EVENT-SYNCHRONOUS PROCESSING OF MISARTICULATED STOP CONSONANTS

VIKRAM C. M.

Event-Synchronous Processing of Misarticulated Stop Consonants

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Vikram C. M.

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

JUNE 2019

Certificate

This is to certify that the thesis entitled "**EVENT-SYNCHRONOUS PROCESSING OF MISARTICULATED STOP CONSONANTS**", submitted by **VIKRAM C. M.** (146102045), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

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

To

The lotus feet of

Goddesses **Sri Sringeri Sharada Devi**

for her infinite blessings

Acknowledgements

I express my sincere gratitude to my research supervisor, Prof. S. R. M. Prasanna for providing me an opportunity to work under his guidance. I am thankful for his continuous guidance in all aspects, constant motivation, and support throughout the doctoral studies. His dedication, discipline, and hard work are the source of motivation for me. Without his support, it would be completely impossible for me to carry out research work and bring the thesis to this level. I also would like to sincerely thank him for providing me with the financial support for attending conferences and workshops.

I am thankful to my doctoral committee members Prof. Samarendra Dandapat, Prof. Rohit Sinha, and Dr. Priyankoo Sarmah for their encouragement and valuable suggestions on my work. I am very much greateful to them for insightful comments and constructive criticisms, which helped me bring my work to the current form. I am very much thankful to my administrative supervisor Dr. Suresh Sudaram for his kind help and support. 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. Also, a heartful thanks to Dr. Abhishek Shrivastava of Design Department for his valuable suggestions and encouragement for my research work.

This thesis would become highly impossible without the help and support of Speech-language pathologists (SLPs) and staffs of All Indian Institute of Speech and Hearing (AIISH), Mysore. Especially, I would like to express my sincere thanks to Prof. M. Pushpavathi and Prof. Ajish K. Abraham for providing a wealth of knowledge about speech-language pathology. Their timely help and valuable suggestions helped me to formulate the objective of this thesis. I would like to acknowledge the help of Dr. Gopi Sankar, Dr. Navya, Dr. Gopi Kishor, Mrs. Deepthi, Ms. Nikitha, and Mr. Girish, during data collection and perceptual evaluation. Further, my thanks go to children, parents, and teachers for their cooperation during data collection.

I owe my special thanks to Dr. Prashanth Wali and Ajey Sir, who created interest in the signal processing subject during my undergraduate course and motivated to pursue higher studies at a premier institute like IIT. I would like to thank Dr. R. Kumara Swamy, and Dr. Veena Karjigi of Siddaganga Institute of Technology, Tumkur, for their help and encouragement to join Ph.D.

I am extremely grateful to my senior Dr. Nagaraj Adiga for his brotherly care, thought-provoking technical discussions, and patiently correcting my papers and thesis. Also, I acknowledge all my

seniors, Dr. Deepak, Ramesh Sir, Dr. Biswajit, Dr. Banriskhem, Dr. Rohan, Dr. Rajib, and Dr. Bidisha for mentoring me during my research life. I never forget my close friends Sishir, Akhilesh, and Protima, for their support since from the beginning of my Ph.D. to thesis correction. Useful technical discussions with them shaped my research in many aspects. A thankful note for my lab-mates Ramesh, Subhasis, Himakshi, Sarfaraz, Shikha, Mrinmoy, Sandeep, Saswati, Balaji, Sriram, Sukanya, Vineeta, Prabhakar, Shoubhik, Deepika, and the rest for their direct/indirect contributions during my stay at IIT Guwahati. Also, a heartful thanks to my project-mates Abhishek, Pankaj, and Abhilash. I would like to thank Mr. Sathish and Mrs. Kavya Sathish for their extreme care and hospitality. I am grateful to my friend Raveesh for his moral support.

During my PhD, I attended Interspeech conference with funding received from International Speech Communication Association (ISCA), for which I will be thankful forever. I convey my sincere thank to Ministry of Human Resources Department (MHRD), Govt. of India and Indian Council for Medical Research (ICMR), Govt. of India for providing fellowship for my Ph.D. thesis work. Also, I would like to acknowledge the Department of Biotechnology (DBT), Govt. of India for providing financial assistance for the data collection.

I salute to my grand-parents, parents, sister, brother-in-law, aunt, and uncle for their care, constant blessings, and silent prayers for my success and moreover, making me to stand in this position. Without their motivation and encouragement, this research work would become impossible.

Finally, my infinite salutes go to God for his showers of blessings in past, present, and future.

VIKRAM C. M.

Abstract

This thesis aims towards the development of an event-synchronous framework for the automatic evaluation of misarticulated stops. In the proposed event-synchronous approach, the speech signal anchored around a specific event is used to analyze the difference between normal and misarticulated stops as well as discrimination among the different categories of misarticulated stops. The work considers two important events, namely, epoch (glottal closure instant) and vowel onset point (VOP) for the processing of misarticulated stops. The applications of event-synchronous methods are demonstrated for the automatic evaluation of misarticulation stops produced by children with cleft lip and palate (CLP).

Automatic detection of epochs is a necessary stage in the epoch-synchronous analysis of speech. The first work of the thesis proposes an epoch extraction algorithm for the pathological children speech. The algorithm considers the vertical striations present in the wide-band spectrogram as the representative candidates of epochs. The time-frequency representation with better spectro-temporal resolution is estimated using single pole filter (SPF) based approach. The SPF-based time-frequency representation is processed to estimate the epoch locations. The performance of proposed epoch extraction algorithm is evaluated on a database containing simultaneously recorded speech and electroglattographic signals of pathological children.

In the next work, detected epochs are used to propose an epoch-synchronous feature to locate the VOPs. The detected VOPs are used for the segmentation of consonant-vowel (CV) transition regions. The SPF-based two-dimensional discrete transform (2D-DCT) coefficients computed from the CV transition regions are used for the classification of normal and misarticulated stops. The classifier shows a better performance for the proposed SPF-based 2D-DCT features than the conventional short-time Fourier transform (STFT) based features. Also, the knowledge of VOP is used for the discrimination of glottal stops from the non-glottal misarticulations.

Further, a special focus is given on the detection on the nasalized voiced stop, which is produced in the presence of velopharyngeal dysfunction and considered as a severe category of misarticulation in CLP speech. The proposed event-synchronous framework analyzes the effect of nasalization from voice bar and CV transition regions of voiced stops. Knowledge of glottal activity, syllable nucleus, low-frequency dominance, and VOP is used to segment the voice bar and CV transition regions. The segmented regions are epoch-synchronously processed to extract the spectral, spectro-temporal,

excitation source, and periodicity features. These features are used to develop a support vector machine based classification system for detection of nasalized voiced stops. The results showed that the combination of information from voice bar and CV transitions helps to detect the nasalized voiced stops in a better way than their individual contributions.

Finally, binary classifiers developed for the classification of normal vs. misarticulated stops, glottal vs. non-glottal, and nasalized vs. non-nasalized misarticulations are combined in a hierarchical manner to demonstrate a system for the multi-class classfication of misarticulated stops.

The major contributions of the current thesis are as follows:

- SPF-based time-frequency representation is proposed for the spectro-temporal analysis of speech. In particular, SPF is used for the epoch extraction and processing of CV transition regions.
- The time-localized vertical striations present in the SPF-based time-frequency representation are considered as the representative candidates of epochs. This vertical striation information is utilized to develop an epoch extraction algorithm for the pathological children speech.
- An epoch-synchronous temporal feature is proposed for the detection of VOP. The detected VOPs are used to develop a method for the screening of misarticulated stops produced by CLP children from normal.
- A method for the detection of nasalized voiced stops is proposed using epoch-synchronously computed spectral, spectro-temporal, excitation source, and periodicity features.
- A hierarchical framework is developed for the automatic classification of glottal, nasalized, and devoiced stops by combining the proposed event-synchronous features.

