respectively. $\beta_p$ is the coefficient for controlling the PF effect in case of spectral peaks, which considered as 1. The PF factor for each spectral peak can be obtained as follows.

$$\alpha_j(i,m) = \frac{P_j^{pf}(i,m)}{P_j(i,m)}, \tag{4.9}$$

The normalized spectrum around $j^{\text{th}}$ peak is weighted symmetrically by the factor $\alpha_j(i,m)$ to obtain the modified peak value as shown in Figure 4.7. The width, for which the spectrum is weighted symmetrically both sides of each peak is equal to half of the distance between the peak and its preceding valley. The weighted spectrum can be seen in dotted line from Figure 4.7.

All the five peaks of each frame are modified in this manner. Similar to spectral peaks, separate post-filters are used to derive the adjusted values of spectral valleys as given in (4.10).

$$V_j^{pf}(i,m) = (1-\beta_d)V_j(i,m) + \beta_d[\frac{\sigma_{V_j,i}^{(N)}}{\sigma_{V_j,i}^{(G)}}(V_j(i,m) - \mu_{V_j,i}^{(G)}) + \mu_{V_j,i}^{(N)}], \tag{4.10}$$

where, $\mu_{V_j,i}^{(N)}$ and $\mu_{V_j,i}^{(G)}$ represents the mean of $j^{\text{th}}$ spectral valleys of the spectrum obtained for $i^{\text{th}}$

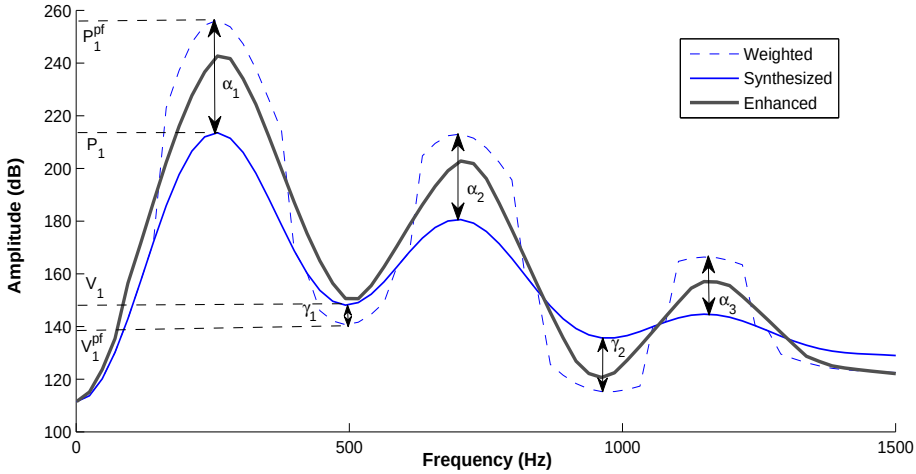


**Figure 4.7:** Illustration of spectral PF method.

category of sound units for natural and generated spectrum, respectively. $\sigma^{(N)}_{V_j,i}$ and $\sigma^{(G)}_{V_j,i}$ represents the standard deviations of the same. $\beta_d$ is the coefficient for controlling the PF effect in case of spectral valleys, which is considered as 1 in this study. The PF factor for each spectral valley can be obtained as follows.

$$\gamma_j(i,m) = \frac{V_j^{pf}(i,m)}{V_j(i,m)}, \tag{4.11}$$

The normalized spectrum around $j^{\text{th}}$ spectral valley is weighted symmetrically by the factor $\gamma_j(i,m)$ to obtain the enhanced spectral valleys as shown in Figure 4.7. In this case, $j$ ranges from 1 to 5 and $i = 1, 2, ..., 6$ for six category of sound units. To avoid the discontinuity at the boundary regions between peaks and valleys, mean smoothing over a window of 250 Hz is performed. The spectrum after performing the smoothing is the enhanced spectrum shown in Figure 4.7 with a solid gray line. After the PF of spectral peaks and valleys, the normalization effect is removed from the spectrum by using previously obtained (during normalization) minimum and maximum values of the spectrum. This procedure is done to preserve the spectral energy. The normalized spectrum obtained using the proposed method is shown in Figure 4.8(a), (b), (c) for a frame of vowel, semivowel, and nasal, respectively. It depicts the natural, synthesized and enhanced spectrum in each case, which shows that using the proposed PF technique, the peaks are made sharper by using different PF factors. The obtained spectrum using the proposed method is closer to the natural counterpart.

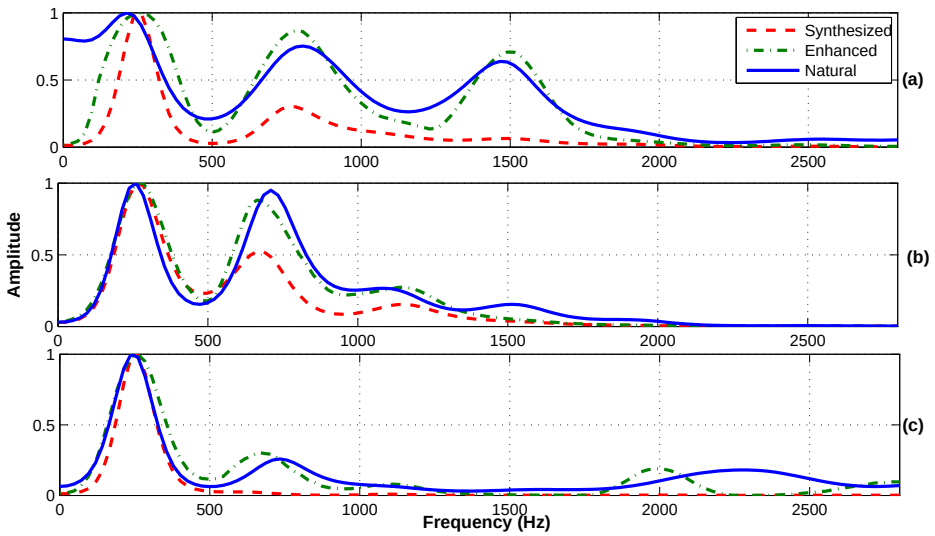


**Figure 4.8:** Synthesized, enhanced and natural spectrum for a frame of (a) vowel (b) semivowel (c) nasal.

### 4.5.3 Experimental evaluation

In this section, the efficacy of the proposed PF method regarding alleviating the difference between the parameters of natural and synthesized speech is analyzed. The quality of the post-filtered speech is compared with different state-of-the-art methods, in terms of acoustic representation, objective and subjective evaluations. For this, SPSS voices are developed using CMU arctic database for male (BDL) and female (SLT) speakers. As mentioned earlier, for each of the voices, first 1000 utterances are used in training of the SPSS. Out of the rest 132 utterances, 100 are used for comparison of natural and generated parameter sequences to derive the dynamic PF factors and developing the SVM classifier. This stage can be referred as training phase of the proposed PF method. The rest 32 utterances are used in the testing phase.

#### 4.5.3.1 Implementation of state-of-the-art methods

The proposed method is implemented by training source, and spectral parameters along with the sonority feature in the SPSS framework, as explained in Section 4.2.2. During testing, depending on the input text to be synthesized, context-dependent HMMs are concatenated to get the sentence HMM. Based on maximum likelihood criterion, source (log $F_0$), spectral (MGC coefficients) and sonority feature parameters are generated from the HMMs [134]. The MGC coefficients and sonority feature of each frame are fed to the SVM classifier to derive the class information as shown in Figure 4.2. The source and spectral parameters of each frame are then fed to the proposed dynamic PF method

based on corresponding sonorant class label. The post-filtered parameters for the entire utterance are further applied to the vocoder to derive the synthesized speech as shown in Figure 4.2. In this case, the 35-dimensional MGC coefficients are used and the vocoder employed is MLSA filter.

Apart from the proposed method, the implementation of the other state-of-the-art methods of PF does not require modeling of the sonority feature. In this case, HMMs are trained using the same source and spectral features. The spectral features, MGC coefficients including the zeroth coefficient (35-dimensional) and their delta and delta-delta coefficients are modeled using continuous density HMMs in the first stream. log $F_0$, its delta and delta-delta are modeled in the next three streams using MSD-HMM. There are total four streams in this case, whereas the system with proposed dynamic PF uses five streams to model additional sonority feature. The remaining training process is same as described in Section 4.2.2. During testing, depending on the text to be synthesized, log $F_0$ and MGC coefficients are generated, which are fed to MLSA filter to render the synthesized waveform.

To check the efficacy of the proposed method in advanced vocoder, the STRAIGHT based HTS framework is also included in our study [53]. During training, additional band aperiodicity parameters (BAP) (26-dimensional) are incorporated along with other parameters mentioned above. In this case, the spectral representation using MGC coefficients is 50-dimensional, which is 150-dimensional with delta and delta-delta coefficients. The source parameter log $F_0$ and voicing decisions are extracted using the STRAIGHT algorithm in this case. BAP and MGC coefficients are modeled along with their delta and delta-coefficients using continuous distribution HMMs, while log $F_0$, its delta and delta-delta are modeled using MSD-HMMs. The remaining training process is similar to the systems as mentioned above. During testing, the generated log $F_0$ and aperiodicity components are used to produce the excitation source, which passed through the STRAIGHT vocoder along with spectral parameters to derive the synthesized speech. For implementing the proposed method, additional sonority feature described above is also integrated into the STRAIGHT based framework. The dynamic PF method is employed to the excitation source and the spectrum before passing through the STRAIGHT vocoder. For the evaluation, the synthesized speech is derived under the following conditions.

- **NONE**: No enhancement method.
- **MCP**: Mel-cepstral post-filter [74].
- **LSP**: LSP based post-filter [76].
- **GV**: Global Variance based PF [23].

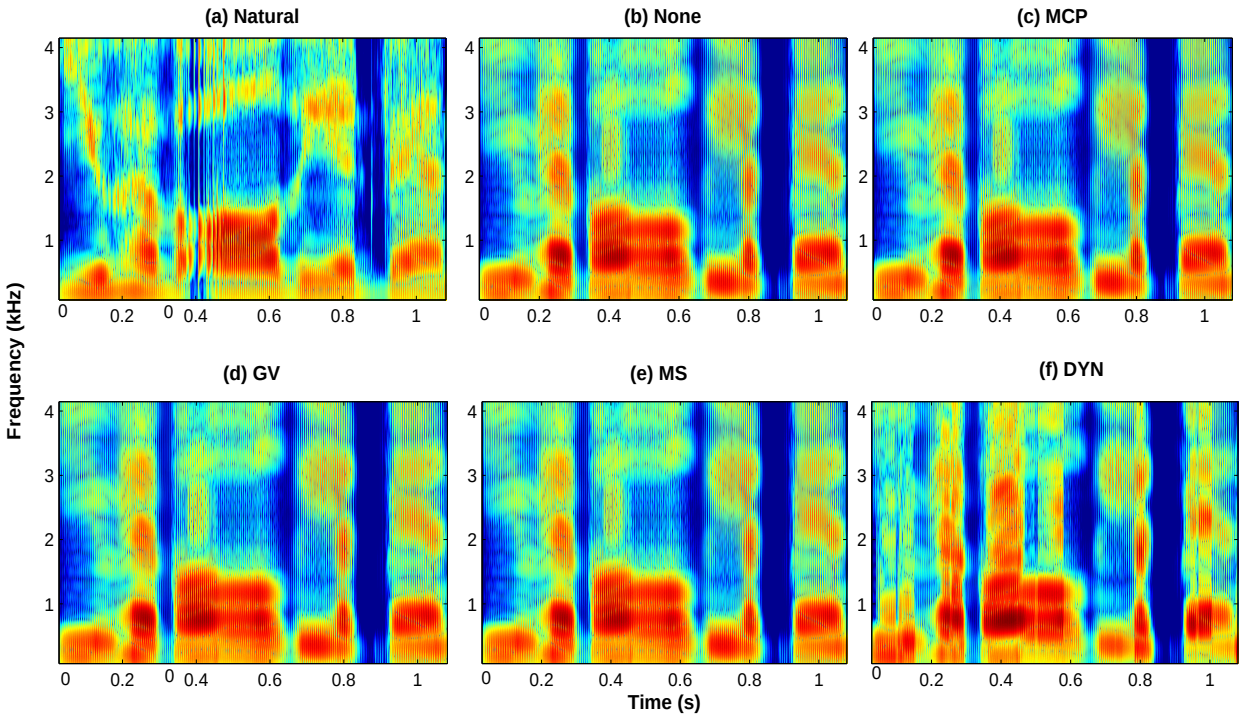


**Figure 4.9:** Spectrograms for the utterance "It was impossible to hoist sail and claw off that shore" for SLT speaker corresponding to (a) natural, (b) without any PF, (c) MCP PF, (d) GV based PF, (e) MS based PF, (f) Dynamic PF.

- **MS**: Modulation spectrum based PF [79].
- **SRC**: Dynamic source enhancement method (proposed).
- **SPEC**: Dynamic spectral enhancement method (proposed).
- **DYN**: Dynamic source and spectral enhancement method (proposed).
- **GV+MS**: Both GV and MS based PF.
- **GV+MS+DYN**: Both GV and MS based PF followed by DYN.
- **STRAIGHT:GV+MS**: Both GV and MS based PF with STRAIGHT vocoder.
- **STRAIGHT:GV+MS+DYN**: STRAIGHT:GV+MS followed by DYN.

The state-of-the-art methods MCP, LSP, GV, and MS, are already integrated with the latest version of the HTS framework, which can be implemented by setting the corresponding flags during training. The $\beta$ value for MCP is set to 1.4, and $\alpha$ value for LSP is set to 0.7 for the experiments.

Figure 4.9 shows the wide-band spectrogram (frame-size=5 ms, frame-shift=1 ms) for different types of synthesized speech (after applying different PF methods). It can be observed that the formants and spectral structure are much more prominent in the case of natural speech in Figure 4.9(a),

compared to synthesized speech without any PF, Figure 4.9(b). Figure 4.9(c), (d), (e) represent spectrograms corresponding to MCP, GV and MS based PF, respectively. The spectral prominence is better in GV and MS compared to MCP. The spectrogram obtained from speech signal obtained after using the proposed PF method is shown in Figure 4.9(f). It can be noted that the spectral prominence is more compared to that of GV and MS methods in Figure 4.9(d),(e). Apart from these pictorial observations regarding the effect of proposed dynamic PF method, objective and subjective evaluations are also performed.

#### 4.5.3.2 Objective evaluation

The objective measures used for evaluation of different post-filtered speech signal are the perceptual evaluation of speech quality-mean opinion score (PESQ-MOS) and log spectral distance (LSD). PESQ is a ITU-T recommendation P.862 method for speech quality evaluation [135]. It requires a reference signal and a candidate signal for which speech quality is to be measured. In this case, the reference signal is the natural speech signal. First, both the signals are aligned in time and then processed through an auditory transform. The auditory transform is a representation of perceived loudness in time and frequency. The distortion parameters are extracted from the difference between transforms of the reference and candidate signals. It gives PESQ-MOS values which are correlated with human perception. For calculating the LSD, firstly the reference and candidate signals are time aligned using dynamic time warping algorithm. For each of the different types of the synthesized speech obtained from different PF methods, the objective evaluation is performed for the 32 test utterances, for both male and female speakers. The corresponding average scores are depicted in Table 4.3. It can be observed from Table 4.3 that, both PESQ-MOS is more in the case of GV and MS compared to MCP and LSP. Again, these scores for SRC are less than both GV and MS. While SPEC has more PESQ-MOS compared to MS, yet it is lesser than that of GV. DYN is the proposed dynamic source and spectral PF method, and it has comparable scores with that of GV. As the state-of-the-art method for PF in SPSS are GV and MS (GV+MS), therefore the proposed PF method is applied to the source and spectral parameters obtained from GV and MS based PF (GV+MS+DYN). In this case, significant enhancement of performance can be observed from both the evaluation parameters, in both male and female cases. The proposed method is also integrated with the STRAIGHT vocoder as mentioned above. However, it yields limited improvement over STRAIGHT synthesized speech.

**Table 4.3:** Objective measure for different types of post-filtering methods.

| Method | Female (SLT) | | Male (BDL) | |
|---|---|---|---|---|
| | **PESQ-MOS** | **LSD** | **PESQ-MOS** | **LSD** |
| NONE | $0.66 \pm 0.12$ | $3.3 \pm 0.12$ | $0.76 \pm 0.15$ | $2.9 \pm 0.22$ |
| MCP | $0.79 \pm 0.28$ | $2.73 \pm 0.14$ | $0.82 \pm 0.23$ | $2.65 \pm 0.19$ |
| LSP | $0.75 \pm 0.23$ | $2.90 \pm 0.23$ | $0.80 \pm 0.15$ | $2.72 \pm 0.23$ |
| GV | $0.98 \pm 0.18$ | $2.34 \pm 0.18$ | $1.12 \pm 0.12$ | $2.12 \pm 0.20$ |
| MS | $0.83 \pm 0.13$ | $2.60 \pm 0.23$ | $0.95 \pm 0.21$ | $2.32 \pm 0.19$ |
| SRC | $0.79 \pm 0.35$ | $2.42 \pm 0.21$ | $0.93 \pm 0.13$ | $2.42 \pm 0.16$ |
| SPEC | $0.85 \pm 0.23$ | $2.38 \pm 0.15$ | $1.03 \pm 0.24$ | $2.31 \pm 0.12$ |
| DYN | $0.97 \pm 0.15$ | $2.29 \pm 0.13$ | $1.09 \pm 0.17$ | $2.02 \pm 0.11$ |
| GV+MS | $1.02 \pm 0.22$ | $2.31 \pm 0.19$ | $1.08 \pm 00.21$ | $2.01 \pm 0.15$ |
| GV+MS+DYN | $1.13 \pm 0.21$ | $2.22 \pm 0.20$ | $1.14 \pm 0.23$ | $1.18 \pm 0.23$ |
| STRAIGHT : GV+MS | $1.14 \pm 0.21$ | $2.19 \pm 0.17$ | $1.15 \pm 0.26$ | $1.17 \pm 0.17$ |
| **STRAIGHT : GV+MS+DYN** | $1.16 \pm 0.18$ | $2.18 \pm 0.23$ | $1.16 \pm 0.16$ | $1.17 \pm 0.18$ |

#### 4.5.3.3 Subjective evaluation

The subjective evaluation is carried out to derive the mean opinion score (MOS) based on the naturalness of the synthesized utterances. In this assessment, the methods considered for comparison are DYN, GV+MS, GV+MS+DYN, STRAIGHT: GV+MS and STRAIGHT: GV+MS+DYN for both SLT and BDL speakers. For each of the methods, 5 speech files are considered along with 5 natural speech data. Thereby, for each of the speakers, non-repetitive 30 speech files (corresponding to different utterances) are obtained. The speech files are randomly coded to avoid bias towards any of the methods. Total 10 subjects took part in this evaluation, who are research scholars having sound knowledge of speech perception. As in this assessment, we are considering only expert listeners, so the number of subjects is limited. The subjects were asked to provide scores against each speech file between 1 to 5 based on naturalness, where 5 corresponds to the best quality. Reference speech files were also provided to the subjects to get the knowledge of naturalness, which is the closeness of the speech file to that of natural speech. Total 50 scores are obtained for each type among the 6 sets. The scores are depicted in the boxplot with 95% confidence intervals as shown in Figure 4.10. Figure 4.10(a) and Figure 4.10(b) represent the scores obtained from BDL and SLT speaker, respectively. Corresponding mean and standard deviations are noted in Table 4.4, where we can observe an increase in MOS from $2.38 \pm 0.77$ to $3.27 \pm 0.51$ after applying the proposed dynamic PF over GV+MS method of PF, in case of the female speaker. However, this improvement is somewhat less in case of the male speaker,

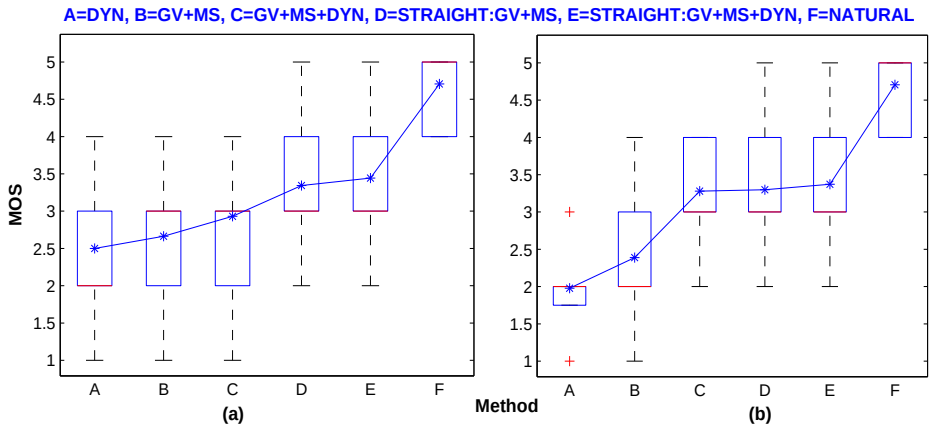


**Figure 4.10:** Boxplot representing MOS corresponding to naturalness for both the speakers. Naturalness for (a) male (BDL), (b) female (SLT). The mean values are represented by the connecting solid lines and median values by red lines in each subplot.

which is $2.66 \pm 0.82$ to $2.93 \pm 0.74$. Again, in case of STRAIGHT, the proposed algorithm seems to have limited gain in performance.

Between each pair of methods in Table 4.4, a series of pairwise t-test is performed to observe the differences between the distributions of MOS scores. All pairs are found to be significantly different at 1% level except DYN vs. GV+MS for the male speaker. This overlapping of the distributions signifies that the application of GV+MS method and proposed dynamic PF method provides an equivalent improvement in the synthesis quality. But in case of the female speaker, the DYN PF method performs poorer compared to GV+MS based PF. However, the application of DYN PF over GV+MS shows quite well in case of female speaker compared to that of the male speaker. Another pair of methods with less than 1% level of difference is STRAIGHT: GV+MS vs. STRAIGHT: GV+MS+DYN in case of both male and female. From our observation, it can be inferred that the impact of spectral PF is less significant in case of STRAIGHT spectrum. The spectral peaks in the STRAIGHT spectrum are more prominent compared to the peaks in the spectrum obtained from Mel-generalized cepstral analysis. Therefore, modifying the lower frequency peaks and valleys does not seem to have much effect on perception.

#### 4.5.3.4 Comparison with DNN based post-filtering

Although GV, MS, and MCP based PF methods are commonly used, recent advances in the literature show successful attempt to design post-filters using the deep generative architecture [81]. The proposed dynamic PF method is compared with DNN based PF, a similar architecture to [81] is

**Table 4.4:** Subjective evaluation result in terms of MOS for different types of PF methods.

| Method | Female (SLT) | Male (BDL) |
|---|---|---|
| DYN | $1.97 \pm 0.69$ | $2.50 \pm 0.79$ |
| GV+MS | $2.38 \pm 0.77$ | $2.66 \pm 0.82$ |
| GV+MS+DYN | $3.27 \pm 0.51$ | $2.93 \pm 0.74$ |
| STRAIGHT : GV+MS | $3.29 \pm 0.92$ | $3.34 \pm 0.92$ |
| STRAIGHT : GV+MS+DYN | $3.37 \pm 0.90$ | $3.45 \pm 0.78$ |
| NATURAL | $4.70 \pm 0.45$ | $4.758 \pm 0.49$ |

implemented for both male (BDL) and female (SLT) speakers using the same database for training and testing. The conditional probability of the parameters of natural speech given generated parameters can be modeled as a post-filter and used to modify the generated parameter sequences. It attempts to compensate the gap between natural and synthesized. This probabilistic post-filter is trained with deep feed-forward neural network using Theano in a GPU system [136]. We have used time aligned 100 parallel utterances from CMU arctic database for training of the post-filters. Both Mel-cepstral domain (35-dimensional) and spectral domain (1024-dimensional) post-filters are designed. In case of Mel-cepstral domain post-filter, 5 frames context is used resulting in a 175-dimensional feature vector. For training, a DNN with 3 hidden layers with 2048 hidden units in each layer.

In case of spectral domain, the spectrum is estimated from speech signal with 48 kHz sampling frequency, using Mel-generalized cepstral analysis [137] and warped in Bark scale before training. Only lower frequency part (upto 8 kHz) of the spectrum is considered for PF. The feature dimension is 350 (fast Fourier transform points) with 6 hidden layers and 2048 units in each layer. However, no context information is used in spectral domain due to the higher dimension. The mini-batch size was set to 10 during training. The learning rate was set to 0.0001 for all models. The momentum and weight decay were also employed to train the models. 250 epochs were executed in training.

The MGC contours derived from the DNN based post-filter and proposed dynamic post-filter along with corresponding natural and generated (GV+MS) contours are shown in Figure 4.11. It can be observed that the DNN based PF is also able to reduce the over-smoothing to a considerable extent. The preference test is conducted to see its impact on the perceptual quality of synthesized speech. In this evaluation, only spectral domain DNN PF is considered as it is reported to have better performance compared to that of cepstral domain. MLSA vocoder is used to synthesize speech corresponding to different PF approaches. The methods considered for the preference test are GV+MS, GV+MS+MCP, GV+MS+DNN, each of which is compared against GV+MS+DYN method. There

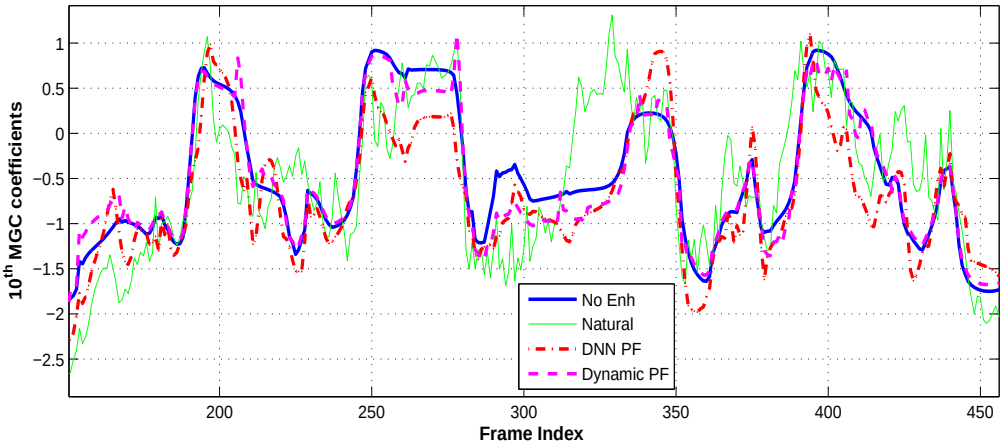


**Figure 4.11:** Contours of $10^{th}$ MGC sequence corresponding to natural, GV+MS and GV+MS+DYN for the utterance "We must give ourselves and not our money alone" of SLT speaker.

**Table 4.5:** Result of preference test in terms of % of preference.

| Female Speaker (SLT) | | | | | | |
|---|---|---|---|---|---|---|
| **Paired Test** | **GV+MS** | **GV+MS+MCP** | **GV+MS+DNN** | **GV+MS+DYN** | **No Pref.** | **p-value** |
| GV+MS Vs GV+MS+DYN | 17.5% | – | – | 78.7% | 3.7% | $8.99\times 10^{-4}$ |
| GV+MS+MCP Vs GV+MS+DYN | – | 25.0% | – | 67.5% | 7.5% | 0.01 |
| GV+MS+DNN Vs GV+MS+DYN | – | – | 31.2% | 61.2% | 7.6% | 0.06 |
| **Male Speaker (BDL)** | | | | | | |
| GV+MS Vs GV+MS+DYN | 25.4% | – | – | 65.5% | 9.1% | $4.5\times 10^{-4}$ |
| GV+MS+MCP Vs GV+MS+DYN | – | 32.3% | – | 62.6% | 5.1% | $1.3\times 10^{-2}$ |
| GV+MS+DNN Vs GV+MS+DYN | – | – | 41.3% | 47.5% | 11.2% | 0.05 |

are total 3 pairs of methods each having 40 pairs of utterances (20 for male and 20 for female). Total 20 subjects took part in this subjective evaluation, where each listener compared 120 utterances. They were asked to mark no preference against a pair, if both the speech files are of equivalent quality. The result of this preference test is shown in Table 4.4 in terms of preference %. The p-values obtained from double-tailed t-test between each pair are also shown. It can be observed that for both the speakers GV+MS+DYN method significantly outperforms GV+MS and GV+MS+MCP. And the corresponding p-values are found to be much lower. From the comparison between GV+MS+DYN and GV+MS+DNN, it can be observed that for the female speaker 61.2% utterances corresponding to the proposed method is preferred against 31.2% utterances obtained from GV+MS+DNN. However, in case of male speaker preference % is almost same for GV+MS+DNN and GV+MS+DYN. Also, % of no preference and p-values are comparatively higher in this case. From the above experiments, we

can say that the proposed dynamic PF improves the naturalness of synthesized speech compared to that of the GV+MS+MCP method. The MCP PF system also enhances the spectral peaks. However, it does not take into account the temporal over-smoothing effect. Although the spectral peaks are enhanced, some frames may be more enhancement, which may lead to unnatural quality. Apart from this, in addition to spectral modification, the proposed method also modifies $F_0$ and SoE aspects related to the excitation source, which may provide an additional advantage. In the preference test, the proposed method is preferred by 61% over the DNN based PF method, in case of the female speaker. In case of the male speaker, the preference for proposed dynamic PF method is 47%, while that DNN based PF method is 41%. This shows the efficacy of the dynamic PF using source and special parameters compared to the automatic data-driven PF method.

#### 4.5.3.5 Discussion

As shown in Table 4.4, there is no significant improvement in performance after applying the proposed PF method in case of the STRAIGHT vocoder. The MOS value for STRAIGHT:GV+MS+DYN is $3.37 \pm 0.90$, which is $3.29 \pm 0.92$ in case of STRAIGHT:GV+MS for female speaker. In case of MLSA vocoder after applying the proposed PF method the MOS value $3.27 \pm 0.51$ which is close to the STRAIGHT vocoder without applying the proposed PF method. Therefore, we can summarize that, the proposed method with MLSA vocoder (GV+MS+DYN) performs similar to STRAIGHT vocoder with GV+MS based PF (STRAIGHT:GV+MS). If we consider the computational complexity, in case of STRAIGHT we model MGC coefficients along with its delta and delta-delta (150-dimensional), BAP (26 dimensional) and $LF_0$ (1-dimensional) features. This results in a 177-dimensional feature vector to be modeled. Whereas, in case of MLSA vocoder with the proposed method (GV+MS+DYN), we model 21-dimensional sonority feature, 105-dimensional MGC coefficients with its delta and delta-delta and 1-dimensional $LF_0$, total 127-dimensional feature vector. Therefore, the proposed method have less computational complexity in statistical modeling compared to STRAIGHT vocoder, with same performance.

## 4.6 Spectral tilt based post-filtering

After performing the PF to conventional source and spectral parameters as mentioned above, another aspect that may effect the quality of synthesized speech is spectral tilt. This section focuses on the analysis of divergence in tilt or slope of vocal-tract spectrum extracted from natural and

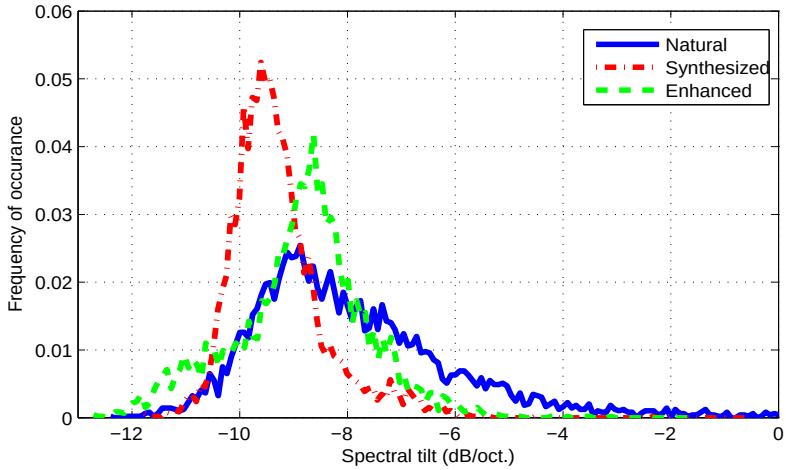


**Figure 4.12:** Distribution of spectral tilt (dB/octave) derived from LP spectrum for natural, synthesized and tilt enhanced voiced speech frames.

synthesized speech. More spectral tilt represents high difference between lower and higher frequency content. Alternatively, a flat spectrum represents near-uniform distribution of energy over the low and high frequency regions. [138] describes that, abrupt closure of vocal folds results in flat spectrum, by increasing loudness of produced speech signal. The spectral tilt has high impact on the perception of speech [139]. Figure 4.12 shows the distribution of spectral tilt (dB/octave) for voiced frames, which conveys that synthesized speech has strong negative slope compared to that of the natural. A method for compensating gap in spectral tilt between natural and synthesized speech may help to improve the quality of synthesized speech.

### 4.6.1 Analysis of spectral tilt in natural and synthesized speech

For comparison of natural and synthesized speech, HTS voices are developed, where context dependent quinphone HMMs are trained by using source and spectral parameters [40]. Here, spectral parameters are MGC coefficients, their delta and delta delta coefficients (105). Source parameters are $\log F_0$ and its dynamic features (3). Depending on the input text to be synthesized, context dependent HMMs are concatenated to get sentence HMM. Based on maximum likelihood criterion, spectral and source parameters are generated which are then applied to MLSA filter to derive the synthesized speech [134]. In this study, HTS voices are trained on CMU arctic database for four speakers, of which two are American male (BDL, RMS) and other two are American female (CLB, SLT) [48].

As the database contains 1132 sentences, 1000 sentences are used in training and the rest 132

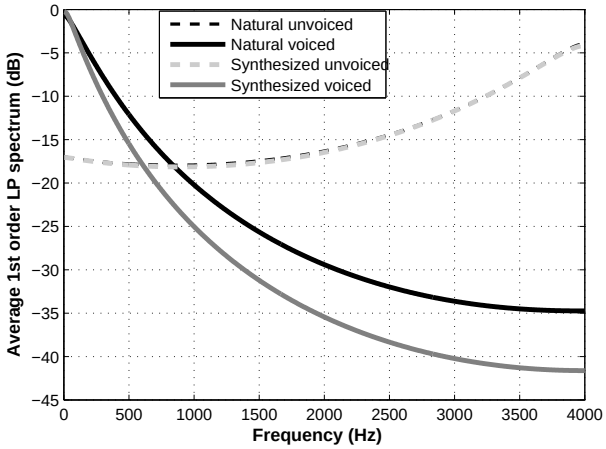


**Figure 4.13:** Average log frequency response of first order LP filter for voiced/unvoiced segments of synthesized and natural speech.

sentences are used for testing. In the interest of comparing spectral tilt of synthesized 132 sentences with that of corresponding natural utterances for different speakers, spectra are modeled using first order LP analysis. This spectrum carries information only about spectral tilt and not the formant peaks. For the same set of synthesized and natural speech files, voiced and unvoiced regions of synthesized and natural speech are analyzed separately to derive appropriate information regarding the change in spectral tilt and its effect on synthesis quality. Average of first order LP spectrum of all the voiced frames corresponding to natural and synthesized speech of SLT speaker is depicted in Figure 4.13. Similar average spectrum is also obtained for all unvoiced frames.

For voiced/ unvoiced detection, the method described in [88] is used. The speech signal is first passed through zero frequency resonator followed by trend removal termed as zero frequency filtered signal (ZFFS). The slope of positive zero crossings of ZFFS represent the SoE as stated in [140]. Frames which have majority SoE more than 1% of mean SoE, are considered as voiced and remaining as unvoiced. From Figure 4.13, it is apparent that there is a large gap in between slope of average spectrum of voiced frames of synthesized and natural speech. This in turn demonstrates that, spectral tilt of synthesized speech is much higher than that of the natural. As the VTS is modulated by the glottal flow spectra, the sharp discontinuities due to abrupt closing of vocal folds in voiced sounds are also reflected in the VTS of natural speech signal. This increases higher frequency energy of the spectrum. While, in case of synthesized speech due to averaging during modeling, these information are not captured well and results in strong negative slope of synthesized speech spectrum. Due to this fact, synthesized speech sounds muffled compared to natural. Again the abruptness of vocal-folds

**Table 4.6:** Spectral tilt (dB/oct.) for different sound categories in case of natural and synthesized speech.

| **Average spectral tilt (dB/oct.)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Speaker→** | **BDL** | | **CLB** | | **RMS** | | **SLT** | |
| **Category↓** | Nat. | Synth. | Nat. | Synth. | Nat. | Synth. | Nat. | Synth. |
| Low-vowels | -3.6 | -3.8 | -6.1 | -7.0 | -6.1 | -6.9 | -6.2 | -6.8 |
| Mid-vowels | -3.4 | -3.8 | -6.7 | -7.1 | -6.5 | -7.1 | -6.6 | -6.8 |
| High-vowels | -3.0 | -3.4 | -5.5 | -6.6 | -5.3 | -6.7 | -6.2 | -6.4 |
| Semi-vowels | -6.9 | -7.1 | -8.1 | -9.2 | -8.3 | -9.0 | -8.0 | -8.7 |
| Nasals | -6.2 | -7.8 | -7.5 | -8.9 | -8.1 | -9.2 | -7.8 | -9.2 |

during glottal closure due to change in vocal-tract constriction varies with different category of voiced sounds, which decreases in the following order: *low-vowels, mid-vowels, high-vowels, semi-vowels and nasals*. Therefore, the spectral tilt deviation between natural and synthesized speech may vary with these categories of sound units.

To further analyze its behavior on different voiced sound categories and different speakers, the average spectral tilt (dB/ octave) is measured for all the four speakers for both natural and synthesized case over the set of 132 test utterances. For each category of sound unit, LP analysis is performed with frame size 20 ms, frame shift 10 ms and LP order ($\frac{F_s}{1000} + 4$), where $F_s$ is sampling frequency. The average of derived LP spectrum is computed for all the frames corresponding to each category and a regression line is fit using least square error method. The slope of the regression line represents the spectral tilt. For each sound category of synthesized and natural speech, average spectral tilt for the same set of utterances are compared for the four speakers and shown in Table 4.6. This shows that, deviation in spectral tilt differs for different classes. In case of all four speakers, the spectral tilt values for low-vowels, mid-vowels and high-vowels are in similar ranges, whereas for semi-vowels and nasals, the spectral tilt is more negative. The traditional method of designing a high pass filter (HPF) with constant coefficient to flatten spectral tilt may not solve the problem.