Keywords: Cleft lip and palate, event-synchronous processing, misarticulated stops, epoch, vowel onset point, and single pole filter.

Contents

List of Figures

List of Tables

List of Acronyms

2D-DCT	Two-Dimensional Discrete Cosine Transform
ASR	Automatic Speech Recognition
BPF	Band-Pass Filter
BPFS	Band-Pass Filtered Signal
BS	Backward Searching
CLP	Cleft Lip and Palate
CN	Controlled Normal
CNN	Convolution Neural Network
CV	Consonant-Vowel
DCT	Discrete Cosine Transform
DEGG	Differenced Electroglottograph
DFT	Discrete Fourier Transform
DR	Detection Rate
DPI	Dynamic Plosion Index
DYPSA	Dynamic Programming Projected Phase-Slope Algorithm
EGG	Electroglottograph
EFT	Exponentially Forgetting Transform
FAR	False Alarm Rate
FS	Forward Searching
HWFIR	Hamming Window-based Finite Impulse Response
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Model
GOP	Goodness of Pronunciation
HELPR	Hilbert Envelope of Linear Prediction Residual
HMM	Hidden Markov Model

IDA	Identification Accuracy
IDR	Identification Rate
IIR	Infinite Impulse Response
ILPR	Integrated Linear Prediction Residual
LOSO	Leave One Subject Out
LP	Linear Prediction
MFCCs	Mel-Frequency Cepstral Co-efficients
MWIP	Maximum Weighted Inner Product
MR	Miss Rate
MSP	Multi-Scale Product
RBF	Radial Basis Function
RoR	Rate of Rise
SCAI	Stop Consonant Articulation Index
SEDREAMS	Speech Event Detection using the Residual Excitation and a Mean-based Signal
SIGMA	Singularity in Electroglottograph by Multi-scale Analysis
SoE	Strength of Excitation
SFF	Single Frequency Filter
SNR	Signal to Noise Ratio
SPF	Single Pole Filter
SR	Spurious Rate
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TFR	Time-Frequency Representation
TM	Time Marginal
VOP	Vowel Onset Point
VOT	Voice Onset Time
VPD	Velo-Pharyngeal Dysfunction
YAGA	Yet Another Glottal Closure Instant Algorithm
ZFFS	Zero Frrequency Filtered Signal
ZFR	Zero Frequency Filter

1

Introduction

Contents

Objective

Misarticulated stops are commonly produced by the individuals living with cleft lip and palate (CLP), hearing disorders, motor speech disorders, and cognitive impairment. Speech technology based system for the evaluation of misarticulated stops can be used as an assistive tool by speech language pathologists during the treatment of individuals with speech sound disorders. Various categories of misarticulated stop exhibit different degree of variations in the acoustic characteristics of closure, burst, and transition regions. Due to the presence of diverse acoustic characteristics, automatic detection of misarticulated stops is a challenging task. Motivated by the importance of events in speech analysis, an event-synchronous framework is proposed for the processing of misarticulated stops. The proposed approach involves the processing of speech signal anchored around a specific event, which carries the significant information about the deviant articulation. The present work considers the glottal closure and vowel onset events for the processing of misarticulated stops. The features extracted around these events are used for the screening of misarticulated stops from normal and classification of the category of misarticulation. The applications of event-synchronous features are demonstrated for evaluation of misarticulated stops produced by CLP children.

1.1 Overview of misarticulated stop consonants

The production of stop consonants involves the dynamic movement of articulators [1]. At the initial stage of stops production, the vocal tract is completely closed by an articulator somewhere along its length (bilabial in /p/, /b/, alveolar in /t/, /d/, and velar in /k/, /g/). The articulatory constriction blocks pulmonic airflow and develops pressure behind the constriction. This pressure is referred to as intraoral pressure. When the constriction gets suddenly released, the intraoral pressure causes an abrupt flow of air, which is reflected as noise-like burst in the speech signal. Due to the involvement of such complex articulatory dynamics, stop consonants are one of the most frequently misarticulated class of sound units in speech sound disorders [2].

Speech sound disorder refers to the difficulties associated with motor production, perception, phonological representation of speech sounds, and speech segments. Speech sound disorders are commonly found in individuals living with structural anomalies (e.g., cleft lip and palate (CLP) and other craniofacial anomalies), motor speech disorders (e.g., apraxia and dysarthria), syndrome-related disorders (e.g., Down syndrome and galactosemia), and sensory-based disorders (e.g., hearing

impairment) [3–5]. Backing of alveolars (/t/→/k/), fronting of velars (/k/→/t/), devoicing errors (/d/→/t/), nasalized stops (/d/→/n/), addition errors ("kit"→"ikit"), and omission of finals ("bat" → "ba") are some of the typical examples for misarticulated stops [6]. Different categories of misarticulated stops provide information about the structural and functional disabilities associated with speech production. For example, nasalized stops provide important information about the improper functioning of velum [7], diagnosis of backing and fronting errors gives information about the tongue-palate contact [8]. Repetition of syllables containing stops, such as [pa-ta-ka, pa-ta-ka,..., pa-ta-ka] is used for the assessment of coordination between oral and laryngeal mechanisms in motor speech disorders [9,10]. Moreover, stop consonants form an important class of speech sounds in a language; hence, the production of misarticulated stops severely degrades the speech intelligibility [11, 12]. Therefore, identification and correction of misarticulated stops are necessary to improve speech intelligibility of the individuals with speech sound disorders.

Conventional approaches: Conventionally, misarticulated stops are perceptually identified by expert speech-language pathologists (SLPs). An SLP plays a central role in screening, assessment, diagnosis, and treatment of persons with speech sound disorders. Perceptual evaluation is often considered as a gold standard for documentation of speech sound disorders. However, it is a time-consuming process and results are subjective. A survey by the American Speech-Language-Hearing Association (ASHA) reported that SLPs have been spent about 67% of their time for the perceptual evaluation and administering speech therapy [13]. Reliability of the perceptual evaluation depends on several factors, such as experience of SLP, mental status of SLP at the time of evaluation, inter-rater, and intra-rater reliability issues [14]. Apart from perceptual evaluation, instrumental techniques such as X-ray, magnetic resonance imagining, ultrasound, electropalatography (EPG), etc. have been used for the diagnosis of the place of articulation related errors [15–17]. Applications of EPG and ultrasound imaging techniques have been demonstrated for the assessment of tongue-palate contact in individuals with CLP and Parkinson's disorder [8,18–20]. The results of the instrumental evaluation are objective and accurate than perceptual judgments. However, instrumental assessment is not a cost-effective approach and requires skilled professionals to handle the equipment. Also, different instruments are required for the evaluation of various categories of misarticulated stops. For example, EPG can be used for the diagnosis of deviant place of articulation in backing and fronting errors, but the same instrument may not be suitable for the evaluation of problems related to nasalized stops. Since speech

therapy is a long durational process and requires a routine assessment of speech production errors, the use of complex instruments may not seem to be appropriate.