### 4.6.2 Modification of spectral tilt

In this work, a novel method of spectral tilt modification is proposed. The first order LP filter can be used to model the spectral tilt of each frame of given speech signal. For each voiced frame in the synthesized and natural speech signal, the first order LP filter spectrum (coefficients) is computed as shown in (4.12) and (4.13). The error vector ($\boldsymbol{E_i}(w)$) is the difference between average power spectrum obtained from the first order LP coefficients of synthesized and natural speech. It is derived

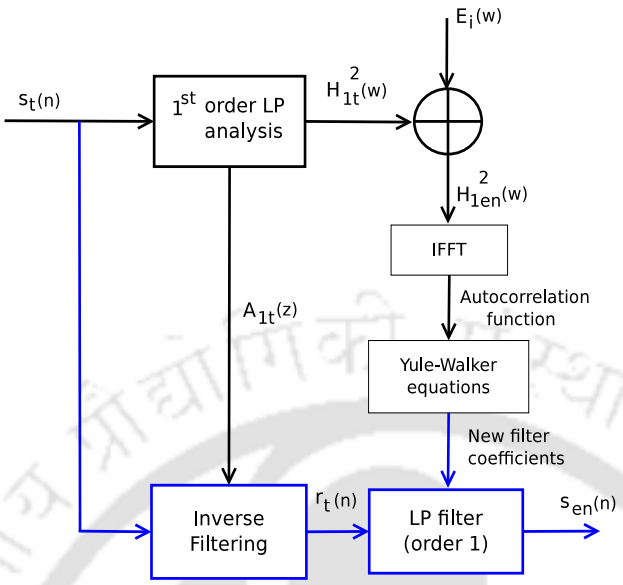


**Figure 4.14:** Block diagram of spectral tilt enhancement framework.

for each class of voiced sounds as shown in (4.14), where $\boldsymbol{E_i}(w)$ is the error vector for $i^{\text{th}}$ class of sound, $i = 1, 2, ..., 5$. $x_i$ and $y_i$ are the number of frames of $i^{\text{th}}$ category of voiced sound present in the database, in natural and synthesized speech respectively. This can be referred as training stage.

$$\boldsymbol{H_{1s}}(z) = \frac{1}{1 + a_s z^{-1}}, \tag{4.12}$$

$$\boldsymbol{H_{1n}}(z) = \frac{1}{1 + a_n z^{-1}}, \tag{4.13}$$

$$\boldsymbol{E_i}(w) = \frac{1}{x_i} \sum_{x_i} \boldsymbol{H^2_{ni}}(w) - \frac{1}{y_i} \sum_{y_i} \boldsymbol{H^2_{si}}(w), \tag{4.14}$$

In the enhancement step, given a synthesized utterance to be tilt modified, primitively voiced sound class information corresponding to current frame is retrieved from the corresponding label file. The power spectrum corresponding to first order LP coefficient of given synthesized speech signal frame is represented as $\boldsymbol{H_{1t}}(w)$ and the $1^{\text{st}}$ order inverse filter spectrum is represented as $\boldsymbol{A_{1t}}(w)$. The corresponding filter equations are shown in (4.15) and (4.16), respectively. Based on the class information ($i^{\text{th}}$ class), corresponding error vector $\boldsymbol{E_i(w)}$ is added to $\boldsymbol{H^2_{1t}}(w)$ of the frame to get $\boldsymbol{H^2_{1en}}(w)$. This is tilt modified first order LP power spectrum for a particular frame of $i^{\text{th}}$ class as shown in (4.17). The residual signal obtained by passing the test utterance frame through $\boldsymbol{A_{1t}}(z)$ is

$\boldsymbol{r_t}(n)$, which is spectral tilt subtracted speech signal.

$$\boldsymbol{H_{1t}}(z) = \frac{1}{1 + a_t z^{-1}}, \tag{4.15}$$

$$\boldsymbol{A_{1t}}(z) = 1 + a_t z^{-1}, \tag{4.16}$$

$$\boldsymbol{H^2_{1en}}(w) = \boldsymbol{H^2_{1t}}(w) + \boldsymbol{E_i}(w), \tag{4.17}$$

For each voiced frame in the test utterance, the above method is employed to derive modified power spectrum $\boldsymbol{H^2_{1en}}(w)$. The modified power spectrum is applied to inverse Fourier transformation (IFT) to obtain a new autocorrelation function. This autocorrelation function is used in the Yule-Walker equations to compute a new LP filter [120]. This kind of modification helps to change the slope of the spectrum, while the other formant informations remain intact in the residual ($\boldsymbol{r_t}(n)$). This residual signal $\boldsymbol{r_t}(n)$ is then passed through the new $1^{\text{st}}$ order LP filter to derive spectral tilt modified synthesized speech signal. Given a synthesized utterance, the proposed framework for spectral tilt modification can be implemented as shown in Figure 4.14.

The effectiveness of proposed method that reduces the spectral tilt difference between synthesized and natural speech, can be seen by comparing the LP spectrum of natural, synthesized and tilt modified speech signals depicted in Figure 4.15. This shows average log magnitude spectrum of normalized LP spectrum of natural, synthesized and modified speech signals for different classes of voiced sounds. As it is obvious from Figure 4.13, the average first order LP spectrum for all voiced sounds have significant difference of spectral tilt between natural and synthesized counterpart. The same can be inferred from Figure 4.15. In addition to this, Figure 4.15 gives the intuition that, this difference increases as we traverse from the low-vowels to nasals. In all the classes, the spectral tilt seems to be modified and becomes close to that of natural. Moreover, Figure 4.12 shows the distributions of spectral tilt derived from LP spectrum of natural, synthesized and tilt modified speech. It shows that, mean and variance of the distribution derived from tilt modified speech is close to that of natural speech.

### 4.6.3 Experimental evaluation

The efficacy of the proposed method is clearly evident from Figure 4.15 that shows LP spectrum for natural, synthesized and enhanced speech. Therefore, the proposed method has the ability to reduce spectral tilt to some extent. Besides flattening the spectrum to required extent, any other changes or distortions are not introduced. The enhanced HTS synthesized speech is compared with

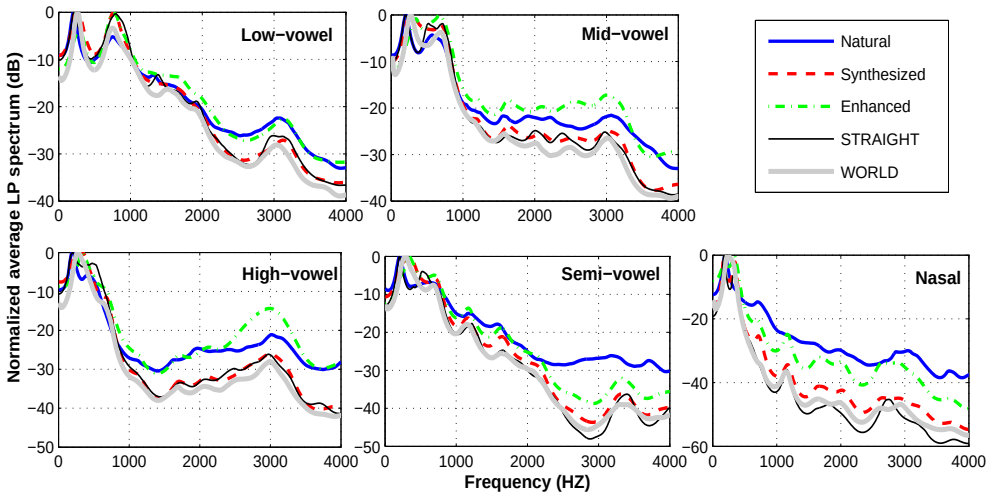


**Figure 4.15:** Average log magnitude normalized LP spectrum for different classes of voiced sounds for same set of natural, synthesized and enhanced speech.

the state-of-the art PF methods (GV and MS based).

#### 4.6.3.1 Objective evaluation

HTS systems are developed for the four speakers using following combinations of methods.

- HMM: HTS generated speech without any PF.
- HMM+GV: HTS generated speech using GV based PF.
- HMM+GV+MS: HTS generated speech using both GV and MS PF.
- HMM+GV+TM: HTS generated speech using GV based PF followed by proposed tilt modification (TM).
- HMM+GV+MS+TM: HTS generated speech using GV and MS PF followed by proposed TM.
- HMM+GV+MS+HPF: HTS generated speech using GV and MS based PF followed by HPF. A first order HPF with coefficient 0.9 is used in this case.
- WORLD+GV+MS: HTS generated speech using WORLD [141] vocoder with GV and MS PF.
- WORLD+GV+MS+TM: HTS generated speech using WORLD vocoder with GV and MS based PF followed by proposed TM.
- STRAIGHT+GV+MS: HTS generated speech using STRAIGHT [53] with GV and MS PF.
- STRAIGHT+GV+MS+TM: HTS generated speech using STRAIGHT with GV and MS based PF followed by proposed TM.

**Table 4.7:** Result for objective evaluation.

| Speaker→ | Male | | | | Female | | | |
|---|---|---|---|---|---|---|---|---|
| | BDL | | RMS | | CLB | | SLT | |
| System↓ | MOS | LQO | MOS | LQO | MOS | LQO | MOS | LQO |
| HMM | 0.94 | 1.16 | 1.19 | 1.24 | 0.67 | 1.10 | 0.85 | 1.01 |
| HMM+GV | 0.96 | 1.17 | 1.15 | 1.22 | 0.73 | 1.11 | 0.90 | 1.20 |
| HMM+GV+MS | 0.98 | 1.17 | 1.03 | 1.08 | 0.84 | 1.13 | 1.12 | 1.21 |
| HMM+GV+TM | 0.98 | 1.16 | 1.06 | 1.20 | 0.79 | 1.14 | 1.21 | 1.16 |
| HMM+GV+MS+TM | 1.10 | 1.22 | 1.24 | 1.25 | 0.89 | 1.15 | 1.20 | 1.24 |
| HMM+GV+MS+HPF | 0.96 | 1.17 | 1.02 | 1.12 | 0.78 | 1.13 | 0.92 | 1.21 |
| WORLD+GV+MS | 1.02 | 1.18 | 1.12 | 1.18 | 0.86 | 1.13 | 1.19 | 1.23 |
| WORLD+GV+MS+TM | 1.11 | 1.22 | 1.19 | 1.27 | 0.88 | 1.19 | 1.21 | 1.25 |
| STRAIGHT+GV+MS | 1.12 | 1.28 | 1.18 | 1.24 | 0.92 | 1.19 | 1.21 | 1.24 |
| STRAIGHT+GV+MS+TM | 1.17 | 1.29 | 1.23 | 1.34 | 1.12 | 1.24 | 1.23 | 1.29 |

The objective measure used for evaluation of different post-filtered synthesized speech are PESQ-MOS and PESQ-listening quality opinion (LQO) scores [135]. The evaluation is carried out for 58 randomly selected sentences and the scores corresponding to each speaker is shown in Table 4.7. The scores obtained from synthesized files of HMM+GV+TM are almost similar to that of HMM+GV+MS. This implies that the proposed TM method has comparable contribution compared to MS based enhancement over GV based enhanced speech files. Furthermore, HMM+GV+MS+TM outperforms HMM+GV+MS. The enhanced speech files obtained from HMM+GV+MS+HPF provide limited improvement over HMM+GV+MS. The proposed TM method also shows significant improvement, when applied to STRAIGHT and WORLD synthesized speech. It gives notable improvement in case of same speaker in both training and testing. However, use of a different speakers in training and testing also gives moderate improvement in the synthesized speech.

#### 4.6.3.2 Subjective evaluation

The subjective evaluation is performed on the SLT voice. As from the objective evaluation HMM+GV+MS is found to be comparable with HMM+GV+MS+TM, these two methods along with HMM+GV+MS+HPF speech files are used for the subjective evaluation. In the latter discussion, HMM+GV+MS, HMM+GV+MS+TM and HMM+GV+MS+HPF are referred as SYNTH, ENH and HPF respectively. A reliable subjective measure of intelligibility is preference test [142]. In this case, three evaluation parameters: intelligibility, naturalness and speaker similarity are used. Intelligibility refers to how well the message is conveyed, naturalness refers to closeness of reference

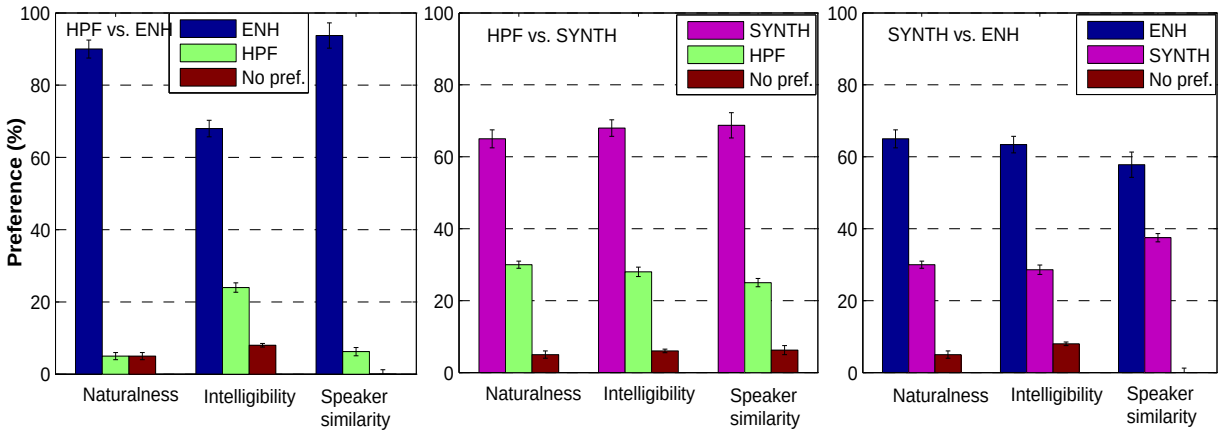


**Figure 4.16:** Result of preference test in terms of naturalness, intelligibility and speaker similarity. Here, SYNTH, ENH and HPF refers to HMM+GV+MS, HMM+GV+MS+TE and HMM+GV+MS+HPF respectively.

file with natural speech. Using the parameter speaker similarity, one can measure whether speaker information is intact after applying the proposed enhancement.

Total 20 sentences are randomly selected for this experiment. Corresponding to each sentence, there are three wave files from three enhancement methods. The three types of wave files for each sentence are played in pair for HPF vs. ENH, HPF vs. SYNTH and ENH vs. SYNTH. The subjects are asked to provide their preference three times per pairing, one for each evaluation parameter. For intelligibility, the subjects were asked to choose the utterance which they could understand with the least effort. For naturalness, they were asked to choose the utterance that was closer to the natural speech. For speaker similarity, first a natural speech is played for the same speaker and subjects were asked to provide their preference among the two wave files in terms of similarity in voice. Total 20 subjects are employed for this task who are native Indian speakers and have knowledge about speech perception. Utterances are unlabeled and played in random order to avoid bias towards any method. If a listener is unable to differentiate between any two wave files for some evaluation parameter, corresponding file is marked as no preference.

The result of the preference test is expressed in terms of percentage of wave files preferred for a method. Preference percentage for each evaluation parameter and each pair of methods along with percentage of no preference in each pair is shown in Figure 4.16. In the preference test for HPF vs. ENH, the latter one shows very good performance in terms of intelligibility, naturalness and speaker similarity. The improvements in naturalness and speaker similarity are particularly large. In other

**Table 4.8:** Result of preference test for babble noise, each for SNR 25 dB, 15 dB and 5 dB.

| **Preference (%)** | | | |
|---|---|---|---|
| **SNR** | **HMM+GV+MS+HPF vs. HMM+GV+MS+TM** | | |
| | HMM+GV+MS+HPF | HMM+GV+MS+TM | No preference |
| 25 dB | 35.0 | 60.0 | 5.0 |
| 15 dB | 48.7 | 50.4 | 1.3 |
| 5 dB | 53.8 | 41.2 | 5.0 |
| **HMM+GV+MS+HPF vs. HMM+GV+MS** | | | |
| | HMM+GV+MS+HPF | HMM+GV+MS | No preference |
| 25 dB | 38.6 | 55.4 | 6.0 |
| 15 dB | 45.3 | 48.5 | 6.2 |
| 5 dB | 55.5 | 44.5 | 0.0 |
| **HMM+GV+MS vs. HMM+GV+MS+TM** | | | |
| | HMM+GV+MS | HMM+GV+MS+TM | No preference |
| 25 dB | 40.0 | 60.0 | 0.0 |
| 15 dB | 48.5 | 50.2 | 1.3 |
| 5 dB | 51.7 | 48.3 | 0.0 |

two cases, HPF vs. SYNTH and ENH vs. SYNTH, SYNTH and ENH are preferred, where ENH is yielding large gain in terms of improving naturalness and intelligibility.

In order to observe the intelligibility gain by the proposed method in noisy environment, babble noise with SNR 5dB, 15dB and 25dB are added to the synthesized, enhanced and high pass filtered speech. The same preference test is carried out for each type of noise with different SNRs which is summarized in Table 4.8. This analysis in presence of noise shows that, as the level of noise increases, the preference percentage increases for HPF in both HPF vs. ENH and HPF vs. SYNTH cases, although the listeners did not prefer HPF in clean conditions. When comparing SYNTH vs. ENH, as the noise level increases, most of the preferences goes towards SYNTH cases. This implies, although in clean condition the HPF method for improving intelligibility is not preferable, in presence of noise it yields higher intelligibility. However, in this work the effort is made towards improving intelligibility, naturalness in clean environment which achieves good performance indeed.

## 4.7 Summary

In this chapter, an effort is made to improve the naturalness of synthesized speech by alleviating the difference in the source and spectral parameters between synthesized and natural speech. To accomplish this, a dynamic post-filter is proposed, which employs PF with varying factors for different

classes of sound units. The intelligibility and naturalness of speech signal are highly dependent on spectral peaks, valleys and SoE that correlate to sonority. Therefore, the sound units are classified into different sonorant classes and corresponding post-filters are applied to each sonorant category. As the sonority feature set is specifically designed to reflect the source and spectral prominence, an SVM classifier using these features is used to derive the class information. In the cases, where only synthesized speech is available without any label information and parameter contours, the proposed dynamic PF method can be reliably employed. The proposed method yields significant improvement in naturalness when combined with the state-of-the-art GV and MS based PF methods. Another method of PF that modifies the spectral tilt of synthesized speech is able to achieve improvement in terms of naturalness, intelligibility and speaker similarity.

# 5

# Significance of Sonority Information for Voiced/Unvoiced Decision

Contents

### Objective

*The quality of synthesized speech obtained from SPSS significantly relies on excitation source generation. Voiced/unvoiced decision is an essential component for generation of excitation source. It is obtained from $F_0$ and other excitation source evidence in the existing literature. The discontinuity at the point of contact in the vocal-folds excites energy into the vocal-tract resulting voicing effect in the produced speech signal. The perceptual reflection of voicing over the sound produced is correlated with the sonority information, which is related to less vocal-tract constriction and significant glottal vibration. Therefore, the possible variation in voicing with the change in supraglottal pressure due to vocal-tract constriction, rate of closing of vocal folds and regularity in structure of the signal are intact in the sonority associated with a sound unit. Voicing and degree of opening of vocal-tract are the two most effective correlate of sonority, that potentially contribute to the sonority hierarchy for sonorants and obstruents uniformly. Therefore, the voicing effect can be captured by the sonority measurement derived from system, source and suprasegmental information in the speech signal. In this chapter, a novel voiced/unvoiced decision method using sonority information is proposed and integrated in the SPSS framework for generation of excitation source. It leads to better voicing decision compared to the existing methods resulting in synthesized speech of improved quality, which is assured from objective and subjective analysis.*

## 5.1 Introduction

The SPSS is the state-of-the-art speech synthesizer in the recent literature that generates synthesized speech with sufficient intelligibility. It provides flexibility in terms of adaptation of statistical behavior of different speaking styles, emotions, speakers, languages [1, 2]; compression factor over the other speech synthesizer like USS [3, 4] and robustness [130]. However, the synthesized speech obtained from SPSS lacks naturalness compared to that of USS due to poor vocoder, deficient excitation source generation, inaccuracy in acoustic modeling and over-smoothing of the generated parameter sequences [5]. The rich characteristics of the speech signal intact in natural speech may not be adequately represented using only limited acoustic features modeled in the statistical environment. The naturalness in synthesized speech is mostly governed by the excitation source component. In the SPSS framework, along with other factors voiced/unvoiced decision plays a key role in excitation source generation module, which is basically impulse train for voiced frames and random noise for unvoiced

frames. Generally, $F_0$ information in employed to know voiced and unvoiced frames in the utterance to be synthesized. Improving the voicing decision may absolutely improve the excitation source and synthesis quality. There are several successful efforts in the literature towards this direction.

There are various time and frequency domain approaches in the literature for voiced/unvoiced detection [88]. These methods include the features related to production characteristics of voiced sounds such as energy, periodicity, short term autocorrelation, zero crossing rate, autocorrelation peak strength, harmonic measure from the instantaneous frequency amplitude spectrum [89,90]. These methods use some threshold obtained from empirical observation. To avoid such thresholding, statistical modeling based approaches have gained popularity using HMM, GMM, DNN [143]. These methods aim to explore better modeling technique using existing features and requires substantial amount of training data with manual segmentation. In conventional SPSS or HMM based speech synthesis, MGC coefficients are used to represent VTS information and $F_0$ is used to model the excitation source aspect. $F_0$ is modeled along with voicing decision using MSD-HMM and consequently error in $F_0$ estimation is propagated to the excitation source generation module [7, 8]. In the voiced regions $F_0$ is modeled as continuous Gaussian distribution and discrete symbol in unvoiced regions [8]. As the voicing decision is dependent on $F_0$, errors in $F_0$ estimation leads to error in MSD-HMM. This results in misdetection of voiced and unvoiced frames. In case of weekly voice sounds with lower energy like voiced fricatives, voiced stops and voiced affricates, the corresponding frames may get classified as unvoiced which will be propagated to the MSD-HMM in training.