Acoustic analysis: Deviant production mechanism associated with misarticulated sounds get reflected in the acoustic characteristics of speech signal. Acoustic characteristics of misarticulated stops are studied using the speech waveform and spectrogram [2,9,21]. These studies showed the presence of a significant difference between normal and misarticulated stops. The existence of acoustic differences between normal and misarticulated phonemes has been motivated several researchers to propose speech technology-based solution for the automatic evaluation of speech sound disorders [22–24]. These methods are developed for most of the phonemes present in a language, including stops. Speech processing based algorithms provide an objective result and help to overcome the limitations of perceptual evaluation. Implementation of speech processing based systems requires simple hardware like microphone and computer. Hence, these systems are simpler and cost-effective than the conventional instrument based approaches [25]. Due to the limited availability of SLPs at remote places, computerized speech evaluation tools can be used for the self-assessment by disordered subject. Also, the results obtained by speech processing algorithms can be presented in the form of interactive games to make therapy sessions more attractive for children [26].

1.2 Issues in the automatic evaluation of misarticulated stops

Automatic evaluation of misarticulated stops is a phone-level approach. Figure 1.1 provides an overview of speech technology-based system for the automatic evaluation of misarticulated stops. In this system, initially, word/sentence level stimuli containing target stop consonants are presented to the subject. For example, if the target is /b/, then word [bell] and sentence [bob is a baby boy] can be considered as stimuli. Response of the subject for the target stimulus is recorded and processed for the evaluation of misarticulated stops. Since the target stop consonant is embedded within a word/sentence, region corresponding to the target needs to be segmented. The segmentation is carried out using hidden Markov model (HMM)-based force-alignment or knowledge-based approaches [10, 23, 27]. From the segmented region, acoustic features such as Mel-frequency cepstral coefficients (MFCCs), spectral moments, etc. are computed. Finally, the extracted features are given to a classifier for the detection of the presence/absence of misarticulation and classification of the category of misarticulation. To develop a system for the automatic evaluation of misarticulated stops, the

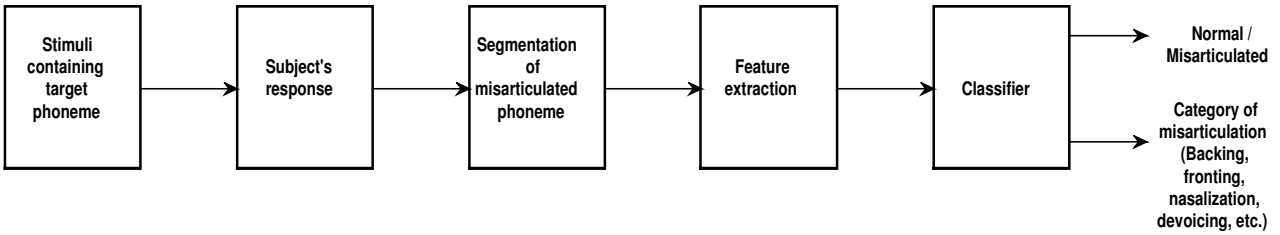

Figure 1.1: An overview of speech technology-based system for the automatic evaluation of misarticulated stops.

following issues need to be addressed.

1.2.1 Segmentation of consonant region

Segmentation of the subject's response for the target is an essential pre-processing step during the automatic evaluation of misarticulated stops. Performance of the misarticulation evaluation system critically depends on the accuracy of segmentation algorithm employed. Most of the phone-level misarticulation detection methods employ HMM-based automatic speech recognition (ASR) system for the segmentation purpose [23]. In the HMM-based method, phoneme boundaries estimated using force-alignment approach. HMM-based ASR system is more suitable for the detection of quasi-stationary vowel-like sounds than dynamic stop consonants [28]. Since acoustic characteristics of stops rapidly change with time, HMM-based methods may not be appropriate for the segmentation of stops [28, 29]. Moreover, acoustic models are trained for normal speech samples, where the models do not have knowledge about the misarticulated sounds. Therefore, in addition to dynamic characteristics of stops, the presence of acoustic mismatch between training and testing data will create another challenge for the segmentation of misarticulated stops using HMM-based methods [22]. Compared to HMM-based methods, knowledge-based approaches are more preferred for the segmentation of stop consonants [28,30]. The knowledge of closure-burst transition is used to segment the stop consonants in knowledge-based processing techniques [28]. But, the existence of burst onset itself a question in case of misarticulated stops. Burst may be absent or weakly evident due to the development of low intraoral pressure in individuals with CLP [2] and weak occlusion in motor speech disorders [10]. Rapidly varying acoustic characteristics and weakly evident burst create challenges for the segmentation of misarticulated stops.

1.2.2 Diverse acoustic characteristics of misarticulated stops

A stop consonant is comprised of multiple sub-phonetic units, such as closure, burst, and transition regions. Acoustic correlates of these sub-phonetic units vary with respect to the nature of voicing and place of articulation of stops. Speech waveform and spectrogram of a syllable containing normal unvoiced stop (/t/) are depicted in Figures 1.2(a) and (b), respectively. Speech waveform and spectrogram of a syllable containing normal voiced stop (/d/) are depicted in Figures 1.2(c) and (d), respectively. Articulatory closure phase of unvoiced stops corresponds to an interval of 'silence' (Figure 1.2(a)). Whereas, the closure phase of voiced stops associated with the low-energy voiced signal, which is referred as 'voice bar'(Figure 1.2(b)). Sudden release of articulatory constriction results in the production of an impulse-like event called 'burst onset'. Closure-burst transition carries information related to the manner of articulation of stops. Different stop consonants are produced by varying the place of articulation. The information regarding the place of articulation gets reflected in the burst and formant transitions [1, 31]. The formant transitions are anchored around the event called 'vowel onset'.

In different categories of misarticulated stops, the voicing, place, and manner of articulation affected in a different extent. Some of the misarticulations differ only in terms of place of articulation (e.g., backing and fronting errors), while few of them exhibit a noticeable difference regarding voicing and manner of articulation (e.g., devoiced and nasalized stops). Figures 1.2(e)-(j) illustrate the waveform and spectrograms of various categories of misarticulated stops. The waveform and spectrogram of devoicing error (/d/→/t/) are shown in Figures 1.2(e) and (f), respectively. Devoicing error is produced with the absence of glottal vibrations, due to which the voice bar is absent. Figure 1.2(g) shows the waveform of nasalization error (/d/→/n/), which indicates an increase in the energy of the voice bar region. Due to deviation in the manner of articulation, the absence of burst onset event is observed in the spectrogram of nasalization error (Figure 1.2(g) and (h)). The waveform and spectrogram of backing error (/d/→/g/) are shown in Figures 1.2(l) and (j), respectively. The voicing and manner characteristics of backing error remain the same as that of the normal stop. So, the waveform and spectrogram of backing error (Figures 1.2(l) and (j)) indicate the presence of voice bar and burst onset like normal voiced stops. However, the acoustic characteristics of the burst and transition region get affected due to the deviant place of articulation. From Figure 1.2, it can be noticed that the different categories of misarticulated stops exhibit a different degree of deviation from the normal.