During synthesis, for the frames corresponding to these sounds erroneous voicing information may be generated that results in creation of random noise as excitation. If these misclassification are repeating for several such frames in the same utterance, the naturalness of the synthesized speech severely degrades. On the other hand, for false voiced regions impulse sequence is used as excitation instead of random noise resulting in buzziness in synthesized speech. [93,94] exploited continuous $F_0$ model instead of MSD, where $F_0$ is always available for both voiced and unvoiced regions. In this case, the voicing decision is modeled in an independent stream. There are other attempts to integrate GMM and multilayer perceptron based voicing decision to make improvement in SPSS [95,96]. [98] proposed MSD-HMM modeling of voiced/unvoiced detection in SPSS by exploiting ZFF in $F_0$ estimation. The voicing detection is performed using heuristic threshold over the SoE and modeled by MSD-HMM along with $F_0$. In the current version of SPSS, voicing decision is pitch dependent and pitch estimation is

made using RAPT approach [87], which performs frame by frame autocorrelation analysis to capture the periodicity information. If the periodicity of the speech signal is not very distinctly evident, the autocorrelation based methods tend to give wrong voicing decision. As discussed in Chapter 2, the effect of wrong voicing decision can be prominently observed in corresponding synthesized speech signal. This leads to degradation in the perceptual quality of synthesized speech.

As voicing strength is generally governed by the movement of vocal-folds, most of the approaches use excitation source information as dominant feature to extract the voicing information. For most of the voiced sounds the main excitation occurs at the closing of the vocal folds. This is followed by the closed phase, where formants are the most prominent with high amplitude, slope and less bandwidth. Although the mechanism of source generation is independent of the vocal-tract shape, many studies have shown that with the variation in supra-glottal pressure due to vocal-tract constriction, the shape of glottal waveform specifically the amplitude changes [20]. Despite this change is not much significant in case of moderate constriction, as the constriction increases resulting in higher supra-glottal pressure, its effect on glottal waveform also increases. Therefore, the openness of vocal-tract may also play significant role in voicing strength as well as voicing decision. The sonority associated with a sound unit can be defined in terms of degree of vocal-tract constriction and voicing strength [9]. Based on voicing associated with the sonorant and obstruent sounds, the degree of associated sonority value changes. The sonority hierarchy can be seen in the increasing order of sonority as : *voiceless stops, voiceless fricatives, voiced stops, voiced fricatives, (voiced) nasals, voiced laterals, voiced r-sounds, (voiced) high vowels, (voiced) mid vowels, (voiced) low vowels* [9]. This correlation between voicing and sonority motivated us to explore the sonority information in the task of effective voicing detection.

In Chapter 3, a set of features that reflects the behavior of VTS, excitation source and suprasegmental information is proposed, which is termed as *sonority feature*. The feature set is specifically designed based on the changing vocal-tract constriction, excitation source strength and periodicity, with the change in articulators movement during production of different sonorant sound units. The sonority feature is capable of representing the sonority hierarchy and also found to be useful in classification of sonorants and obstruents. Therefore, it can be hypothesized that adopting the VTS information which has capability to correctly delineate sounds with varying vocal-tract constriction, along with the excitation source may further improve the accuracy of voicing decision. Most of the existing methods use only excitation source aspect for voicing decision. This work focuses on inte-

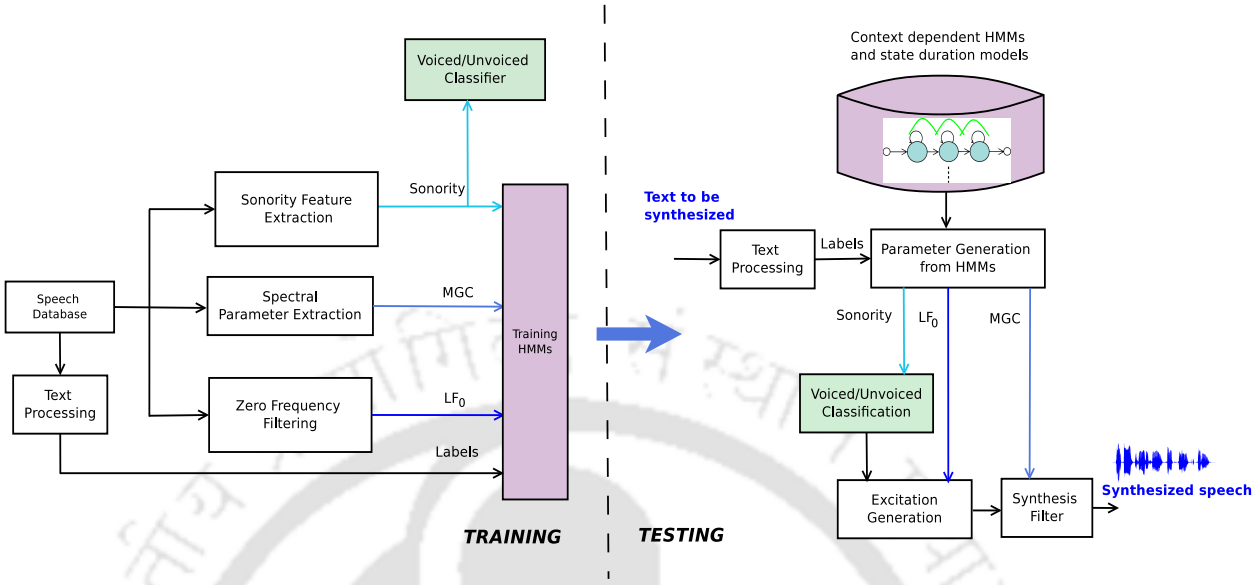


**Figure 5.1:** Block diagram representing the proposed framework.

gration of the sonority feature in the SPSS framework and explore its capability to improve voicing decision in the excitation source generation component during synthesis.

The rest of the chapter is arranged in the following sections. The integration of sonority feature in SPSS framework to derive voicing decision is explained in Section 5.2. Section 5.3 analyses efficacy of sonority feature in the task of voice/unvoiced detection. The effect of proposed voicing decision algorithm in synthesized speech is evaluated in Section 5.4.

## 5.2 Sonority feature in SPSS

The SPSS provides a unified framework to model vocal-tract, excitation and duration parameters simultaneously in HMM [8]. In this work, continuous $F_0$ modeling approach is followed instead of MSD in order to make the voicing decision independent of $F_0$ [93]. The HMM based TTS synthesis system is developed using the well known HTS toolkit that involves training and testing processes [8]. In the training process, excitation, vocal-tract and sonority features are extracted from the speech signal corresponding to the training database. Using these features all the phonemes are modeled using HMMs with 5 states. In each state, 3 streams are used to model different parameters extracted for each phoneme, where the first stream consists of vocal-tract parameters i.e. MGC coefficients including the zeroth coefficient and their delta, delta-delta coefficients. The source parameters, log $F_0$, its delta and delta-delta are modeled in a single (second) stream using continuous HMMs. In

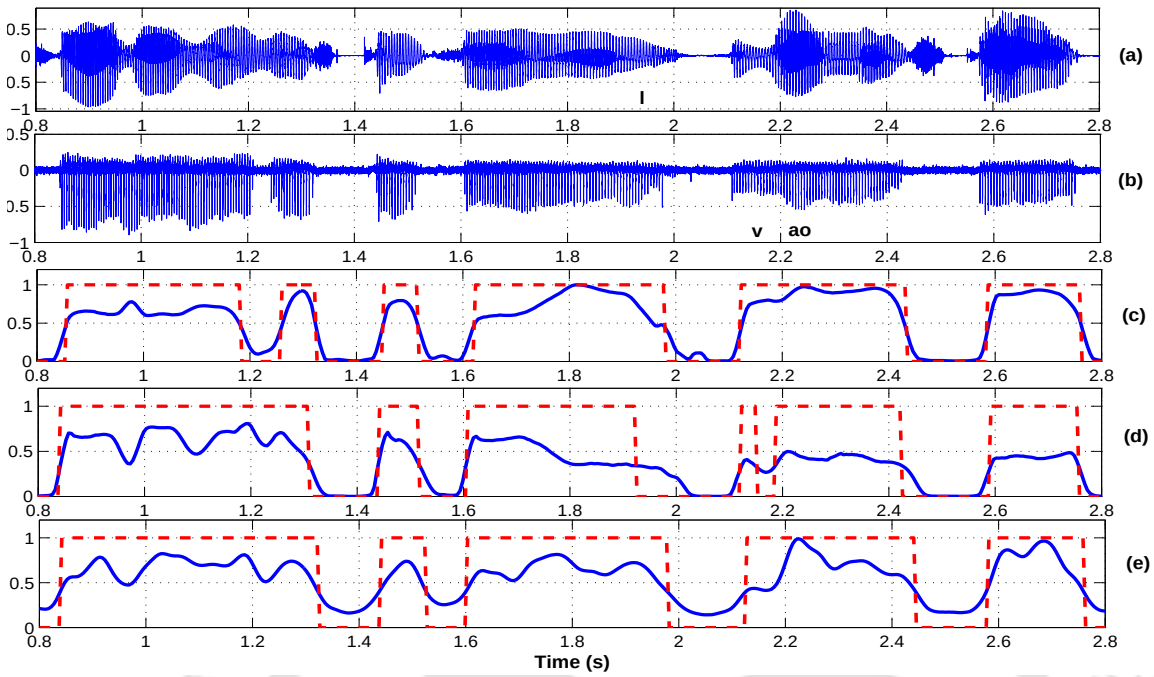


**Figure 5.2:** (a) Natural speech signal for the utterance, (b) DEGG signal, (c) strength of excitation derived from DEGG with reference voiced/unvoiced marking, (d) strength of excitation derived from speech signal and corresponding voiced/ unvoiced marking, (e) combined vocal-tract spectrum feature with corresponding voiced/unvoiced marking for the utterance "He was a head shorter than his companion, of almost delicate physique" for SLT speaker.

the third stream, the sonority feature along with its delta and delta-delta coefficients are modeled using continuous distribution. For each phoneme, parameters mentioned above are extracted along with their corresponding labels as shown in Figure 5.1. The log $F_0$ parameter is extracted using ZFF method as in [107] with 25 ms frame-length and 5 ms frame-shift. In the training part, the maximum likelihood estimation of each parameter is performed using BW re-estimation algorithm. During synthesis, as per input text, using the maximum likelihood parameter generation algorithm, frame-wise MGC coefficients, sonority feature and $F_0$ are computed by maximizing the output probability [8]. The generated sonority feature set $f_1$, $f_2$, $f_3$, $f_4$, $f_5$, $f_6$ and $f_7$ for an utterance "*become an automaton*" is shown respectively in dashed lines in Figure 4.1(b)-(h) of Chapter **??**. This feature set is further used in voiced/unvoiced classification as shown in Figure 5.1

## 5.3 Analysis of sonority feature in voiced/unvoiced detection

The sonority feature explained above has the ability to delineate the voicing associated with a sound unit. It represents source, system and supra-segmental information, among which excitation source

aspect is abundantly studied in the literature of voiced/unvoiced classification. Although the process of generation of excitation source is independent of the filter to a large extent, but due to the interaction between vocal-tract and source the VTS evidence in sonority feature also may give some additional information regarding voicing decision. As described in Chapter 3, the first 5 dimensions of the sonority feature represent VTS ($f_i$, $i = 1, 2..5$), that can be combined as $f_{VTS} = f_1+f_2+f_3+f_4+(1-f_5)$, where each feature is normalized to get the feature value in the range 0 to 1. As $f_5$ is inversely proportional to sonority as well as other features, we included $(1-f_5)$ during combination. The combined VTS feature for the utterance in Figure 5.2(a) is shown in Figure 5.2(e). Figure 5.2(b) and Figure 5.2(c) show corresponding DEGG and SoE derived from DEGG with reference voicing information. Figure 5.2(d) shows SoE derived from speech signal and corresponding voicing markings. In both Figure 5.2(c) and Figure 5.2(d), the frames with the feature value greater than mean value of the feature are selected as voiced and others are selected as unvoiced. If we compare Figure 5.2(c),(d),(e) the effectiveness of combined VTS evidence over SoE evidence can be observed, as it has similar behavior as that of SoE derived from DEGG. In fact, around 2.2 $s$ in Figure 5.2, some voiced frames are misclassified as unvoiced in case of SoE, whereas the VTS feature is able to correctly detect the corresponding voiced region. Therefore, impact of VTS evidence intact in sonority feature can lead to robust voiced/unvoiced detection if combined with SoE and periodicity.

To interpret the potency of sonority feature in voiced/unvoiced detection compared to that of $F_0$, the frames of the utterances corresponding to CMU arctic database (SLT speaker) are classified into voiced and unvoiced categories. The voiced categories include the frames of vowels, semivowels, nasals, voiced fricatives, voiced stops, voiced affricates and other frames are labeled as unvoiced except silence regions. Figure 5.3(a) to (g) elucidate distributions obtained for voiced and unvoiced frames corresponding to each dimension of the sonority feature and Figure 5.3(h) shows the same for $F_0$ obtained from RAPT algorithm. Significant deviation within distributions corresponding to each pair can be observed for all dimensions of the sonority feature. On the other hand, the distributions corresponding to $F_0$ in Figure 5.3(h) seems to have more overlapping area. This indicates relative inefficiency of $F_0$ in case of voiced/unvoiced detection compared to that of sonority feature. In case of $f_1$, $f_2$, $f_3$, $f_4$, $f_6$, $f_7$, it can be observed that the feature values for voiced frames are higher than that of the unvoiced frames in average sense. While in case of $f_5$, average value corresponding to the unvoiced is higher that that of the voiced. This is due to the fact that, the bandwidth parameter

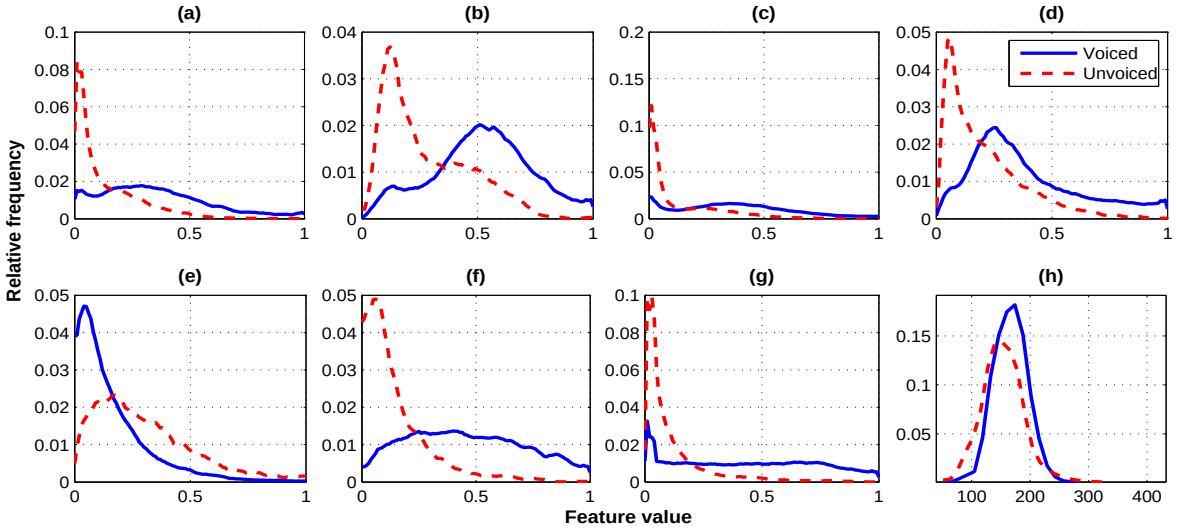


**Figure 5.3:** Distributions of sonority feature and $F_0$ for voiced and unvoiced frames of arctic database of SLT speaker; (a) $f_1$, (b) $f_2$, (c) $f_3$, (d) $f_4$, (e) $f_5$, (f) $f_6$, (g) $f_7$, (h) $F_0$.

represented by $f_5$ is inversely proportional to the sonority hierarchy, whereas all other dimensions are directly related to the degree of sonority. One useful way to quantify the distance between a pair of Gaussian probability density function is KLD measure [122] as given by (5.1).

$$D_{KL}(A,B) = \frac{1}{2}\left\{\frac{\sigma_A^2}{\sigma_B^2} + \frac{\sigma_B^2}{\sigma_A^2}\right\} - 1 + \frac{1}{2}\{\mu_A - \mu_B\}^2\left\{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}\right\}, \tag{5.1}$$

where, $A$ and $B$ are two univariate Gaussian distributions with means $\mu_A$, $\mu_B$ and standard deviations $\sigma_A$, $\sigma_B$ respectively. Here $A$ and $B$ represent samples of one feature for voiced and unvoiced classes respectively. For each pair of the distributions corresponding to each feature, KLD is calculated and depicted in Table 5.1. It can be observed from Table 5.1 that apart from $f_5$, the KLD values corresponding to other sonority features are higher compared to that of $F_0$. To include the efficacy of each dimension of sonority feature for voiced/unvoiced detection, corresponding weights ($w_i$) are derived from the KLDs, such that

$$\sum_{i=1}^{7} w_i = 1, \tag{5.2}$$

where,

$$w_i = \frac{[D_{KL}(A,B)]_{f_i}}{\sum_{i=1}^{7}[D_{KL}(A,B)]_{f_i}}, \tag{5.3}$$

**Table 5.1:** KLD measures corresponding to different features between normal distributions of voiced and unvoiced frames for CMU arctic database (SLT speaker).

| Feature | KLD | $w_i$ |
|---|---|---|
| $f_1$ | 2.06 | 0.14 |
| $f_2$ | 1.47 | 0.09 |
| $f_3$ | 1.89 | 0.13 |
| $f_4$ | 1.04 | 0.07 |
| $f_5$ | 0.72 | 0.05 |
| $f_6$ | 3.48 | 0.23 |
| $f_7$ | 4.27 | 0.29 |
| $F_0$ | 0.78 | - |

The weights assigned to each of the seven features according to their potential to classify different voiced and unvoiced frames are also shown in Table 5.1. Thus a competent representation of sonority feature in term of voiced/unvoiced classification is derived in this work. Further to analyze the efficacy of the sonority feature with respect to different voiced classes, the same distributions are derived for vowels, semivowels, nasals, voiced fricatives, voiced stops, voiced affricates and unvoiced individually. The KLD measures between the distributions of each of the voiced sound category and that of unvoiced are determined for each dimension of sonority feature. The average KLD across all the 7 dimensions are obtained for each voiced sound category. Similarly, the KLD distances are derived between distributions of $F_0$ extracted from each voiced sound category and unvoiced frames. These values are shown in Table 5.2. This gives information regarding ability of the sonority feature to differentiate a particular category of voiced sounds with the unvoiced. The more the KLD value is, more is the capability of the feature to correctly classify corresponding categories against unvoiced sounds. It can be observed from Table 5.2 that $F_0$ has higher ability to differentiate voiced fricatives, voiced stops, voiced affricates with unvoiced. However, in most of the databases or in common use, these phonemes are less likely to occur compared vowels, semivowels and nasals. Therefore, it will have minimal effect on the overall classification performance, which is assured from further experiments.

### 5.3.1 Voiced/unvoiced classification using sigmoidal function

Before proceeding to train a statistical classifier using the sonority feature, we want to emphasize the ability of the feature based on empirical methods of voiced/unvoiced classification. These classifiers are beneficial over the statistical classifiers, when there is no training data available. Given a particular

**Table 5.2:** Average KLD measure corresponding to sonority features and $F_0$ between normal distributions of different categories of voiced and unvoiced frames for CMU arctic database (SLT speaker).

| Category | Sonority | $\mathbf{F}_0$ |
|---|---|---|
| Vowels | 2.37 | 0.55 |
| Semivowels | 2.51 | 1.32 |
| Nasals | 3.09 | 2.62 |
| Voiced Fricatives | 0.60 | 1.00 |
| Voiced Stops | 1.02 | 1.33 |
| Voiced Affricates | 0.36 | 1.01 |

speech signal the 7-dimensional sonority feature is extracted and each dimension is normalized to the range 0 to 1. As bandwidth ($f_5$) is inversely proportional to the sonority feature, $(1 - f_5)$ is considered as the $5^{\text{th}}$ dimension. The feature contour corresponding to each dimension is then multiplied with corresponding weight given in Table 5.1. All the seven weighted features are added together to derive the combined sonority feature ($S(n)$) as shown in Figure 5.4(d) corresponding to the speech signal shown in Figure 5.4(a). It is troublesome to set the threshold directly on $S(n)$ in Figure 5.4(d) due its high variability, therefore a gross level feature is derived by passing the normalized $S(n)$ through a sigmoidal function, given by

$$S_g(n) = (1 - S_{gm})\frac{1}{1 + \exp(-\lambda(S(n) - T_h))} + S_{gm}, \tag{5.4}$$

where, $S_g(n)$ is the sigmoidal function of $S(n)$, $\lambda$ is the slope parameter of the sigmoid which is set to 20, the threshold $T_h$ is derived from the mean value of $S(n)$. The minimum value of the sigmoid function $S_{gm}$ is set as 0. The obtained gross level sonority feature, $S_g(n)$ can be used for voiced/unvoiced classification by setting a suitable threshold. In this case, the frames with values of $S_g(n)$ greater than 20% of its maximum value are considered as voiced and others are classified as unvoiced. Figure 5.4(e) shows $S_g(n)$ obtained from $S(n)$ shown in Figure 5.4(d). It can be observed that the gross shape of the sonority feature after passing through the sigmoid function is highly correlated with the SoE contour in Figure 5.4(c), derived from the DEGG signal in Figure 5.4(b). The corresponding speech signal is shown in Figure 5.4(a). The reference voicing boundaries are also shown in Figure 5.4(b). From the comparison of reference boundaries and the obtained voicing decision shown in Figure 5.4(c), it is observed that voicing decision can be precisely obtained from $S_g(n)$ by setting a suitable threshold. The performance voiced/unvoiced classification is evaluated in

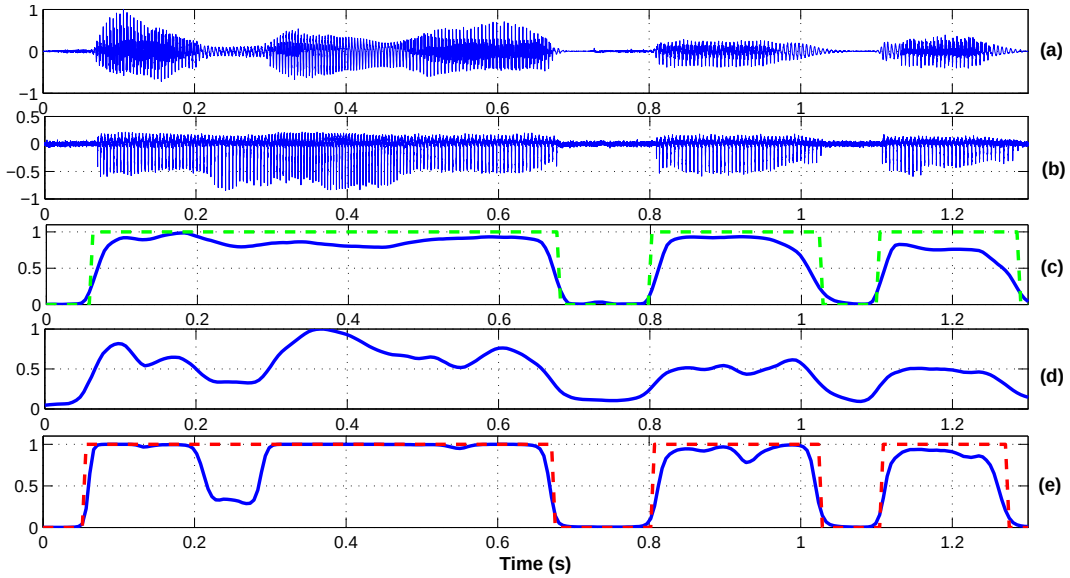


**Figure 5.4:** (a) Natural speech signal, (b) corresponding DEGG signal, (c) strength of excitation derived from DEGG along with reference voiced/unvoiced marking, (d) combined sonority evidence, (e) sonority evidence after passing through a sigmoid function with derived voiced/unvoiced marking; for the utterance "Hardly were our plans made public before we were met by powerful opposition" for SLT speaker.

terms of voicing error ($V_E$) and unvoicing error ($U_E$) between reference and detected voicing regions.

$$V_E = N_V/N_{ref};\ U_E = N_U/N_{ref}, \tag{5.5}$$

$$VU_E = V_E + U_E, \tag{5.6}$$

where, $N_V$ is the number of voiced frames classified as unvoiced, $N_U$ is the number of unvoiced frames classified as voiced, $N_{ref}$ the total number of frames present in the testing database and $VU_E$ is the total voicing error. A reliable performance is obtained by applying the combined sonority feature to the sigmoidal function with $V_E = 2.98\%$ and $U_E = 2.22\%$ for the SLT speaker in CMU arctic database. The performances of this empirical method for BDL speaker as well as other standard databases are shown in Table 5.3, which shows that the empirical method using sonority feature performs consistently well for all the databases. Therefore, the sonority feature can be further exploited to develop a higher level classifier for voiced/unvoiced detection.

### 5.3.2 Voiced/unvoiced classification using SVM

To avoid heuristic thresholding, the derived sigmoidal signal ($S_g(n)$) is applied to an SVM classifier (with RBF kernel). Using the normalized $S_g(n)$ feature, 5-fold cross-validation is performed to derive the appropriate values of $c$ and $\gamma$. The SVM model is trained for randomly chosen 80% frames of the

**Table 5.3:** Comparison of the different methods for voicing detection in terms of the percentage of voicing error ($V_E$) and unvoicing error ($U_E$).

| Database → | SLT | | | BDL | | | KEELE | | | CSTR (RL) | | | CSTR (SB) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method ↓ | $V_E$ | $U_E$ | $VU_E$ | $V_E$ | $U_E$ | $VU_E$ | $V_E$ | $U_E$ | $VU_E$ | $V_E$ | $U_E$ | $VU_E$ | $V_E$ | $U_E$ | $VU_E$ |
| **Sigmoid** | 2.98 | 2.22 | 5.20 | 7.60 | 4.56 | 12.16 | 6.43 | 2.59 | 9.02 | 1.99 | 8.47 | 10.46 | 7.32 | 5.06 | 12.38 |
| **SVM Sigmoid** | 2.20 | 2.40 | 4.60 | 4.57 | 3.42 | 7.99 | 6.40 | 5.05 | 11.45 | 6.62 | 7.15 | 14.37 | 6.72 | 3.91 | 10.63 |
| **SVM (Proposed)** | **1.62** | **1.27** | **2.89** | **1.47** | **1.72** | **3.19** | **2.05** | **1.95** | **4.00** | **2.77** | **4.10** | **6.87** | **2.70** | **2.60** | **5.30** |
| **RAPT** | 3.26 | 4.26 | 7.52 | 7.16 | 4.51 | 11.67 | 7.22 | 2.02 | 9.24 | 6.21 | 1.20 | 7.41 | 2.53 | 1.00 | 3.53 |
| **STRAIGHT** | 5.39 | 8.41 | 13.80 | 8.77 | 4.21 | 12.98 | 6.49 | 4.25 | 10.74 | 10.82 | 4.27 | 15.09 | 1.27 | 3.40 | 4.67 |
| **SRH** | 3.91 | 5.02 | 8.93 | 16.09 | 5.25 | 21.34 | 11.4 | 9.45 | 20.85 | 12.22 | 6.64 | 18.86 | 4.14 | 15.53 | 19.67 |
| **REAPER** | 1.61 | 4.94 | 6.55 | 2.39 | 7.55 | 9.94 | 8.58 | 6.57 | 15.15 | 2.03 | 5.41 | 7.44 | 3.30 | 2.31 | 5.61 |

database and the rest 20% frames are used for testing. From the predicted voiced/unvoiced labels during testing, the error values $V_E$ and $U_E$ are obtained. The classification errors using this method (SVM Sigmoid) is shown in Table 5.3 for different databases. In most of the cases, improvement can be observed over the empirical method (Sigmoid). To achieve better classification accuracy, SVM based classifier is developed using the normalized 7-dimensional sonority feature (SVM Proposed) with RBF kernel and $c$, $\gamma$ values are obtained in same way as mentioned above. The voiced/unvoiced classification performance obtained using this classifier is superior compared to other methods as observed from Table 5.3. The proposed method is compared with the state-of-the-art methods RAPT, STRAIGHT [53], SRH [101] and REAPER [100] for the databases CMU arctic (BDL and SLT) [48], KEELE [144] and CSTR (RL and SB) [145]. The improved performance depicted in Table 5.3 assures that the proposed voiced/unvoiced classification can be integrated with the SPSS framework to obtain better excitation source during synthesis.

The overall framework can be seen from the block diagram shown in Figure 5.1. In the SPSS framework, along with extraction of the generic acoustic features like MGC coefficients and $LF_0$, the sonority feature is also extracted from speech signals corresponding to the training database. All the features are modeled using HMM as mentioned in Section 5.2 in the training phase. During testing, given the text to be synthesized using the trained HMMs MGC coefficients, $LF_0$ and sonority features are generated. The sonority feature is fed to the previously trained SVM classifier to obtain the voicing decision corresponding to all the frames. This voicing decision along with continuous $LF_0$ values are used in excitation source generation component. The resultant excitation source along with MGC coefficients are passed through the synthesis filter (vocoder) to derive the synthesized speech.

## 5.4 Experimental observations

To evaluate the performance of the proposed voiced/unvoiced detection method in terms of quality of synthesized speech, HMM based SPSS systems are developed using CMU arctic database for SLT and BDL speakers. As the database contains 1132 utterances corresponding to each speaker, 1000 utterances are used in training of SPSS system while the rest 132 sentences are used in testing. The sonority feature along with log $F_0$ and MGC coefficients are extracted from the speech signal with 25 ms frame-size and 5 ms frame-shift. In case of the proposed method, the $F_0$ parameter is extracted using ZFF algorithm [107] with continuous values in both voiced and unvoiced frames and continuous $F_0$ modeling approach is followed [93]. For comparison RAPT, REAPER, SRH and STRAIGHT methods of voicing decision are also implemented. For these cases the same $F_0$ extracted from ZFF is used, only voicing decision is adopted from individual methods. Except the proposed method, MSD-HMM is used to model $F_0$ as well as voicing decision for all other approaches. In both RAPT and REAPER methods autocorrelation analysis of speech signal and dynamic programming are used to derive the voicing decision. SRH uses harmonic estimation of the residual signal for pitch tracking and local thresholding of pitch contour is used for voicing decision. The STRAIGHT algorithm employs wavelet transformation approach, where the voicing strength in different bands is used to derive the voicing decision. In this work, version 40 of STRAIGHT is used. For deriving the synthesized speech using the excitation source and spectrum MLSA filter is used. As the excitation source model used in conventional MLSA based SPSS framework is simple impulse and noise based, it is greatly effected by wrong voicing decision. However the advanced excitation generation algorithms like STRAIGHT [53] that uses both periodic and aperiodic components along with the phase component, may be comparatively less prone to voicing decision. Therefore the same voicing decision algorithms are also employed in the STRAIGHT based SPSS framework. In this case, only the voicing decision and $F_0$ information is being adapted from the proposed as well as RAPT, REAPER, SRH and STRAIGHT algorithms and ZFF respectively, whereas all other modules remain same as [53]. In case of STRAIGHT, along with traditional SPSS parameters 25-dimensional BAP are also trained in a separate stream. Therefore, there are total 4 streams in this case. Apart from this, the training process remains same as described in Section 5.2. During testing, from the generated continuous $F_0$, voicing decision obtained from generated sonority feature and aperiodicity component, excitation source is generated. Using this excitation source and MGC coefficients synthesized speech is obtained from STRAIGHT vocoder

algorithm. Due to the complex analysis synthesis procedure in STRAIGHT, the synthesized speech is of higher naturalness. The synthesized speech files obtained from both MLSA and STRAIGHT based SPSS with different voicing decision algorithms for both SLT and BDL speakers are considered for objective and subjective evaluations. The experiments establishes the efficacy of the proposed method.

### 5.4.1 Objective evaluation

The objective measures used for evaluation of synthesized speech using different voicing algorithms are $VU_E$ and LSD. As the length of natural and corresponding synthesized speech files are not same, frame level alignment between natural and synthesized speech is performed using DTW before calculating these measures. In case of each speaker and each of the voicing algorithms such as proposed, RAPT, REAPER, SRH and STRAIGHT for both MLSA and STRAIGHT synthesizers, 132 test utterances are used. The average $VU_E$ and LSD across 132 utterances for each method are shown in Table 5.4. As the accuracy of voicing decision improves, $VU_E$ and LSD decrease. It can be observed from Table 5.4 that in most of the cases synthesized speech using the proposed and STRAIGHT based voicing decision methods achieve better performance compared to others. The performance of SRH method is poorer compared to other methods. Another observation is that RAPT and REAPER have similar trend of performance. Compared to the MLSA synthesized speech, the LSD values for all methods are less in case of STRAIGHT synthesizer. This is due to higher performance of STRAIGHT vocoder compared to MLSA.

### 5.4.2 Subjective evaluation

As seen from the objective evaluation, voicing error during excitation source generation has a significant impact on naturalness of synthesized speech. In the objective evaluation, the SRH method performs poorer compared to all other, while performance of REAPER is similar to that of RAPT to some extent. As in the subjective evaluation, number of subjects is limited and listening too many speech files may be a tedious task, therefore only proposed, RAPT and STRAIGHT based methods of voicing decision are considered in this case. Two types of subjective evaluations are carried out namely MOS and PT for all the three methods. Total 10 subjects took part in this evaluation, who are research scholars having sound knowledge of speech perception. In case of MOS evaluation, the subjects were asked to provide a score against each speech file between 1 to 5 based on the naturalness, where 5 corresponds to the best quality and 1 for the least. Reference speech files were

**Table 5.4:** Objective evaluation for synthesized speech using different voiced/unvoiced decision methods with MLSA and STRAIGHT vocoder.

| Vocoder ↓ | Voicing ↓ Method | SLT | | BDL | |
|---|---|---|---|---|---|
| | | $VU_E$ | LSD | $VU_E$ | LSD |
| MLSA | Proposed | 6.52 | 2.55 | 5.96 | 3.17 |
| | RAPT | 12.21 | 2.52 | 10.32 | 3.26 |
| | REAPER | 10.11 | 2.96 | 8.93 | 3.37 |
| | SRH | 15.50 | 2.85 | 11.90 | 3.34 |
| | STRAIGHT | 6.99 | 3.15 | 6.22 | 3.24 |
| STRAIGHT | Proposed | 5.41 | 2.14 | 4.35 | 2.41 |
| | RAPT | 10.52 | 2.19 | 6.22 | 2.57 |
| | REAPER | 8.15 | 2.18 | 6.54 | 2.51 |
| | SRH | 11.20 | 2.31 | 8.79 | 2.52 |
| | STRAIGHT | 3.25 | 2.13 | 4.32 | 2.47 |

also provided to the subjects to get the knowledge of naturalness, which is defined as the closeness of the synthesized speech to natural speech. For each method 5 speech files along with 5 natural speech files corresponding to different utterances are considered for deriving MOS. By considering both MLSA and STRAIGHT vocoder approaches, there are total 35 speech files for each of the female (SLT) and male (BDL) speakers. The speech files were randomly coded to avoid bias towards any method. Thus total 50 scores are obtained for each type among the 7 sets. Table 5.5 depicts mean values of the scores corresponding to each method. It can be observed that the proposed method achieves MOS of 2.5 against 1.7 using RAPT algorithm for female speaker. In case of male speaker, the improvement lower, which is 3.0 against 2.2 using RAPT. Again, for STRAIGHT vocoder, the improvement is limited. The MOS value is 3.5 using proposed method, whereas it 2.4 and 3.5 using RAPT and STRAIGHT based voicing decision. In this STRAIGHT based voicing decision achieves better than RAPT and equivalent to the proposed method. Although the performance of the proposed method seems to be not improved to a significant extent, however the proposed method is convenient in terms of computational complexity compared to STRAIGHT.

Also to show the overall distribution, boxplot corresponding to the scores are shown in Figure 5.5. In both the cases, it is evident that after applying the proposed voicing decision, the naturalness of the synthesized speech increases significantly. Although mixed excitation source is used in case of STRAIGHT vocoder, from Table 5.5 and Figure 5.5 we can infer that the synthesized speech quality

**Table 5.5:** Subjective evaluation result for mean opinion score.

| Vocoder | Speaker | Proposed | RAPT | STRAIGHT |
|---|---|---|---|---|
| MLSA | SLT | 2.5 | 1.7 | 1.4 |
| | BDL | 3.0 | 2.2 | 2.1 |
| STRAIGHT | SLT | 3.8 | 2.4 | 3.5 |
| | BDL | 3.4 | 2.3 | 2.1 |

still depends on voicing decision. In case of the MLSA vocoder, the improvement is found to be more for the female speaker, whereas for STRAIGHT vocoder, male speaker is found to achieve higher performance after using the proposed method.

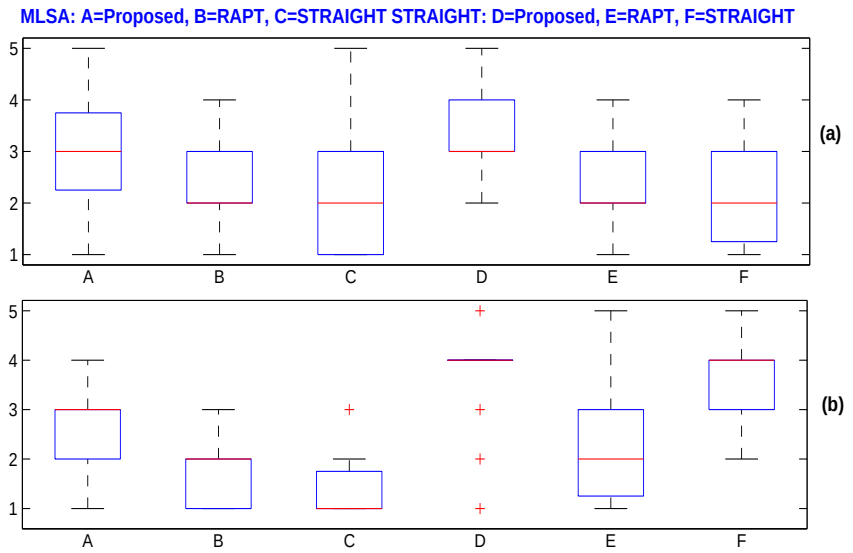


**Figure 5.5:** Boxplot representing distribution of mean opinion scores of different methods for (a) SLT, (b) BDL speaker.