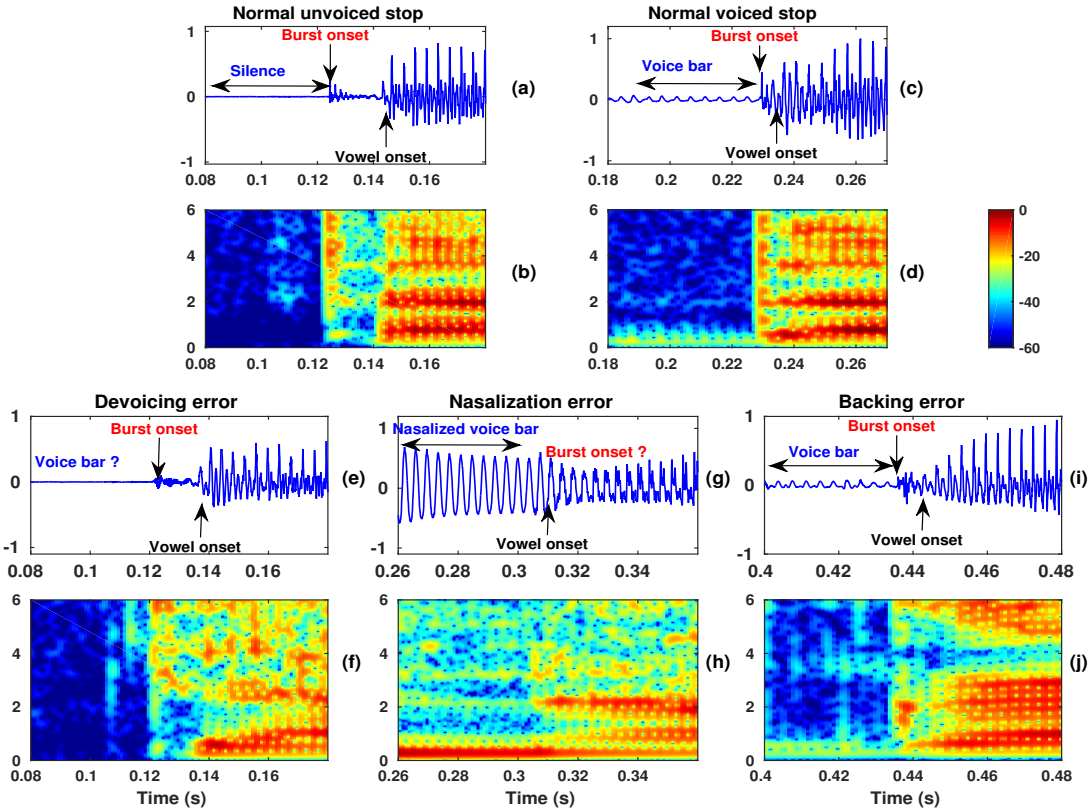

Figure 1.2: Illustration of speech waveform and spectrograms of syllables containing normal and misarticulated stops. (a) Waveform of normal unvoiced stop /t/ and (b) its spectrogram. (c) Waveform of normal voiced stop (/d/) and (d) its spectrogram. (e), (g), and (l) represent the waveform of devoicing , nasalization, and backing errors (substitution of /d/ by /g/), respectively and corresponding spectrograms are depicted in (f), (h), and (j).

Due to the presence of diverse acoustic characteristics, choice of acoustic feature for the processing of misarticulated stops becomes a challenge. For example, acoustic features related to voicing can discriminate the devoicing errors from normal, whereas the same features may not be appropriate for the detection of backing errors. Hence, different acoustic features required to represent the deviant characteristics of misarticulated stops.

1.2.3 Block-processing techniques

Most of the phone-level misarticulation detection systems employ block-processing based MFCCs or linear prediction cepstral coefficients to represent the acoustic-phonetic information [23]. Block-processing of speech involves the analysis of signal using a fixed window of 20-30 ms duration. Stop consonants comprised of multiple sub-phonetic units. In case of misarticulated stops, the information regarding the misarticulation is heterogeneously distributed over different sub-phonetic units. Some

of the sub-phonetic units have high similarity with normal, while only a few of them reflect the deviant production characteristics. For example, in case of backing errors, only burst and transition regions get affected, whereas the closure region is similar to that of normal stops. In such cases, the analysis should focus only on the burst and formant transitions, rather than considering entire stop consonant. The sub-phonetic units of stops span a very small duration. Often bursts of voiced stops have 2-5 ms. Therefore, to analyze the misarticulation information from different sub-phonetic units, the temporal resolution of analysis window plays an important role. However, in conventional block-processing methods, a single analysis of 20-30 ms frame may contain multiple sub-phonetic units and their transitions. Hence, using block-processing technique, it is difficult to analyze the dynamically varying acoustic characteristics of misarticulated stops.

1.3 Motivation for the present work

Speech technology based system for the automatic evaluation of misarticulated stops has a potential application in the clinical environments. The system can be used as an assistive tool by SLPs during the assessment of speech sound disordered speakers. In order to develop a system for the detection of misarticulated stops, it is required to address the issues discussed in Section 1.2. These challenges include mainly the segmentation of misarticulated regions, choice of acoustic features, and temporal resolution of conventional block-processing techniques. The thesis addresses the aforehead mentioned challenges using the event-synchronous approach [32,33]. The present work considers two major events of speech, namely, epoch and vowel onset point (VOP) for the processing of misarticulated stops. The thesis mainly focuses on the development of algorithms for the automatic detection of epoch and VOP events and the extraction of features around these events for the evaluation of misarticulated stops. The applications of proposed event-synchronous features are demonstrated on the database containing misarticulated stops produced by children with CLP. The motivation for the present work in the thesis is explained in this section.

1.3.1 Event-synchronous processing

In event-synchronous processing methods, the analysis is focused around a specific event, rather than giving equal importance to all the regions in the speech signal. Whereas, in conventional block-processing techniques, all the regions of speech are treated equally and analyzed using a single feature vector like MFCCs. In event-synchronous processing techniques, it is possible to apply different analy-

sis methods around different events. Various categories of misarticulated stop exhibit different degree of variations in the acoustic characteristics of closure, burst, and transition regions. Therefore, the event-synchronous methods provide great flexibility to select and analyze the speech signal around a specific event for the detection of misarticulation. Recently, transition regions anchored around vowel offset and onset events are used for the analysis of severity levels of Parkinson's disease [34]. Different features extracted around burst and vowel onset events are used for the detection of articulatory disorders in Parkinson's disease [10]. The advantages of event-synchronous approaches over conventional block-processing methods are the main motivation for the current work. In this thesis, the speech signal anchored around a specified event is used to analyze the differences between normal and misarticulated stops as well as among the different categories of misarticulated stops. Hence, the proposed approach is referred as *event-synchronous approach*.

1.3.2 Importance of epoch and vowel onset point events

The present work considers epoch and VOP events for the processing of misarticulated stops. Glottal closure instants are referred as the epochs of voiced speech [35]. In epoch-based processing of speech, analysis is focused only around the epoch, rather than the frame of 20-30 ms duration as in case of block-processing techniques. Therefore, epoch-based methods provide a better temporal resolution than block-processing techniques. Epoch-synchronously computed features are widely used for the acoustic-phonetic segmentation of speech [27, 36, 37]. With reference to stops, knowledge of epochs is useful in the accurate detection of burst onset and VOP events [33,38]. Epoch-synchronously estimated spectrum helps to analyze the vocal tract characteristics of each glottal cycle. The epoch-based spectrum is used for the analysis of dynamic trill sounds [36]. The epoch-synchronously estimated formant trajectories are used for the study of transition characteristics of stops [39].

The VOP corresponds to the instant at which vowel begins. In a consonant-vowel (CV) unit containing stop consonant, the CV transition region anchored around VOP contains the information about burst, frication, and formant transitions. VOPs are used as the anchor points for the recognition of CV units containing stop consonants in Indian languages [40, 41]. Experiments in [42–44] revealed that CV transitions carry the information about the voicing, place, and manner related aspects of stop consonants. The misarticulated stops exhibit deviation in terms of voicing, place, and manner of articulation. Therefore, CV transition regions anchored around VOP can be considered for the analysis of misarticulated stops. Also, VOP event is present in different categories of misarticulated

stops as demonstrated in Figures 1.2(e)-(j). Therefore, VOP can be considered as an alternative event for the burst onset for the segmentation of misarticulated stops.