In the PT, 10 pairs of speech files corresponding to 5 for proposed vs. RAPT and 5 for proposed vs. STRAIGHT, for each of MLSA and STRAIGHT based vocoders are played to the subjects. Therefore each subject listened to of 20 pairs (10 for MLSA vocoder and 10 for STRAIGHT vocoder) of speech files in random order for each of SLT and BDL speakers. Therefore each subjected evaluated total 40 pairs of speech files. They were asked to listen the speech files pairwise and provide their preference within each pair. No preference is given if both the speech files seem to have same perceptual quality. Table 5.6 shows the mean values of percentage of speech files chosen by the listeners corresponding to each method. It can be observed from Table 5.6 that, in each case the preference % for the proposed method is more than 65%, except the case of BDL speaker for STRAIGHT vocoder. In case of SLT

**Table 5.6:** Result of preference test represented in terms of % of speech files subjects have preferred while comparing proposed method with RAPT and STRAIGHT method for voicing decision.

| Vocoder | Speaker | Proposed | RAPT | STRAIGHT | No pref. |
|---|---|---|---|---|---|
| MLSA | SLT | 71.4 | 11.4 | - | 17.2 |
| | | 88.6 | - | 5.7 | 5.7 |
| | BDL | 65.7 | 20 | - | 14.2 |
| | | 88.6 | - | 2.9 | 8.6 |
| STRAIGHT | SLT | 94.3 | 5.7 | - | 0 |
| | | 71.4 | | 11.4 | 17.1 |
| | BDL | 45.7 | 45 | - | 9.2 |
| | | 42.8 | - | 28.6 | 28.6 |

speaker MLSA vocoder, the proposed is preferred 71.4% compared 11.4% for RAPT. In same case the proposed method is preferred 88.6% over 5.7% for STRAIGHT based voicing decision method. In case of STARIGHT vocoder, the preference % is comparatively less for the proposed method. The proposed sonority based voicing decision achieves comparatively lower performance in case of BDL speaker STRAIGHT vocoder. Although the preference % is higher for the proposed method, the % of no preference is increasing this case. In both the subjective evaluations, the proposed method is found to have remarkable impact on the synthesized speech quality.

## 5.5 Summary

The contributions made in this chapter explore the scope of sonority as acoustic-phonetic correlate of voicing decision. The conventional voiced/unvoiced detection algorithms use excitation source based features by considering that voicing is independent of the vocal-tract information. However, due to source and system coupling during open phase of glottal vibration, the voicing may also be associated with the extent of vocal-tract constriction. The sonority aspect associated with a segment of speech signal can be represented as a combination of source, spectral and suprasegmental evidence. Moreover sonority is highly correlated with both vocal-tract constriction and voicing strength. Based these facts, the voiced/unvoiced classification is performed using sonority feature and it is found to yield improvement as compared to the existing methods. Both empirical and statistical classifications are presented in this work. The proposed voiced/unvoiced classification algorithm using sonority feature achieved better accuracy compared to the existing methods. As the voicing decision plays a great role in synthesized speech quality obtained from SPSS, therefore the proposed voicing decision algorithm

using sonority feature is integrated with the SPSS framework. With the increase in accuracy of voicing decision, the naturalness quality of synthesized speech is also found to be improved from both objective and subjective evaluations.

# 6

# Combined Framework for Improving Synthesized Speech

Contents

### Objective

*The objective of this chapter is to bring out a combined framework for improving quality of synthesized speech obtained from SPSS using sonority information. The degradation in the quality of synthesized speech of SPSS in terms of naturalness is a consequence of different issues involved with the individual modules of the framework. In the previous chapters, parameterization of speech signal in terms of sonority feature along with the conventional source and spectral features is proposed. The sonority feature is further used in dynamic PF, spectral slope modification and improved voicing decision, which are discussed in detail. The contributions of each of these modules carry different attributes related to the quality of synthesized speech. These contributions are studied independently and each of them achieves stand-alone improvement. In this chapter, the combined performance of these different modifications is analyzed and their comparative significance is presented.*

## 6.1 Introduction

In the previous chapters, the focus is made to improve synthesized speech quality by incorporating modifications to different elements of the SPSS. It can be expected that bringing these modifications under a common framework may lead to further improvement in terms of naturalness. This chapter discusses the combined framework, where the first component is improved parametric representation of speech signal in terms of the additional sonority feature described in Chapter 3. The extracted sonority feature is incorporated in the SPSS framework and employed in dynamic PF algorithm discussed in Chapter 4. This sonorant class based dynamic PF reduces the over-smoothing in source and spectral parameters. During excitation source generation using the post-filtered source parameters, the improved voicing decision is incorporated as proposed in Chapter 5. This assists in generating correct voiced/unvoiced frames of the excitation source. The generated excitation source along with post-filtered spectral parameters are passed through the vocoder to render the synthesized speech waveform. In dynamic PF method, the focus is only to modify the spectral peak prominence, $F_0$ and reduce over-smoothing of the generated source and spectral parameters. However, there is some amount of discrepancy in the spectral slope of synthesized speech, if we compare with that of the natural speech. Therefore, the synthesized speech obtained is further passed through the spectral slope modification algorithm. The output synthesized speech signal is expected to have improved naturalness due to the combined effect of each component.

The rest of this chapter is organized as follows. Section 6.2 elaborates the combination of individual modules and the proposed architecture. The experiments performed to find efficacy of the proposed framework and observations are discussed in Section 6.3. Section 6.4 summarizes the work.

## 6.2 Proposed framework

The proposed combined framework essentially exploits the independent modules discussed in individual chapters simultaneously, that are incorporation of additional feature, dynamic PF, improved voicing decision and spectral tilt modification. The proposed framework is shown in Figure 6.2.

### 6.2.1 Feature extraction

For developing the HMM based SPSS system, the HTS toolkit version 2.3 [134] is used, that involves training and testing processes. In training stage, the first step towards developing SPSS is representing the speech database in parametric form. For this, the conventional features used are $F_0$, its delta and delta-delta (3-dimensional), MGC coefficients, its delta and delta-delta (105-dimensional). Along with these features the sonority feature, its delta and delta-delta (21-dimensional) are also extracted from training speech database. The $F_0$ parameter is extracted using ZFF algorithm with continuous values in both voiced and unvoiced frames. The extraction of sonority feature is explained thoroughly in Chapter 3. Delta and delta-delta sonority feature is discussed in Chapter 4. In each case, the dynamic features are first and second order derivatives of speech parameters used to include dependency of features of one frame over its nearby frames. All the features are extracted for a window length of 25 ms with 5 ms window shift. The features corresponding to the label sequence obtained from text processing module are used for training HMMs. All the phonemes are modeled with 5 states, each state having 3 streams for modeling different parameters. The first stream consists of MGC coefficients including the zeroth coefficient and their delta, delta-delta coefficients. Continuous $F_0$ modeling approach is followed as mentioned in [93]. Therefore, $F_0$, its delta and delta-delta are modeled in a single (second) stream using continuous HMMs. In the third stream 21-dimensional sonority feature is modeled. It can be seen from Figure 6.2 that apart from using the sonority feature in training of HMMs, it is also used to develop voiced/unvoiced classifier and sonorant classifier. The classes considered in the sonorant classification are: low-vowels, mid-vowels, high-vowels, glides, liquids and nasals. Each dimension of the sonority feature is normalized to get values in between the range of $0-1$. The normalized sonority feature is employed to develop the SVM classifiers with RBF kernel.

5-fold cross-validation is performed to derive the appropriate values of $c$ and $\gamma$. These classifiers will be employed in the subsequent modules for improving quality of synthesized speech.

### 6.2.2 Dynamic post-filtering

As explained in Chapter 4, the aim of the dynamic PF method is to reduce the over-smoothing of the generated parameters, imposed due to statistical modeling. As shown in Figure 6.2 based on the obtained label sequence from text to be synthesized, log $F_0$, MGC coefficients, and sonority features are generated from trained HMMs using the parameter generation algorithm. The generated log $F_0$ and MGC coefficients are passed through the dynamic PF module. The PF mechanism has also two parts: training and testing.

- **Training:** As shown in Figure 6.2, a background dataset is used for training module of the dynamic PF method. The CMU arctic database contains 1132 utterances, among which 1000 utterances are used to develop an HMM based speech synthesis system that models conventional excitation source, spectral and sonority features, as described above. Using this system, the source (log $F_0$), spectral (MGC coefficients) and sonority parameters along with synthesized speech are derived for 100 test utterances. The natural counterparts (log $F_0$, MGC coefficients, speech signal) are also available for the same. These parameters and speech signal corresponding to natural and synthesized are referred as the background set. From the background set, the mean and standard deviations of values of spectral peaks, valleys and $F_0$ for frames of each sonorant class corresponding to natural and synthesized are accumulated, which is elaborately described in Chapter 4. This process can be referred as training of the dynamic PF method.
- **Testing:** In the testing phase, the rest 32 utterances are used for synthesis. As per the input text, using the maximum likelihood parameter generation algorithm, framewise MGC coefficients, sonority feature and $F_0$ sequences are computed by maximizing output probability of static and dynamic features [8]. For each frame, the generated sonority feature and MGC coefficients are passed through the SVM based sonorant classifier and the class label is obtained. Based on the sonorant class information and corresponding means and standard deviations obtained during training, the source and spectral parameters are subjected to the dynamic PF as discussed in Chapter 4. The dynamic PF results in enhanced $F_0$ contour, improved spectral prominence and temporal variation across frames.

### 6.2.3 Improved voicing decision

As mentioned earlier, the $F_0$ parameter used in this framework is extracted from the ZFF method and of continuous valued. To generate the excitation source using post-filtered $F_0$ sequence, the voicing information is essential. To make the voicing decision independent of $F_0$, it is obtained by applying the generated sonority feature through the SVM based voiced/unvoiced classifier developed in the training phase. Based on the class label for voiced/unvoiced frames and $F_0$ values, the excitation source is generated. This excitation source along with post-filtered spectral parameters are passed through the synthesis module to derive the synthesized speech.

### 6.2.4 Spectral tilt modification

Apart from the prominence of spectral peaks and valleys, spectral slope also plays a significant role in the perception of synthesized speech. To alleviate the gap in the spectral slope of natural and synthesized speech, we apply the spectral slope modification method on the synthesized obtained from the previous module. As discussed in Chapter 4, the spectral slope modification method consists of training and testing phases. As shown in Figure 6.2, using the background set of natural and synthesized speech signal, the error vector corresponding to each sonorant class is obtained during training. In the testing phase, using these error vectors and knowledge of sonorant class, the spectral slope of each frame is modified to make it similar to that of the natural. The spectral slope modified speech signal achieves additional improvement.

## 6.3 Experimental evaluation

In order to analyze the efficacy of the proposed framework, the evaluation is carried out at different levels of modification for both SLT and BDL speakers. 32 utterances from CMU arctic database are used in the assessment that are not considered in the training stage. The proposed methods are implemented in both MLSA and STRAIGHT vocoder based SPSS framework. However, the combined structure depicted in Figure 6.2 corresponds to MLSA vocoder based SPSS. In case of STRAIGHT, apart from other features the aperiodicity parameter is also modeled in another stream of HMM, which is used for STRAIGHT based excitation source generation while the other modules remain same as described above. Initially, the synthesized speech files are obtained without any modification, where MSD-HMM is used for modeling $F_0$ and voicing decision simultaneously (BASELINE). In

the next stage, $F_0$ is extracted using ZFF algorithm, followed by continuous $F_0$ modeling approach along with sonority based voicing decision, keeping all other modules same (VOICING). In the third case, the dynamic PF method is applied to spectral peaks, valleys and $F_0$ sequences as explained in Chapter 4. In this case the $F_0$ corresponding to ZFF and sonority based voicing is used as well (VOICING+POST-FILTERING). Therefore, this shows the performance when the improved voicing decision and PF methods are used at the same time. In the fourth type, the synthesized speech obtained after employing the proposed voicing decision and PF methods, is further passed through the spectral tilt modification module (VOICING+POST-FILTERING+TILT). These three kinds of synthesized speech will show significance of each modification in combination with the others. The spectral tilt modification and dynamic PF of $F_0$, spectral peaks and valleys may effect each other while used in fusion. Therefore, the spectral tilt modification method is also applied to the synthesized speech obtained from only improved voicing decision ( VOICING+TILT). Along with these, PF and spectral tilt modification methods are evaluated independently with all other modules being same as BASELINE (POST-FILTERING, TILT). To summarize, following types of speech files are used in the evaluation, in case of both MLSA and STRAIGHT vocoder for each of SLT and BDL speakers. Both objective and subjective assessments are performed to justify the efficacy of each method.

- BASELINE: MSD-HMM based modeling of voicing decision and $F_0$.
- VOICING: Improved voicing decision and continuous $F_0$ modeling.
- POST-FILTERING: Dynamic PF of $F_0$, spectral peaks and valleys.
- TILT: Spectral tilt modification.
- VOICING+POST-FILTERING: Improved voicing decision along with dynamic PF of $F_0$, spectral peaks and valleys.
- VOICING+TILT: Improved voicing decision along with spectral tilt modification.
- VOICING+POST-FILTERING+TILT: Improved voicing decision along with dynamic PF of $F_0$, spectral peaks, valleys and spectral tilt modification.

### 6.3.1 Subjective evaluation

In the previous chapters, the evaluation is performed for two parameters, naturalness and intelligibility. It is observed that the intelligibility of SPSS synthesized speech is better than that of the naturalness and the proposed approaches are towards improving naturalness. Therefore, in this evaluation, we primarily focus on the naturalness parameter. For assessment, 5 speech files are used from

**Table 6.1:** MOS for individual proposed methods and their fusion.

| Speaker → | SLT | | BDL | |
|---|---|---|---|---|
| Method ↓ | MLSA | STRAIGHT | MLSA | STRAIGHT |
| BASELINE | **2.16** | **3.28** | **2.27** | **2.68** |
| VOICING | 2.37 | 3.45 | 2.38 | 2.94 |
| POST-FILTERING | 2.95 | 3.45 | 2.59 | 3.10 |
| TILT | 2.48 | 3.35 | 2.35 | 2.81 |
| VOICING+POST-FILTERING | 3.30 | 3.59 | 2.71 | 3.50 |
| VOICING+TILT | 2.61 | 3.48 | 2.53 | 3.05 |
| VOICING+POST-FILTERING+TILT | **3.38** | **3.70** | **2.75** | **3.55** |

each of the above 7 methods, along with 5 natural speech files. Therefore, total 40 speech files are obtained for each vocoder and each speaker. As both MLSA and STRAIGHT vocoders are used, 80 files are obtained for each speaker. For both, the speakers 160 speech files are randomly coded to avoid bias towards any method. Using these examples MOS based evaluation is performed with 20 subjects, having the knowledge of speech perception. They were asked to provide a score between 1 to 5 based on naturalness, by listening to the speech files carefully using headphone. Naturalness can be defined as the closeness of synthesized speech to human speech. Apart from this reference speech files were demonstrated to them to indicate the effect of voicing error and over-smoothing. Based on observation of these factors along with the quality of synthesized speech, the subjects were asked to provide score against each speech file. The MOS corresponding to each method is shown in Table 6.1. It can be observed that among each of the individual techniques, the TILT has a lower contribution towards improvement compared to VOICING and POST-FILTERING. Again, VOICING+POST-FILTERING always achieves better performance compared to that of VOICING+TILT. However, after fusion VOICING+POST-FILTERING+TILT has shown improvement over VOICING+POST-FILTERING. When BASELINE and VOICING+POST-FILTERING are compared, significant improvement is observed using the proposed combined framework. The MOS value obtained for BASELINE is 2.16 and after applying proposed framework is 3.38 in case of MLSA vocoder and female speaker. We can observe an absolute improvement 1.22 after incorporating the contributions made in this thesis. The improve in MOS value in case of the male speaker is 0.48. It can be observed that the contribution of the PF method is more for the female speaker and less for the male speaker. The distributions corresponding to the scores obtained for each method are shown in terms of boxplot in Figure 6.1. In case of MLSA vocoder the improvement achieved is more for female speaker and in case of STRAIGHT

vocoder, it is more for the male speaker. To observe the statistical significance of results discussed above a series of pairwise t-test is performed, between each pair of methods in Table 6.1. All the pairs are found to be significantly different at 1% level except VOICING+POSTFILTERING vs. VOICING+POSTFILTERING+TILT for SLT STRAIGHT, BDL MLSA AND BDL STRAIGHT. Another pair of methods with less than 1% level of difference is BASELINE vs. TILT.

Another subjective evaluation carried out is preference test, where the comparison is made between the methods BASELINE vs. VOICING, VOICING vs. VOICING+POST-FILTERING, VOICING+POST-FILTERING vs. VOICING+POST-FILTERING+TILT and BASELINE vs. VOICING+POST-FILTERING+TILT. For each comparison, 5 pairs of speech files are given in random order. Therefore, total 20 pairs of speech files are provided (for each vocoder and each speaker), and the subjects are asked their preference among each pair. If both speech files seem to be similar, they are asked to mark those as no preference. The % of choice along with % of no preference corresponding of each comparison are shown in Table 6.2. The p-values obtained from double-tailed t-test between each pair are found to be less than 0.01. It depicts that in contrast to VOICING+POST-FILTERING (42.22%) vs. VOICING+POST-FILTERING+TILT (51.11%), the preference is almost same towards each (for SLT speaker and MLSA vocoder). This result signifies limited contribution of spectral tilt modification compared to improved voicing and dynamic PF methods. However, in BASELINE vs. VOICING+POST-FILTERING+TILT, the latter one is most preferred. This observation shows the quality improvement in SPSS synthesized speech after employing the proposed framework. However, the proposed combined framework achieves preference of 95.56% compared to 2.22% for the baseline method, in case of female speaker MLSA vocoder. The preference obtained for the male speaker is 77.11% for the proposed framework against 22.22% for the baseline method.

**Table 6.2:** Result of preference test in terms of % of preference.

| Comparison | Vocoder | SLT | | | BDL | | |
|---|---|---|---|---|---|---|---|
| | | A | B | No pref | A | B | No pref |
| BASELINE (A) vs. VOICING (B) | MLSA | 11.11 | 75.56 | 13.33 | 35.56 | 57.78 | 6.67 |
| | STRAIGHT | 15.50 | 63.22 | 21.28 | 26.67 | 57.78 | 15.56 |
| VOICING (A) vs. VOICING+POST-FILTERING (B) | MLSA | 11.11 | 75.56 | 13.33 | 31.11 | 60.00 | 8.89 |
| | STRAIGHT | 7.33 | 78.49 | 14.18 | 35.56 | 46.67 | 17.78 |
| VOICING+POST-FILTERING (A) vs. VOICING+POST-FILTERING+TILT (B) | MLSA | 42.22 | 51.11 | 6.67 | 28.89 | 68.89 | 2.22 |
| | STRAIGHT | 38.33 | 56.32 | 5.35 | 33.33 | 55.56 | 11.11 |
| BASELINE (A) vs. VOICING+POST-FILTERING+TILT (B) | MLSA | **2.22** | **95.56** | **2.22** | **22.22** | **71.11** | **6.67** |
| | STRAIGHT | **7.23** | **84.35** | **8.42** | **28.89** | **55.56** | **15.56** |

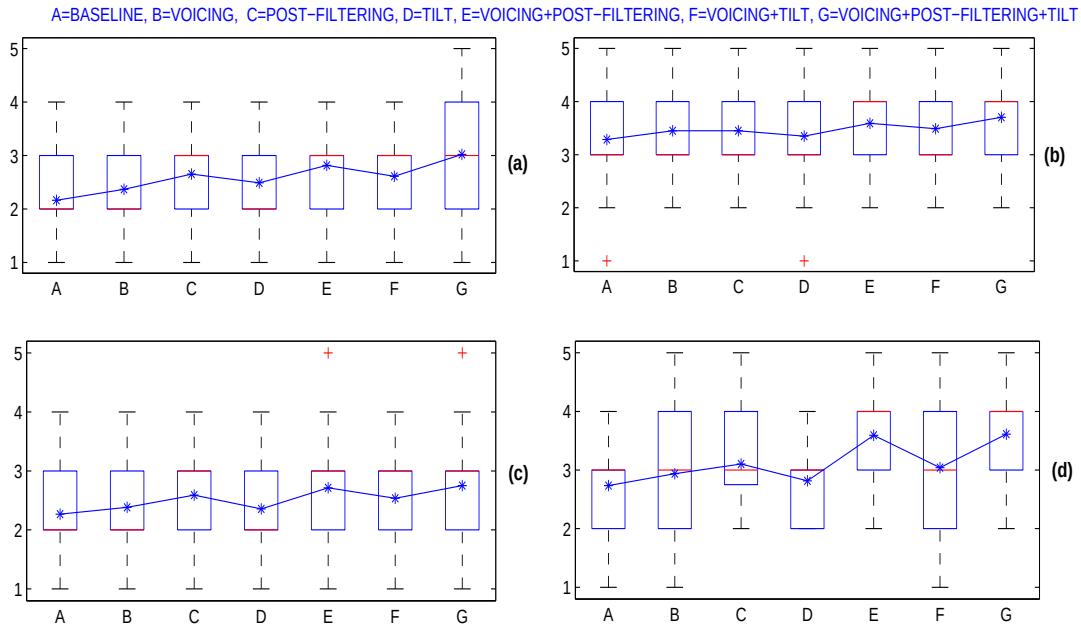


**Figure 6.1:** Boxplot representing distribution of mean opinion scores of different methods for (a) SLT MLSA, (b) SLT STRAIGHT, (c) BDL MLSA, (d) BDL STRAIGHT.