Motivated by the importance of epochs and VOP, the present work explores the knowledge of these events for the processing of misarticulated stops. In particular, the knowledge of epochs is used for the accurate detection of VOPs, extraction of spectral features, and glottal activity detection. The VOP is used as an anchor point for the segmentation of CV transitions, whose applications are demonstrated for the detection of misarticulated stops.

1.3.3 Importance of misarticulated stops in CLP Speech

The thesis considers misarticulated stops produced by CLP children to demonstrate the applications of event-synchronous methods. Motivation for the applications developed in this thesis are explained as follows. CLP is a common most craniofacial anomaly and congenital disorder, which occurs about one in 700 live births across the world [45]. Individuals with CLP often associated with velopharyngeal dysfunction (VPD), where the velopharyngeal valve is not able to close the nasal port completely and consistently during the production of oral sounds [7]. Due to the presence of VPD, speakers with CLP are unable to build intra-oral pressure during the production of stops. In addition to VPD, the presence of anatomical abnormalities and mislearned articulation lead to the production of misarticulated stops in individuals with CLP [7, 46]. Weak, nasalized, glottal, pharyngeal, palatal, velar, and devoicing errors are the important categories of misarticulated stops produced by CLP speakers [47]. In CLP speakers, the glottal, pharyngeal, and nasalized stops are considered as severe consonant production errors. Whereas, palatalization and weak distortions are categorized as mild errors. Correction of severe errors requires both surgery and speech therapy, while speech therapy is sufficient for the correction of mild errors [46]. Therefore, the detection of category of misarticulated stops helps to plan for speech therapy and surgical interventions. Also, different speech therapy strategies have been developed for different categories of misarticulations [7]. Therefore, the major objectives for the assessment of misarticulated stops in CLP speech are [48]

- Screening of various categories of misarticulated stops from the normal.
- Identification of the category of misarticulation.

To identify the presence or absence of misarticulated stop consonant, the system should discriminate various categories of misarticulations (glottal, nasalized, weak, etc.) from normals. Also, to detect a

particular type of misarticulation, it is required to differentiate the desired category of misarticulated stop from other errors. However, to the best of our knowledge these issues are not yet addressed in the existing literature. Motivated by the importance of misarticulated stops in the treatment of individuals with CLP, the present work demonstrates the applications of event-synchronous methods for the screening of misarticulated stops from normal and classification of various categories of misarticulated stops in CLP speech.

1.4 Organization of the thesis

In the event-synchronous processing of speech, detection of events is a very important step and the performance of the system critically depends on the accuracy of the detected events. Hence, the algorithms for automatic detection of the considered events, i.e., epochs and VOPs are developed. Subsequently, the detected events are used to propose an event-synchronous framework for the detection of misarticulated stops. In each chapter, the impact of accuracy of the event detection algorithms and acoustic features on the automatic evaluation of misarticulated stops is studied. The organization of the thesis is explained as follows.

- Chapter 2 reviews the various event detection algorithms proposed in the literature for the analysis of stop consonants. Importance of different acoustic cues such as burst, VOT, formant transitions in the study of normal and misarticulated stops are reviewed. Based on the pros and cons of the existing misarticulated stop processing methods, the scope for the present thesis is discussed.

- In chapter 3, an epoch extraction algorithm is proposed by exploiting the spectro-temporal characteristics of impulse-like excitation at the glottal closure instants. Single pole filter (SPF) based filter-bank approach is used for the estimation of time-frequency representation (TFR), which gives better time-frequency resolution than conventional window-based approaches. SPF-based TFR is processed to locate the epochs. The performance of SPF-based epoch extraction algorithm is evaluated on the database containing pathological children voices and compared with the state-of-the-art epoch extraction methods. The SPF and epoch extraction are the necessary components of the algorithms proposed in subsequent chapters.

- Significance of VOP event in the processing of misarticulated stops in CLP is explained in Chapter 4. Epoch-synchronously computed temporal feature is proposed for the accurate detection of

VOP. The performance of the VOP detection algorithm is evaluated on the database containing the syllables containing normal and misarticulated stops, and compared with the state-of-the-art methods. The VOPs are used as anchor points to segment the CV transition regions, from which spectro-temporal features are extracted using SPF-based TFR. The spectro-temporal features for the screening of various categories of misarticulated stops in CLP speech from normal. Further, the application of VOP events in the discrimination of glottal vs non-glottal misarticulations is demonstrated.

- Detection of a particular category of misarticulation is addressed in Chapter 5 by taking the case of nasalized voiced stop. The nasalized voiced stops are analyzed using voice bar and CV transition region. A knowledge-based algorithm is proposed for the segmentation voice bar and CV transition regions. From the segmented regions, the spectral, spectro-temporal, excitation source, and periodicity aspects of nasalized voiced stops are analyzed. Finally, the proposed features are used for classification of normal vs nasalized voiced stops, and nasalized vs non-nasalized misarticulations.

- In Chapter 6, the combination of different event-synchronous features proposed in the previous chapters is carried out to demonstrate a system for the assessment of misarticulated stops. Spider plot-based visualization tool is developed for the evaluation of overall stop consonant articulation capability of an individual. By combining normal vs. misarticulated, glottal vs. non-glottal, and nasalized vs. non-nasal binary classifiers, a hierarchical system is proposed for the classification of normal, glottal, non-glottal, nasalized, non-nasalized and devoiced stops in CLP speech.

- Chapter 7 summarizes the contributions of the thesis and mentions possible future directions.

2

Processing of Normal and Misarticulated Stops: A Review

Contents

2.1 Introduction

The primary objective of this thesis is to develop a speech processing based system for the automatic evaluation of misarticulated stops. For the detection of misarticulated stops, it is necessary to understand the various acoustic cues and signal processing methods proposed in the literature for the analysis of stop consonants in normal and disordered speech. This chapter reviews the different acoustic features proposed in the literature for the analysis of normal and misarticulated stops. First, the existing ASR-based methods for the detection of misarticulated phonemes are discussed in Section 2.2. Methods for the automatic detection of burst and voice onset time (VOT), and their significance in the analysis of normal and misarticulated stops are discussed in Section 2.3. Importance of CV transition regions and methods for their processing are reviewed in Section 2.4. Application of articulatory closure phase in the analysis of misarticulated stops is explained in Section 2.5. A detailed review about the acoustic characteristics of misarticulated stops in CLP speech is provided in Section 2.6. The summary of the pros and cons of different acoustic features for the processing of misarticulated stops and scope for the present thesis are explained in 2.7.

2.2 Statistical speech recognizers

Hidden Markov model (HMM) based system and its extensions are the widely used statistical modeling based automatic speech recognition (ASR) systems. In the statistical approach, acoustic-phonetic knowledge of different sound units is captured by a single feature vector like Mel-frequency cepstral coefficients (MFCCs). The MFCCs are computed using block-processing approach, and they capture the gross spectral shape information of speech. In HMM-based ASR systems, the production knowledge of speech sounds is implicitly incorporated in the form of MFCC feature, and its first and second order derivatives. In the literature, ASR is used for the automatic scoring of misarticulated sounds [24, 49–51], estimation of boundaries of misarticulated phoneme [23, 52], and Substitution, Omission, Distortion, and Addition (SODA) error analysis [53, 54].