### 6.3.2 Objective evaluation

As the number of subjects and speech files are limited in case of the subjective evaluation process, therefore objective measures are also used to validate the improvement achieved after incorporating the proposed algorithms. The objective measures used are PESQ-MOS and LSD. Both the measures are calculated for each type of synthesized speech files corresponding 32 utterances for both the speakers and vocoders. The reference files used are the corresponding natural speech files. As the length of natural speech files and synthesized are not same, the frames are aligned using DTW before calculating the LSD. Both the objective measures calculated corresponding to each method are noted in Table 6.3. After incorporating the proposed method the MOS values are found to be increased, while the LSD is decreasing. This refers to increase in naturalness and decrease in distance with the natural speech, after incorporating the proposed methods. The MOS value is found to be highest for VOICING+POST- FILTERING+TILT method (1.41), while the same for BASELINE is 1.16 corresponding to SLT STRAIGHT. The observations from both objective and subjective evaluation show a significant improvement in synthesis quality in case of female speaker.

**Table 6.3:** Objective evaluation results for proposed methods and their fusion.

| **Speaker →** | **SLT** | | | | **BDL** | | | |
|---|---|---|---|---|---|---|---|---|
| **Method ↓** | **MLSA** | | **STRAIGHT** | | **MLSA** | | **STRAIGHT** | |
| | MOS | LSD | MOS | LSD | MOS | LSD | MOS | LSD |
| BASELINE | **1.02** | **2.52** | **1.16** | **2.19** | **1.08** | **2.30** | **1.13** | **1.85** |
| VOICING | 1.41 | 2.50 | 1.19 | 2.14 | 1.10 | 2.25 | 1.19 | 1.83 |
| POST-FILTERING | 1.13 | 2.46 | 1.22 | 2.10 | 1.14 | 2.23 | 1.16 | 1.83 |
| TILT | 1.15 | 2.45 | 1.20 | 2.12 | 1.10 | 2.27 | 1.17 | 1.82 |
| VOICING+POST-FILTERING | 1.18 | 2.39 | 1.26 | 2.08 | 1.17 | 2.19 | 1.21 | 1.81 |
| VOICING+TILT | 1.14 | 2.43 | 1.18 | 2.10 | 1.18 | 2.21 | 1.22 | 1.82 |
| VOICING+POST-FILTERING+TILT | **1.21** | **2.35** | **1.41** | **2.04** | **1.20** | **2.17** | **1.24** | **1.78** |

## 6.4 Summary

In this chapter, the proposed modules in the previous chapters are fused together to introduce a combined scheme for improvement in the quality of synthesized speech. From the subjective and objective evaluations, it is observed that the PF algorithm has the highest impact on quality of synthesized speech. As two different perceptual aspects of synthesized speech i.e. spectral prominence and excitation source components are modified, it gives significant enhancement in synthesized speech quality. Improved voicing decision and PF mechanism in fusion give advantage from two different prospectives. The spectral-tilt modification method provides added benefit, although it is less while in fusion with the dynamic PF method. However, in combination with the improved voicing algorithm the spectral tilt modification method shows significant contribution. Finally, when all the three different perspectives are combined together, a commendable performance is achieved.

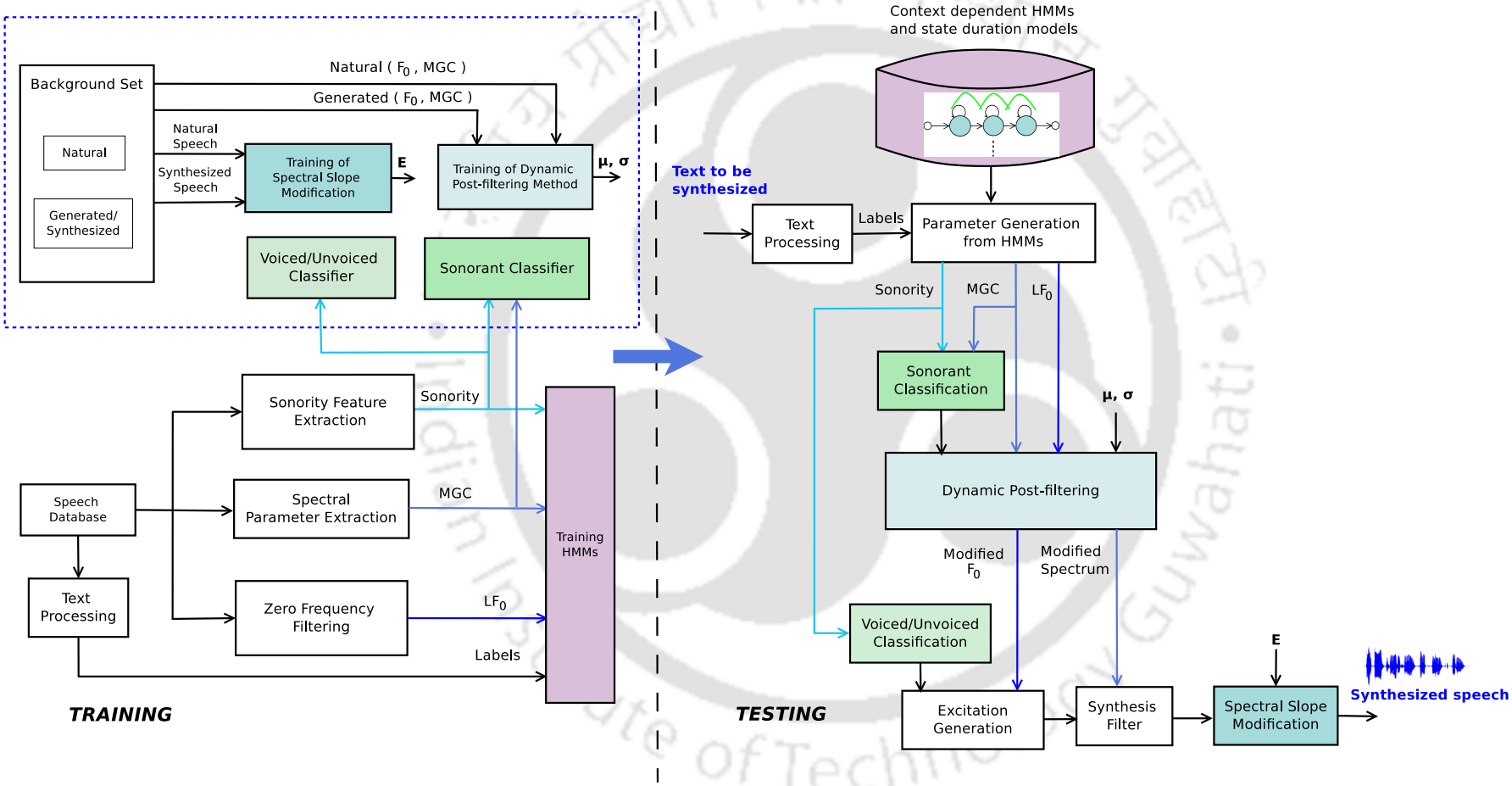


**Figure 6.2:** Block diagram representing the combined framework.

# 7

# Summary and Conclusions

Contents

*In this chapter, we have summarized and concluded the work presented in this thesis towards improving the quality of statistical parametric speech synthesis using sonority information.*

## 7.1 Summary

The objective of this thesis is to improve the quality of synthesized speech obtained from SPSS approach. The drawbacks of SPSS attempted to solve in this thesis include, poor representation of speech signal in terms of acoustic parameters, over-smoothing of generated parameter sequence and inaccurate voicing decision. In order to overcome these issues, the sonority information is incorporated in the SPSS framework. The sonority information is further used in reducing the over-smoothing and improving voicing decision. The contributions incorporated in this thesis are summarized below.

(i) **Sonority information:** In the second chapter of this thesis, a detailed review is made on the issues associated with the SPSS framework that limit the synthesized speech quality. The first issue discussed is the use of limited number of parameters as representation of the speech signal. The next issues are over-smoothing of generated parameter contours and error in voicing decision during excitation source generation. The existing efforts towards alleviating these issues along with their pros and cons are discussed in the second chapter. Apart from the conventional source and spectral features, incorporation of additional information and its subsequent utilization in improving other modules may contribute towards enhancement of synthesized speech quality. As the sonority information has high correlation with the speech perception, therefore a study is made to extract the feature representing sonority associated with a sound unit. The sonorant sounds can be characterized by higher spectral prominence, excitation source strength and longer duration. To capture the vocal-tract information the usefulness of HNGD spectrum is exploited. Different measures related to the spectral peaks of HNGD spectrum are used to represent the formant prominence. The HE of LP residual has the ability to represent the excitation source. The peak to side-lobe ratio of each peak (corresponding to GCI) derived from HE of LP residual represents the excitation source aspect of sonority. The periodicity aspect is represented by the correlation measure of the speech signal over several pitch periods. These three aspects derived from the speech signal is defined as sonority feature. The extracted sonority feature is shown to be useful in different speech processing applications. The

usefulness of the sonority feature in sonorant/non-sonorant segmentation and six class sonorant classification tasks verify its efficacy in representing the degree of sonority. The proposed feature also improves the performance of DNN based phoneme recognition and VOP detection. As the degree of sonority is associated with the speech perception, the sonority feature is used in different modules of SPSS to improve synthesized speech quality.

(ii) **Dynamic post-filtering:** As mentioned earlier, another major shortcoming of SPSS is over-smoothing of the generated parameter contours. In the literature, PF methods are found to be useful to alleviate this. The existing methods use constant PF factor for all the categories of sound units. However, based on the difference in production behavior, different categories of sound units may require enhancement to different extent. The sounds with higher sonority have higher spectral sharpness and SoE that effect the perception. Based on this fact, different PF factors are derived for different categories of sonorant sounds. Along with conventional $F_0$ and MFCCs, the sonority feature is also modeled in a separate stream in the HTS. The source parameters considered for PF are $F_0$, SoE, and the spectral parameters are first five spectral peaks and valleys. For each of these parameters, a comparison is carried out between natural and synthesized counterpart, using a background dataset of parallel natural and synthesized utterances. This comparison shows that the deviation between natural and synthesized parameters may vary with different sonorant categories. Therefore, means and variances of the post-filters are obtained for the above mentioned parameters corresponding the sonorant categories. An SVM classifier is trained using the sonority feature to classify the speech sound to different sonorant classes. This can be referred as the training phase.

During synthesis, based on the text to be synthesized, along with the conventional features the sonority feature is also generated framewise. The sonorant class information of each frame is obtained from the SVM classifier using the generated sonority feature. The corresponding post-filter mean and variance is used to derive the modified source and spectral parameter values. This dynamic modification of the source and spectral parameters helps to reduce the over-smoothing to improve the quality of synthesized speech. Introducing dynamic PF factor in the spectral domain helps to preserve the fine spectral structure, whereas the dynamic PF over successive frames reduces the over-smoothing. This post-filter tries to reduce the deviation between natural and synthesized speech in terms of $F_0$, SoE, spectral peaks and valleys.

(iii) **Spectral tilt modification:** The PF of spectral features such as peaks, valleys and source features $F_0$, SoE have significant impact on the perceptual quality of synthesized speech. As stated in the previous studies, the spectral tilt also has an effect on the perception. Suppression of higher harmonics results in more negative slope that makes sounds muffled in quality. Enhancement of higher harmonics to some extent results in clear sounds. An analysis is carried out to compare the spectral tilt corresponding to voiced frames between natural and synthesized speech. The synthesized speech is found to have more negative spectral tilt compared to that of the natural. This behavior seems to be consistent over different speakers. The same is analyzed for different sonorant categories (low-vowels, mid-vowels, high-vowels, semivowels, nasals). It is observed that the divergence is more in case of low-vowels, mid-vowels, high-vowels compared to that of the nasals. Therefore, a spectral tilt modification method is proposed that flattens the spectral tilt to required extent depending on the class of sound unit. For a parallel set of natural and synthesized speech frames, the first order LP spectrum is determined. Corresponding to each class an error vector is obtained by subtracting the average of first order LP spectrum of all the frames of the synthesized from that of the natural. This is the training phase. During testing, given the frame of synthesized speech for a particular class the tilt of the spectrum is corrected using the knowledge of corresponding error vector. This method of spectral slope modification corrects the spectral tilt, without effecting the other parameters of the spectrum. This method is found to provide improvement in terms of naturalness, intelligibility and speaker similarity.

(iv) **Improved voicing decision:** The naturalness of synthesized speech greatly depends on the excitation source, which is generated based on the voicing decision corresponding to each frame. In conventional approach, voicing decision and $F_0$ are modeled using MSD-HMM, which may carry errors due to $F_0$ estimation and modeling. As the sonortiy notion is highly correlated with voicing, therefore the sonority feature is analyzed to find its efficacy in representing voicing information. It is found that the feature has better capability to differentiate between voiced and unvoiced classes compared to conventional features. An SVM classifier developed to decide voiced/unvoiced frames for the generation of excitation source. In this case, the $F_0$ information is extracted from the ZFF algorithm with continuous values in both voiced and unvoiced frames. Therefore, continuous $F_0$ modeling is used instead of MSD-HMM. During synthesis, the voicing

decision corresponding to each frame is obtained by applying the generated sonority feature to the SVM based voiced/unvoiced classifier. The derived voicing decision and $F_0$ are used to generate the excitation source which is applied to the vocoder to render the synthesized speech. This improved voicing decision is found to have compelling effect on the naturalness.

(v) **Combined framework:** After exploring the sonority feature and its prospective applications in different modules of SPSS, the work focuses on combining all these modules into a common framework. In the training phase, the sonority feature is also extracted along with log $F_0$ and MGCs, from the training speech corpora. These are modeled simultaneously in the HTS framework. At the same time, the dynamic post-filters for the source and spectral parameters are trained with respect to different sonorant classes using the background parallel data. The error vectors corresponding to spectral tilt modification method are also obtained for different classes. During the same training phase, two SVM based classifiers for sonorant and voiced/univoiced classifications are trained using the sonority feature.

During testing, the sonority, log $F_0$ and MGCs are generated from the HMMs. The generated sonority feature is applied to the SVM based sonorant classifier to derive the class information. Accordingly, the dynamic PF is performed to derive modified source and spectral parameters. Using the modified source parameters and voicing information derived from the sonority based voiced/unvoiced classifier, excitation source is generated. The synthesized waveform is obtained from the vocoder using the excitation source and spectrum. Further, the spectral tilt modification method is applied to enhance the synthesized speech quality, by correcting the spectral slope of the spectrum.

## 7.2 Conclusions

The conclusions referred from the work proposed in this thesis are as follows:

- Based on the production characteristics of the sonorant sounds, a feature level of representation of degree of sonority is extracted from the speech signal. Less vocal-tract constriction, higher excitation source strength and periodicity are the three mostly cited correlates of sonority. A feature set derived from the vocal-tract system, excitation source and suprasegmental aspects is proposed to quantify the sonority associated with a sound unit. The proposed feature has certain level of capability to correctly delineate the sonority hierarchy. It also shows efficacy

towards different applications like phoneme recognition and VOP detection.

- A dynamic post-filter that applies different modification factors for frames corresponding to various sonorant classes, apparently enhances the temporal variation of the parameter contours. Use of separate PF factors for different spectral peaks and valleys helps to retain the fine spectral structure. This method enhances the synthesized speech quality to a great extent. It is found that the effect of dynamic PF on the synthesized speech is more compared to improvements achieved in other modules of this work.
- To minimize the difference in spectral tilt between natural and synthesized speech, a spectral tilt modification method is proposed that leads to better quality of synthesized speech. As the synthesized speech has more negative spectral slope compare to that of the natural, the modification of spectral tilt leads to a flatter spectrum resulting in clearer sound quality.
- Apart from these parameter modification or enhancement methods, voiced/unvoiced decision is another component that effects the synthesized speech quality. In addition to the excitation source based features, the VTS information present in the sonority feature leads to improved voicing decision. This has a significant impact to naturalness of the synthesized speech.
- A combined framework is exploited that effectively combines each of the above modules and the performance is analyzed. It shows that the dynamic PF method achieves highest preference among each of the individual modules. The combination of dynamic PF and improved voicing decision achieves a countable gain in the quality of synthesized speech. Further a small improvement in observed, if we apply the spectral tilt modification over the synthesized speech.

## 7.3 Contributions

The contributions made as a part of this thesis are as follows:

- Proposing an efficient feature set from vocal-tract system, excitation source and suprasegmental aspects of the speech signal having capability to represent degree of sonority.
- As sonority is associated with production aspects of sound units, its subsequent applications in improved phoneme recognition and VOP detection are presented.
- A sonorant class based dynamic PF method is proposed to reduce the over-smoothing as well as the deviation between natural and synthesized speech parameters.
- To minimize the difference in spectral tilt of natural and synthesized speech, a spectral tilt

modification method is proposed that leads to better quality of synthesized speech.

- Use of sonority feature in voicing decision during excitation source generation in the SPSS framework results in improved naturalness of the synthesized speech.
- A combined framework is proposed that models the sonority feature and further employs it in PF of spectral peaks and valleys, $F_0$, SoE; voicing decision and spectral tilt modification at the same time.

## 7.4 Criticism

The efforts made in this thesis towards improving quality of synthesized speech achieve the goal to some extent. The methods proposed in this work essentially focus on signal processing based approaches instead of contribution to statistical modeling techniques. It firstly derives a set of features representing degree of sonority that can be used in diverse speech processing algorithms. The sonority feature is extracted from the knowledge formant prominence, excitation source strength and periodicity. However, apart from these there are several other correlates of sonortiy broadly studied in phonetics which are not included in the presented work. This results in limited improvement using the sonority feature in addition to MFCC in case of the sonorant classification. Further, the sonority feature is modeled using HMMs and adopted in the dynamic PF method. In this module, the sonority feature is only used in the classification task to derive information about the sonorant class corresponding to each frame. Based on the class information corresponding PF factors are applied to parameters of that frame. This is indirect utilization of sonority feature in PF. The direct modification of different aspects of sonority would have been more useful. In the dynamic PF method, first five spectral peaks and valleys are modified that may also effect the slope of the spectrum. The consequence of PF on the spectral slope is not studied in this work, that may be the reason of limited improvement when dynamic PF and spectral slope modification methods are incorporated at the same time. The next exploration is the improved voicing decision using the sonority information. The major drawback in this case is we require the baseline voicing boundaries to train the SVM classifier, that may not be available for the database used in development of TTS. Based on these drawbacks of the work presented in this thesis, some prospective future work directions are also explicated.

## 7.5 Directions for future work

- The distributions of sonority feature corresponding to different classes of sonorant sounds are found to be overlapping to significant extent. This may be due to the effect of change in the feature value with different speaker. As the resonance peaks may be changed with the variation in vocal-tract length, this effect can be normalized using vocal-tract length normalization techniques. This may lead to speaker independent sonority feature with improved accuracy.
- The sonortiy feature is extracted from the speech signal based on the formant prominence, SoE and periodicity. On the other hand, there are several other correlates of sonority studied in the literature of phonology. Such correlates include duration, continuability, harmonic phases, tonality, pitch and so on. These aspects of sonority can be also studied and a more potent feature of sonority can be developed.
- In case of the dynamic PF method, first 5 spectral peaks are enhanced with different factors. These factors are derived based on the comparison between natural and synthesized counterpart during training. The modification of 5 peaks mostly effects the lower frequency region of the spectrum. To modify the entire spectrum, it can divided into number of bands. The spectral peaks corresponding to different bands can be modified with different factors.
- As the use of SVM classifier in the speech synthesis framework may not be a suitable approach, we aim to add decision tree-based phoneme-level clustering method using sonority class questions instead of the SVM in the future.
- The enhancement of spectral peaks may also effect the spectral tilt. The trade off between PF factors for modification of spectral peaks and change in spectral tilt can be analyzed in detail. A combined method by modifying both spectral tilt and peak prominence to correct the spectrum can be proposed.
- As the idea of the class specific PF method provides better modification of the generated parameter sequences, the state-of-the art methods of PF can also be designed in the same way. A comparison of class independent and class dependent PF in case of each method will help to understand significance of class information in a better way.
- The voicing decision algorithms that require baseline voicing boundaries can be developed using the empirical methods. The issue in this case is setting of the threshold based on the database used. Development of automatic threshold setting method based on different signal character-

istics could be helpful in this scenario.