Phoneme likelihood values obtained from ASR are used to obtain the pronunciation scores. Goodness of pronunciation (GOP) is one of the most commonly used ASR-based measures for the assessment of misarticulated phonemes [24, 49–51, 55]. GOP is computed as the ratio between the likelihood of an expected phoneme sequence to the most likely observed phoneme sequence [49]. Using GOP, the correctness of each phoneme present in a word/sentence can be analyzed. Text-constrained force-

alignment using ASR is the first step in the computation of GOP. This step forces the ASR to detect a predetermined phoneme sequence of the target word. In addition to that, force-alignment gives the time-boundaries of each phoneme present in the target word. Using the phoneme likelihood values of force-aligned and recognized sequence, the GOP (λ) is computed as [51]

$$\lambda(q_i) = log\left(\frac{P(\mathbf{O}|q_i)P(q_i)}{\sum_{j\in J} P(\mathbf{O}|q_j)P(q_j)}\right)/N_{q_i} \tag{2.1}$$

where q_i is the expected phoneme, $\mathbf{O}$ represents the observation (phoneme segment), $q_j,\ j = 1, 2...J$ corresponds to phoneme set, and N_{q_i} corresponds to the duration of observed phoneme. GOP is widely used in the computer-assisted pronunciation teaching in second language learners [49]. Subsequently, applications of GOP are developed for the evaluation of misarticulated phonemes in speech sound disorders. In the literature, GOP is used for the detection of the phone-level misarticulations in apraxia [55] and Autism [51]. Combination of phonological features with conventional MFCCs for the computation of pronunciation scores is proposed for the dysarthria speech [54].

Conventionally, GOP is computed using the ASR trained for normal adult's speech. Therefore, to use ASR for the evaluation of misarticulations in children speech, acoustic mismatch between children and adult speech need to be addressed. Several feature space normalization techniques (e.g. vocal tract length normalization) and model adaptation techniques (e.g. maximum a posteriori and maximum likelihood linear regression adaptation) have been used to solve the issues related to the mismatch between children and adult speech [56]. In addition to this, the training data set of ASR contains only the data of normal speakers. Since the acoustic models are not trained for the misarticulated phonemes, GOP becomes less reliable for the assessment of phone-level misarticulations [51]. To overcome these problems, recently, explicit models are developed for the all possible categories of misarticulations produced for the target and these models are used for the computation of GOP [51]. However, a large amount of data is required for the development of acoustic models for each class of misarticulation, which is difficult under practical conditions.

In some works [23, 52], force-alignment is used for the estimation of boundaries of misarticulated phonemes. From the segmented regions, acoustic features are extracted for the evaluation of misarticulated phonemes. Force-alignment approach is used for the detection of nasalized, pharyngealized, and laryngealized consonants in CLP speech [23]. Force-aligned boundaries are used for the analysis /k/ and /t/ production errors in speech sound disorders [52]. The major limitation of HMM-based force

alignment for phonetic segmentation is that phone boundary information is not explicitly represented in the acoustic models. The estimated boundaries are resulted by the likelihood maximization over possible state alignments. On the TIMIT database, HMM-based force-alignment reported 80-89% agreement within 20 ms of the manual labels [57, 58]. Several knowledge-based and articulatory and acoustic measures are used to improve the accuracy of force-aligned boundaries [57–59].

HMM-based speech recognizer is used for the automatic detection of substitution errors in hearing disorders [53]. Authors of [53] have compared the decoded phoneme sequence by HMM with respect to the target's transcription. The comparison method is used for the detection of typical misarticulated stops in hearing disorders, such as devoicing (/b/→/p/) and backing errors (/t/→/k/). However, HMM-based speech recognizer showed confusion between voiced and unvoiced stops in normal speech [60, 61]. Due to the presence of confusion among various classes of stops, ASR-based approach may not be suitable for the detection substitution errors.

2.3 Burst and voice onset time

Burst and voice onset time (VOT) are the important acoustic cues used in the event-based processing of stop consonants. In the literature, several methods have been proposed for the automatic detection of burst and VOT [28, 33, 62]. Burst and VOT carry the information regarding place of articulation and voicing. Hence, these cues are widely used for the detection of erroneous place of articulation and devoicing errors in speech sound disorders [63].

2.3.1 Burst detection methods

Burst is a transient event, characterized by an abrupt rise in the acoustic energy. The energy difference between the adjacent widowed speech frames is used for the detection of burst onset [64]. The rate-of-rise (RoR), i.e., the energy difference between successive speech frames computed using six specific frequency bands is proposed for the burst detection [28]. King and Taylor proposed a neural network based burst detection algorithm using short-time energy, MFCCs, along with its velocity and acceleration coefficients (39-dimensional feature) [65]. A number of spectral and temporal measures such as energy ratio, zero-crossings, linear prediction (LP) coefficients with the combination of multilayer perception and Bayesian classifier are used for the burst detection [66]. In [67], periodicity measure and energy difference feature are used for burst detection. Usage of the periodicity measure avoids the false detection of voice onset as a burst. Burst onset is considered to be impulse-like in

nature, and the impulse-like signal has a flat power spectrum. Hence, spectral flatness is used to characterize the burst onset in [68]. Log energy of the signal, log energy of high-pass filtered (3kHz) signal, and Wiener entropy is used to locate the burst. Joint spectro-temporal features and random forest classifier are explored for the burst detection in [30]. A non-linear measure called plosion index computed from Hilbert envelope of high-pass filtered speech is proposed for the burst detection in [29]. Plosion index based approach uses normalized cross-correlation coefficient computed using the speech segments of consecutive pitch periods. The normalized cross normalized cross-correlation coefficient is used to avoid the false detection of vowel onset as burst onset. Burst onset detected by plosion index showed higher temporal accuracy than the frame-based energy difference operators.

2.3.2 Burst related features

Spectrum estimated from the 20 ms speech signal followed by the burst onset is used to analyze the place of articulation of stops [69]. The analysis showed that the spectral energy of bilabials (/p/, and /b/) is concentrated in the range of frequencies 500-1500 Hz. Dental and alveolar stops have burst energy concentrated in the higher frequencies (above 4000 Hz), whereas velars showed a concentration of burst energy in the range of 1500-2500 Hz. The global shape of the spectrum computed over the first 20-50 ms from the burst onset has been analyzed for the detection place of articulation [1, 70].

Studies by Blumstein and Stevens [70] revealed that velars have a compact spectrum (strong spectral peak at mid-frequency), whereas bilabial and alveolar stops showed diffuse-falling and raising spectra, respectively. Forrest et al. [71] analyzed the spectral moments of burst. In their study, the first four spectral moments are computed: *first moment*-mean or centre of gravity of spectrum, *second moment*-distribution of energy around mean, *third moment*-spectral tilt, and *fourth moment*-degree of peakedness of the spectrum. The analysis of spectral moments showed a significant difference among velar, alveolar, and bilabial stops [71]. Authors of [72] have analyzed the burst spectral moments for the discrimination of voiced and unvoiced stops in American English.

Burst spectrum encodes the information about the place of articulation. Hence, it is mainly used in the analysis of backing and fronting errors. Using spectral moments, /t/ and /k/ production errors in phonologically disordered children are analyzed [73]. In [73], first, a linear discriminant function derived for the spectral moments of /k/ and /t/ produced by normal children. The derived discriminant function is applied to the spectral moments derived from phonologically disordered children. There is no discrimination observed between /t/ and /k/ produced by phonologically disordered children.

In [74, 75], burst spectral moments of /t/ and /k/ produced by CLP speakers are analyzed. Results showed that there is a reduced difference between the spectral moments of /t/ and /k/ produced by CLP speakers. This reduction is due to the shift in the place of articulation from alveolar to the velar region.

2.3.3 Voice onset time

VOT is defined as the interval between the onset of stop-burst and the onset of laryngeal vibrations [76]. In case of unvoiced stops, voicing leads the burst onset and VOT is referred to as lead VOT. Whereas, in case of voiced stops, voicing lags the burst onset and VOT is referred to as lag VOT. Thus, VOT is considered as an important cue for the discrimination of voiced and unvoiced stops. VOT is successfully used in clinical applications for the detection of voicing errors in speech sound disorders [63].

Estimation of voice onset time: The locations of both burst and voice onsets are required for the estimation of VOT. In wide-band spectrogram, VOT is characterized by the onset of quasi-periodic vertical striations followed by burst onset [76]. Onset of visible energy in the first formant or higher formants in the wide-band spectrogram is considered as an acoustic cue for the computation of VOT [77]. Motivated by the spectrographic characteristics of burst and voice onsets, a VOT estimation method is proposed using reassigned spectrogram [78]. Hansen et al. proposed Teager energy operator based approach for the detection VOT in word initial unvoiced stops [79]. In [80], HMM and random forest-based approach is proposed for the VOT estimation [80]. Here, MFCC-HMM-based force-alignment is used to get an initial estimate of phoneme boundaries of stops. Further, joint spectro-temporal features computed from short-time Fourier transform (STFT) are used to train a random forest classifier for the estimation of VOT. A discriminative structured prediction approach is proposed in [62], where several acoustic features, namely, log of total energy, log of low-frequency energy (500-1000 Hz), log of high-frequency energy (> 300Hz), Wiener entropy, zero crossing rate over 5 ms and binary voicing decision are used for the VOT estimation. In the above-mentioned approaches, the features for VOT estimation are computed by the frame-wise processing of speech [62, 78]. Since voice onset is an instantaneous event, the windowing operation affects the temporal accuracy of estimated VOT. By addressing the issues related to frame-based features, the knowledge of epoch is used for the VOT estimation in [81]. Here, epoch-synchronously computed weighted inner product and zero-crossing difference measures are used for the detection of voice onset event. Recently, deep recurrent

neural network-based approach is proposed for the estimation of lead and lag VOTs in voiced stops [82], where manually marked prevoicing onset, burst onset, and voice onset events. To train the neural network, features proposed in [62] are used. Bayesian Step Change-point Detection operator is used for the estimation of VOT in Parkinson's disorder [10].

Significance of VOT: VOT has been used for the analysis of the place of articulation of stops [76]. VOT vary with respect to the mass of articulator [83]. Therefore, among velar, alveolar and bilabial stops, VOT values showed a decreasing trend from velar to bilabial. VOT has been used as an important temporal cue for the classification of place of articulation [31, 84]. Lead VOT of voiced stops is smaller than unvoiced stops; hence, VOT is used to improve the recognition rate of voiced and unvoiced stops in HMM-based ASR [60].

In the literature, VOT is widely studied for the analysis of errors related to voicing and place of articulation [9, 21]. Word-initial voiced stops have negative VOT, while unvoiced stops have positive VOT. Due to this contrast, VOT is used as a potential cue for the analysis of voicing errors [21]. A comparative study of VOT values of alveolar stops produced by normal and CLP children is carried out in [85], where the VOT values found to be greater in CLP children than normal. VOT is used to analyze the coordination between laryngeal and supra-laryngeal activities in dysarthria [9, 10]. VOT estimated using smoothed Hilbert envelope of speech is used for early detection of Parkinson's disorder [86]. In these studies, the speakers with dysarthria indicated the largest VOT values than normal. Therefore, increased VOT value is considered as a potential indicator of the slow articulatory movements in dysarthria.

2.4 Transition regions

If the stop consonant is followed by a vowel or voiced sound, then after the release of oral constriction, vocal tract requires some time to attain the gesture of following sound unit. Due to variation in the vocal tract shape, the formant frequencies rapidly change over the transition duration. Similarly, if the stop is preceded by a vowel, then the formant transitions around the vowel offset carry information about the place of articulation of stops [1]. The acoustic cues related to onset and offset transitions are used in the literature for the analysis of stop consonants [31, 59, 87]. In this section, various approaches proposed in the literature for the study of the transition region in normal and misarticulated stops are discussed.

2.4.1 Formant transitions

Delattre et al. analyzed that significance of formant transitions for the recognition of the place of articulation [88]. F_2 and F_3 transitions are considered as the essential cues for the analysis of place-related information. In [42], the effect of voicing on the formant transitions is studied. This study revealed that F_1 transition of unvoiced stops is relatively abrupt than voiced stops. Formant transitions are also used to study the difference between nasals and voiced stops [44]. The analysis showed that F_1 transition is relatively smooth for nasals than voiced stops.

Classical formant estimation methods are based on the LP and cepstrum analysis of speech [89]. Peaks present in the spectral envelope are considered as the formant locations. To avoid spurious peaks, formant tracking algorithms are proposed in the literature [90,91]. In case of children speech, the presence of high pitch affects the estimation of formants using LP analysis. Weighted LP and extended weighted LP methods are proposed for the estimation of formants from high pitched speech [92, 93]. Two-dimensional localized Fourier transform based approach is proposed for the estimation of formants from high-pitched speech [94]. In most of the methods, formants are estimated by processing speech frames of 20-30 ms duration; hence the methods are most suitable for the quasi-periodic vowel-like sounds [95]. In the stop-vowel transition region, formant values change rapidly, where the usage of 20-30 ms durational window smears the temporal dynamics of formants. To deal with such rapid transitions, epoch-synchronous LP analysis method is proposed in [39, 96]. Here, the speech signal anchored around each epoch is used for the LP analysis.

The formant frequencies extracted from the LP spectrum, and their first and second order derivatives are proposed for the classification of unvoiced stops in [39]. The nature of formant transitions varies according to the following vowel. Hence, formant transitions are considered as context-depended cues. Sussman et al. proposed acoustic locus equations as a context-independent cue [97]. The locus equations are constructed by measuring the second formant values computed at the vowel onset and mid point of the vowel. Locus equation computed for the second formant (F_2) as

$$F_{2_v} = kF_{2_o} + C \tag{2.2}$$

where F_{2_v} and F_{2_o} are the formant values at the middle portion of vowel and onset of the vowel, respectively. k and C are slope and intercepts of the locus equations, respectively. k and C are estimated using the stops-vowel CV units of different age, gender and contexts. The slope and intercept

of the F_2 locus equations are used for the detection of place of articulation [97, 98].

Formant transitions are used for the analysis of dysarthria in amyotrophic lateral sclerosis [99]. In [99], several acoustic measures are proposed by measuring F_1 and F_2 formant trajectories from speech spectrogram. These measures are (a) transition extent: amount of frequency change along the transitional segment of a trajectory, (b) transition-duration: duration of the transitional segment of the trajectory, (c) averaged transition rate or slope, and (d) starting frequency. Disordered speakers showed relatively flat formant trajectories than the normals. F_2 transition is used for the analysis of the distorted place of articulation in dysarthric speakers [9]. F_2-slope index is used for the assessment of severity of intelligibility in dysarthric speakers [9]. Lindblom et al., estimated the duration of formant transition in normal and dysarthric speakers using exponential curve fitting approach [100]. Second formant trajectory $F_2(t)$ is approximated using the following equations.

$$F_2(t) = (F_{2L} - F_{2T})e^{-\alpha t} + F_{2T} \tag{2.3}$$

where F_{2T} and F_{2L} are the second formant values at begining of the transition region and middle part of the vowel, respectively. The α represent the time constant, which is determined using the equation,

$$\alpha = -\frac{1}{t} ln(\frac{F_2(t) - F_{2T}}{F_{2L} - F_{2T}}) \tag{2.4}$$

Time constant values estimated for the normal and dysarthric cases. The analysis showed that time consonant values are higher for dysarthric speakers than normal. The increased time consonant values revealed the presence of slower articulatory movements in dysarthric speakers. Stop consonantal space constructed by measuring F_2 and F_3 at vowel onset of bilabial, alveolar, and velar stops. The stop consonantal space is used for the analysis of CLP speech [101]. The analysis showed that the area of stop consonantal space gets reduced for CLP speakers when compared to that of normal. The decreased consonantal space area indicates the reduction in the articulation capability in CLP speakers. Formant values estimated using LP-analysis are used for the analysis of glottal stops in cleft palate speech [102]. F_1 and F_2 transitions are used for the assessment of stop consonant production errors in Parkinson's disorder [10, 103].