- Although combining all the modules together is done well, the interactions among the various components will need to be considered more carefully to determine if a they work against each other Interactions among the various components.

# Bibliography


[1] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 805–808.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system." *Journal of the Acoustical Society of Japan (E)*, vol. 21, no. 4, pp. 199–206, 2000.

[3] S.-J. Kim, J.-J. Kim, and M. Hahn, "HMM-based Korean speech synthesis system for hand-held devices," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 4, pp. 1384–1390, 2006.

[4] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, "Quantized HMMs for low footprint text-to-speech synthesis," in *Interspeech*, 2010, pp. 837–840.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[6] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1315–1318.

[7] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1999, pp. 229–232.

[8] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[9] S. G. Parker, "Quantifying the sonority hierarchy," Ph.D. dissertation, University of Massachusetts Amherst [Published by the GLSA.], 2002.

[10] K. Schutte and J. R. Glass, "Robust detection of sonorant landmarks." in *Interspeech*, 2005, pp. 1005–1008.

[11] S. Parker, "Sound level protrusions as physical correlates of sonority," *Journal of phonetics*, vol. 36, no. 1, pp. 55–90, 2008.

[12] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[13] J. Blevins and J. Goldsmith, "The syllable in phonological theory," *Phonology: Critical Concepts: Syllables and Multi-level Analyses*, vol. 3, pp. 75–120, 2001.

[14] M. Gouskova, "Relational hierarchies in optimality theory: the case of syllable contact," *Phonology*, vol. 21, no. 02, pp. 201–250, 2004.

[15] P. De Lacy, "Markedness conflation in optimality theory," *Phonology*, vol. 21, no. 02, pp. 145–199, 2004.

[16] E. Moreton, G. Feng, and J. L. Smith, "Syllabification, sonority, and perception: new evidence from a language game," in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, vol. 41, no. 1. Chic Ling Society, 2005, pp. 341–355.

[17] S. Topbas and H. KopkallI-Yavuz, "Reviewing sonority for word-final sonorant+ obstruent consonant cluster development in Turkish," *Clinical linguistics & phonetics*, vol. 22, no. 10-11, pp. 871–880, 2008.

[18] M. Miozzo and A. Buchwald, "On the nature of sonority in spoken word production: Evidence from neuropsychology," *Cognition*, vol. 128, no. 3, pp. 287–301, 2013.

[19] I. Deschamps, S. R. Baum, and V. L. Gracco, "Phonological processing in speech perception: What do sonority differences tell us?" *Brain and language*, vol. 149, pp. 77–83, 2015.

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

[21] D. H. Klatt, "Review of text-to-speech conversion for english," *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

[22] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[23] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[24] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5, pp. 453–467, 1990.

[25] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1996, pp. 373–376.

[26] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis." in *Proc. Eurospeech*, 1997, pp. 601–604.

[27] N. Mizutani, K. Tokuda, and T. Kitamura, "Concatenative speech synthesis based on HMM," in *Proc. Autumn Meeting of ASJ*, 2002, pp. 241–242.

[28] C. Allauzen, M. Mohri, and M. Riley, "Statistical modeling for unit selection in speech synthesis," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004, p. 55.

[29] S. Sakai and H. Shu, "A probabilistic approach to unit selection for corpus-based speech synthesis," in *INTERSPEECH*, 2005, pp. 81–84.

[30] Z.-H. Ling and R.-H. Wang, "HMM-based unit selection using frame sized speech segments," in *ICSLP*, 2006.

[31] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[32] R. E. Donovan and P. C. Woodland, "Improvements in an HMM-based speech synthesiser," in *Eurospeech Proceedings: 4th European Conference on Speech Communication and Technology*, vol. 1, 1995, pp. 573–576.

[33] X. Huang, A. Acero, J. Adcock, H.-W. Hon, J. Goldsmith, J. Liu, and M. Plumpe, "Whistler: A trainable text-to-speech system," in *ICSLP*, vol. 4, 1996, pp. 2387–2390.

[34] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T next-gen TTS system," in *Joint meeting of ASA, EAA, and DAGA*, 1999, pp. 18–24.

[35] P. Taylor and A. W. Black, "Speech parameter generation from HMM using dynamic features," in *EUROSPEECH*, 1999, pp. 1531–1534.

[36] H. Segi, T. Takagi, and T. Ito, "A concatenative speech synthesis method using context dependent phoneme sequences with variable length as search units," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[37] S. P. Kishore and A. W. Black, "Unit size in unit selection speech synthesis." in *Interspeech*, 2003, pp. 1317–1320.

[38] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to expressive speech synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[39] A. W. Black, "Unit selection and emotional speech." in *EUROSPEECH*, 2003, pp. 1649–1652.

[40] T. Yoshimura, K. Tokuda, T. Kobayashi, T. Masuko, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *EUROSPEECH*, 1999, pp. 2347–2350.

[41] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 660–663.

[42] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[43] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information and Systems*, vol. 85, no. 3, pp. 455–464, 2002.

[44] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis." in *ICSLP*, vol. 98, 1998, pp. 29–32.

[45] Z. Heiga, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 825–834, 2007.

[46] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[47] A. Black, P. Taylor, R. Caley, and R. Clark, "The festival speech synthesis system," 1998.

[48] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[49] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[50] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE transactions on Speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

[51] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3. IEEE, 1997, pp. 1611–1614.

[52] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE transactions on information and systems*, vol. 88, no. 11, pp. 2484–2491, 2005.

[53] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[54] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis." in *Interspeech*, 2001, pp. 2263–2266.

[55] O. Abdel-Hamid, S. M. Abdou, and M. Rashwan, "Improving Arabic HMM based speech synthesis quality." in *Interspeech*, 2006, pp. 1332–1335.

[56] D. O'Brien and A. I. Monaghan, "Concatenative synthesis based on a harmonic model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 11–20, 2001.

[57] C. Hemptinne, "Integration of the harmonic plus noise model into the hidden Markov model-based speech synthesis system," *Master thesis*, 2006.

[58] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *ISCA SSW6*, 2007, pp. 113–118.

[59] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering." in *Interspeech*, 2008, pp. 1881–1884.

[60] T. Masuko, K. Tokuda, and T. Kobayashi, "A study on conditional parameter generation from HMM based on maximum likelihood criterion," in *Autumn Meeting of ASJ*, 2003, pp. 209–210.

[61] Y. Qian, H. Liang, and F. K. Soong, "Generating natural F0 trajectory with additive trees," in *Interspeech*, 2008, pp. 2126–2129.

[62] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis." in *Interspeech*, 2008, pp. 2274–2277.

[63] Y. Qian, Z. Wu, and F. K. Soong, "Improved prosody generation by maximizing joint likelihood of state and longer units," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 3781–3784.

[64] C.-C. Wang, Z.-H. Ling, B.-F. Zhang, and L.-R. Dai, "Multi-layer F0 modeling for HMM-based speech synthesis," in *ISCSLP*, 2008, pp. 1–4.

[65] W. Yi-jian and W. Ren-hua, "HMM-based trainable speech synthesis for Chinese," *Journal of Chinese Information Processing*, vol. 4, p. 010, 2006.

[66] B. Gao, Y. Qian, Z. Wu, and F. K. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood," in *Interspeech*, 2008, pp. 2266–2269.

[67] S. Tiomkin and D. Malah, "Statistical text-to-speech synthesis with improved dynamics," in *Interspeech*, 2008, pp. 1841–1844.

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

[69] J. Latorre, K. Iwano, and S. Furui, "Combining Gaussian mixture model with global variance term to improve the quality of an HMM-based polyglot speech synthesizer," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 2007, pp. IV–1241.

[70] 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.

[71] Y.-J. Wu, H. Zen, Y. Nankaku, and K. Tokuda, "Minimum generation error criterion considering global/local variance for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 4621–4624.

[72] T. Toda and S. Young, "Trajectory training considering global variance for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 4025–4028.

[73] Z.-H. Ling, Y. Hu, and L. Dai, "Global variance modeling on the log power spectrum of LSPs for HMM-based speech synthesis," in *Interspeech*, 2010, pp. 825–828.

[74] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis," *Systems and Computers in Japan*, vol. 36, no. 12, pp. 43–50, 2005.

[75] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "CELP coding based on mel-cepstral analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 33–36.

[76] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[77] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for HMM-based speech synthesis." in *SSW*, 2010, pp. 334–339.

[78] Y. Ohtani, M. Tamura, M. Morita, T. Kagoshima, and M. Akamine, "Histogram-based spectral equalization for HMM-based speech synthesis using mel-lsp," in *Interspeech*, 2012, pp. 1155–1158.

[79] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 290–294.

[80] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, and Z.-H. Ling, "DNN-based stochastic postfilter for HMM-based speech synthesis." in *Interspeech*, 2014, pp. 1954–1958.

[81] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, Z.-H. Ling, and J. Yamagishi, "A deep generative architecture for postfiltering in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 11, pp. 2003–2014, 2015.

[82] D. Erro, "Two-band radial postfiltering in cepstral domain with application to speech synthesis," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 202–206, 2016.

[83] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for stft spectrograms," in *Interspeech*, 2017, pp. 3389–3393.

[84] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[85] E. Jokinen, U. Remes, M. Takanen, K. Palomaki, M. Kurimo, and P. Alku, "Spectral tilt modelling with extrapolated GMMs for intelligibility enhancement of narrowband telephone speech," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 164–168.

[86] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Interspeech*, 2011, pp. 1837–1840.

[87] D. Talkin, "A robust algorithm for pitch tracking," *Speech Coding and Synthesis, Elseview Science B.V.*, 1995.

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

[89] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201–212, 1976.

[90] D. Arifianto, "Dual parameters for voiced-unvoiced speech signal determination," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),*, vol. 4, 2007, pp. IV–749.

[91] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "A multifeature voiced/unvoiced decision algorithm for noisy speech," in *IEEE International Symposium on Circuits and Systems*. IEEE, 2006, pp. 4–pp.

[92] A. P. Lobo and P. C. Loizou, "Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2003, pp. 820–823.

[93] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[94] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 3773–3776.

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

[96] U. Ogbureke, J. Cabral, and J. Berndsen, "Using multilayer perceptron for voicing strength estimation in HMM-based speech synthesis," in *11th International Conference on Information Science, Signal Processing and their Applications (ICASSP)*, 2012, pp. 683–688.

[97] X. Yin, M. Lei, Y. Qian, F. K. Soong, L. He, Z.-H. Ling, and L.-R. Dai, "Modeling F0 trajectories in hierarchically structured deep neural networks," *Speech Communication*, vol. 76, pp. 82–92, 2016.

[98] N. Narendra and K. S. Rao, "Robust voicing detection and F0 estimation for HMM-based speech synthesis," *Circuits, Systems, and Signal Processing*, vol. 34, no. 8, pp. 2597–2619, 2015.

[99] H. Kawahara, H. Katayose, A. d. Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[100] K. Sjölander and J. Beskow, "Wavesurfer-an open source speech tool." in *Interspeech*, 2000, pp. 464–467.

[101] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics." in *Interspeech*, 2011, pp. 1973–1976.

[102] M. E. Beckman, J. Edwards, and J. Fletcher, "Prosodic structure and tempo in a sonority model of articulatory dynamics," *Papers in laboratory phonology II*, pp. 68–86, 1992.

[103] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Communication*, vol. 1, no. 3, pp. 167–184, 1982.

[104] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering,*, vol. 41, no. 7, pp. 663–671, 1994.

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

[106] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[107] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

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

[109] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2061–2071, 2009.

[110] S. Puppel, "The sonority hierarchy in a source-filter dependency framework," in *Phonological investigations (Linguistic & Literary Studies in Eastern Europe, volume 38.)*, J. Fisiak and S. Puppel, Eds. Amsterdam and Philadelphia: John Benjamins Publishing Company, 1992, pp. 467–483.

[111] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *ICSLP*, vol. 2. IEEE, 1996, pp. 1261–1264.

[112] J. Yuan and M. Liberman, "Robust speaking rate estimation using broad phonetic class recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4222–4225.

[113] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[114] A. R. Arrabothu, N. Chennupati, and B. Yegnanarayana, "Syllable nuclei detection using perceptually significant features." in *Interspeech*, 2013, pp. 963–967.

[115] S. H. Dumpala, B. T. Nellore, R. R. Nevali, S. V. Gangashetty, and B. Yegnanarayana, "Robust features for sonorant segmentation in continuous speech," in *Interspeech*, 2015.

[116] A. Jansen and P. Niyogi, "Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1739–1758, 2008.

[117] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1638–1640, 1978.

[118] G. A. Seber, *Multivariate observations*. John Wiley & Sons, 2009, vol. 252.

[119] J. R. Schott, "Principles of multivariate analysis: A user's perspective," *Journal of the American Statistical Association*, 2011.

[120] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

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

[122] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[123] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[124] Kaldi Toolkit: http://kaldi.sourceforge.net.

[125] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. ASRU*, Dec 2011.

[126] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[127] S. R. M. Prasanna, B. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 556–565, 2009.

[128] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Seventh International Conference on Spoken Language Processing*, 2002.

[129] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for hmm-based expressive speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 9, pp. 1406–1413, 2007.

[130] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[131] V. C. Tartter, H. Gomes, and E. Litwin, "Some acoustic effects of listening to noise on speech production," *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 2437–2440, 1993.

[132] C. Binns and J. F. Culling, "The role of fundamental frequency contours in the perception of speech against interfering speech," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1765–1776, 2007.

[133] N. Adiga and S. R. M. Prasanna, "Significance of instants of significant excitation for source modeling," in *Interspeech*. ISCA, 2013, pp. 1677–1681.

[134] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0." in *SSW*. Citeseer, 2007, pp. 294–299.

[135] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.

[136] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: A CPU and GPU math compiler in python," in *9th Python in Science Conf*, 2010, pp. 1–7.

[137] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation." in *ICSLP*, vol. 94, 1994, pp. 18–22.

[138] D. G. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *the Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[139] J. M. Alexander and K. R. Kluender, "Spectral tilt change in stop consonant perception," *The Journal of the Acoustical Society of America*, vol. 123, no. 1, pp. 386–396, 2008.

[140] K. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters,*, vol. 16, no. 6, pp. 469–472, 2009.

[141] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[142] V. Grancharov and W. B. Kleijn, "Speech quality assessment," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 83–100.

[143] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[144] F. Plante, G. Meyer, and W. Ainsworth, "A pitch extraction reference database," *Children*, vol. 8, no. 12, pp. 30–50, 1995.

[145] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching." in *EUROSPEECH*, 1993, pp. 1003–1006.

# List of Publications

### Journal Publications

- Published and communicated:
  1. **Bidisha Sharma** and S. R. M. Prasanna, "Enhancement of Spectral Tilt in Synthesized Speech," in IEEE Signal Processing Letters, vol. 24, no. 4, pp. 382-386, April 2017.
  2. **Bidisha Sharma** and S. R. M. Prasanna, "Sonority Measurement Using System, Source, and Suprasegmental Information," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 3, pp. 505-518, March 2017.
  3. **Bidisha Sharma**, N. Adiga and S. R. M. Prasanna, "Dynamic Post-filtering using Source and Spectral Features for Statistical Parametric Speech Synthesis," submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing (Under Revision).
  4. **Bidisha Sharma** and S. R. M. Prasanna, "Significance of Sonority Information for Voiced/Unvoiced Decision in Speech Synthesis," submitted to Speech Communication (Under Revision).
  5. **Bidisha Sharma** and S. R. M. Prasanna, "Features, Post-filtering and Voicing Decision for Improving Statistical Parametric Speech Synthesis: A Review," submitted to IETE Technical Review (Under Revision).

### Conference Publications

1. **Bidisha Sharma**, "Scope of Sonority Information in Statistical Parametric Speech Synthesis," 3rd Doctoral Consortium, Interspeech 2017.
2. **Bidisha Sharma** and S. R. M. Prasanna. "Vowel Onset Point Detection using Sonority Information," Interspeech, 2017.
3. **Bidisha Sharma** and S. R. M. Prasanna. "Speech Synthesis in Noisy Environment by Enhancing Strength of Excitation and Formant Prominence," in Proc. Interspeech, 2016.

## Other related publications during thesis work

### Journal Publications

1. Rohan Kumar Das, **Bidisha Sharma** and S. R. M. Prasanna, "Significance of Duration Modification for Speaker Verification Under Mismatch Speech Tempo Condition",IJST, 2017.
2. **Bidisha Sharma** and S.R.M. Prasanna, "Polyglot Speech Synthesis: A Review," IETE Technical Review, vol. 34, no. 4, pp. 366-389, 2017.

### Conference Publications

1. Loitongbam Gyanendro Singh , Nagaraj Adiga, **Bidisha Sharma**, Sanasam Ranbir Singh, S. R. M. Prasanna, "Automatic Pause Marking for Speech Synthesis," in proc. TENCON, 2017
2. **Bidisha Sharma** and S. R. M. Prasanna. "Pause Insertion in Assamese Synthesized Speech Using Speech Specific Features," in Proc. NCC, 2017.
3. Deepshikha Mahanta , **Bidisha Sharma**, Priyankoo Sarmah and S. R. M. Prasanna, "Text to Speech Synthesis System in Indian English," in Proc. TENCON, 2016.
4. **Bidisha Sharma**, Nagaraj Adiga and S. R. M. Prasanna "Development of Assamese Text to Speech Synthesis System," in Proc. TENCON, 2015.
5. Biswajit Dev Sarma , **Bidisha Sharma** , S. Ashwin Shanmugam , S. R. M. Prasanna and Hema A. Murthy , "Exploration of Vowel Onset and Offset Points for Hybrid Speech Segmentation," in Proc. TENCON, 2015.
6. **Bidisha Sharma**, and S. R. M. Prasanna. "Improvement of syllable based TTS system in Assamese using prosody modification," in Proc. INDICON, 2015.
7. **Bidisha Sharma** and S. R. M. Prasanna, "Faster Prosody Modification using Time Scaling of Epochs," in Proc. INDICON, 2014.

# Suggestions & Discussions During the Oral Examination

We would like to express my deepest appreciation to the viva-voce committee and external examiner for providing their constructive comments towards improving my contributions made in this thesis. Following are some major comments and suggestions that I will include in my future research.

- In Chapter 3, we have extracted the sonority feature from vocal-tract system, excitation source and suprasegmental features from the speech signal. The experiments carried out for the analysis and evaluation are based on the TIMIT database. The reference labels corresponding to different sound units are also obtained from TIMIT database segmentation. It is suggested that many sound units in the TIMIT database are not properly uttered and the segmentation is also not very accurate. Therefore, these boundaries are required to be examined before the analysis of the sounds units.
- In Chapter 3 we have not used the power spectrum to calculate the bandwidth. For each of the first three spectral peaks $(P_1, P_2, P_3)$, the difference between the frequencies in both side of the peaks, where amplitude of the HNGD spectrum is 0.707 of maximum value of corresponding peak is calculated. This difference is termed as bandwidth associated of the first three formant peaks $(B_1, B_2, B_3)$ in the rest of the thesis.
- Another crucial point is to analyze the duration of these sound units. The duration of vowels is much higher compared to the glides, liquids and consequently the number of pitch period for a segment of sound unit also changes. This fact needs to be taken into account while averaging the feature values corresponding to epochs within a frame.
- It is to be noted that all the experiments done in this work for synthesized speech are on held out sentences from the database used. However, in TTS a better way to evaluate the improvement in naturalness and intelligibility is to find mean opinion score for semantically unpredicted sentences. The duration of the test sentences should me long enough to perceive the synthesis quality.