2.4.2 Dynamic spectral cues

Studies by Kewley-port [104] claimed that a static burst spectrum does not provide complete information regarding the place of articulation. Kewley-port [104] analyzed the time-varying properties of

the stops such as the tilt of the burst spectrum, the existence of a sustained peak in the mid-frequency region and a delayed F_1 onset. Lahiri et al., [105] showed that both dental and alveolar stops have diffuse raising burst spectrum; however, dynamic variation of spectral shape from burst onset to vowel can discriminate them. Nossair et al., [106] analyzed the time-varying spectral characteristics speech segment considered over 60 ms duration from the burst onset. The dynamic spectral cue showed better performance in the classification of place of articulation when compared to static burst spectrum and formant transitions. Suchato [31] proposed several features based on the difference between spectral characteristics of burst and vowel. These acoustic cues are used for the classification of place of articulation. The spectrogram patch considered over few milliseconds from burst onset is analyzed using two-dimensional discrete cosine transform (2D-DCT) and bivariate polynomial coefficients based joint spectro-temporal features [59]. Further, the joint spectro-temporal features are used for the classification of place of articulation. In [59], burst onset is computed automatically using energy difference operation [28] and vowel onset point is located using Hilbert-envelope of LP-residual proposed in [38].

Joint spectro-temporal features are used to study the discrimination between /k/ and /t/ produced by speech sound disordered children [52]. Here, two-dimensional cosine transform (2D-DCT) based joint spectro-temporal features are computed from the Bark-scaled spectrogram patch, which is considered over 60 ms from the burst onset. 2D-DCT features of normal /t/ and /k/ are used to train SVM and GMM models. Using these models, proximity index is computed to analyze the goodness of /t/ and /k/ production in phonologically disordered children.

2.4.3 VOP based analysis of CV transition

In a CV unit containing stop consonant, vowel onset point (VOP) separates the vowel and consonant regions. VOP corresponds to the beginning of the first glottal cycle in the vowel region [61]. If the unvoiced stop is followed by a vowel, then the voicing onset coincides with the vowel onset. Whereas in case of voiced stops, vowel onset is characterized by the onset of formants and high-frequency energy [61]. In a CV unit containing stop consonant, region before the VOP includes the burst, frication, and aspiration noise. Whereas, the region after the VOP carries information about formant transitions. VOPs are used as the anchor points to segment the CV transition region. Further, the segmented transition regions are used for the recognition of CV units containing stop consonants [40, 41, 107]. In these methods, the detection of VOP is a primary task. Based on the knowledge of VOP, the transition region is segmented and analyzed for the recognition for the CV units.

VOP detection methods: VOP detection using energy, zero-crossing rate, fundamental frequency, and vocal tract resonances is carried out in [108]. Neural network and dynamic time warping methods are proposed for the VOP detection in [109]. Abrupt changes in the excitation source and vocal tract system characteristics are considered as the cues for the detection of VOP [38]. The excitation source information derived using LP residual is analyzed for the detection of VOP [110]. Combination of the excitation source, spectrum, and modulation spectral energies for the detection of VOP is proposed in [111]. Here, Hilbert envelop of LP residual, the sum of 10 largest peaks in the discrete Fourier transform (DFT) spectrum, and energy of modulation are used to capture the excitation source, spectrum, and modulation spectral characteristics, respectively. Authors of [112] proposed a VOP detection algorithm by processing the spectral energy present in the band 500-2500 Hz, where the spectral energy is computed around the glottal closure instants. Epoch-synchronously computed amplitudes of the first three formants using group delay spectrum are used for the detection of VOP [113]. Detection of VOP using excitation source features computed using zero-frequency filtered signal, and Hilbert envelope of LP residual is carried out in [114]. Evidence derived from the amplitude envelope of the vowel emphasized amplitude modulated (AM)-frequency modulated (FM) signal is used for the VOP detection [115].

Applications of VOP events: The transition region anchored around VOP is used for the detection of CV units in Indian languages [40,41,112]. In these works, first, 39-dimensional MFCCs are computed using the conventional block-processing approach. Further, frame-wise computed MFCCs of transition region are concatenated to form a feature vector. For example, if MFCCs are calculated using a frame size of 20 ms with a shift of 5 ms, then 10-frames of MFCCs (5 left and 5 right sides of VOP) are concatenated to form a 390-dimensional feature vector. The concatenated feature vector is used for the training of SVM classifier for the recognition of the CV units.

Recently, the processing of transition regions using convolution neural network (CNN) is carried out for discrimination of normal and Parkinson's disordered speech [34,116]. The transition regions of [pa-ta-ka] repetition are segmented using PRAAT-based voicing decision [117]. The wavelet transform based time-frequency representation of the segmented transitions used to train a CNN. Further, CNN features computed from transition regions are used for the detection of normal and misarticulated stops in Parkinson's disorder. In [34], CNN features derived from the transition regions are used for the analysis of different stages of Parkinson's disorder.

2.5 Closure region

Articulatory closure phase of unvoiced stops is characterized by the presence of silence, whereas voiced stops by the presence of low-energy voiced signal (voice bar). The voice bar is an essential cue for the discrimination of voiced stops from unvoiced ones. In addition to voicing, duration of closure provides information about the place of articulation. As the place of articulator moves away from the glottis, the volume of the supra-glottal cavity increases, which reduces the intraoral pressure. Hence, to acquire desired intraoral pressure, the closure duration increases as the place of articulation moves away from the glottis [42]. Due to this reason, the closure duration for velar, dental and bilabial stops follows an order of velar<dental<bilabial [33]. Automatic detection of closure bar using plosion index evidence and burst onset is carried out in [33]. Automatic detection of voice bar is carried out in [118]. To detect low-energy voice bars, the excitation source and spectral features are proposed [118]. Low-pass filtered signal energy (cut-off frequency 1.5 kHz) is used for the detection of closure region in Parkinson's disorder [10]. The usage of low-pass filtering helped to remove the high-frequency noise caused due to weak occlusion in Parkinson's disorder.

The presence or absence of voicing in the closure region and the duration of closure region are used for the analysis of misarticulated stops. In [21], closure duration of normal and phonologically disordered children is analyzed, where phonologically disordered children indicated a longer closure duration than normal. Also, due to devoicing errors, an absence of voice bar is noticed for the phonologically disordered children. Impact of VPD on the closure duration in CLP children is analyzed in [119]. Due to reduced oral pressure, the individuals with CLP showed prolonged closure duration. In Parkinson's disorder, incomplete occlusion increases the turbulence noise and voicing energy during the articulatory closure phase. Hence, the energy of closure bar is used for the discrimination between normal and Parkinson's speakers [10].

2.6 Acoustic characteristics of misarticulated stops in CLP speech

The present thesis focused on the analysis of misarticulated stops in CLP speech. Therefore, it is necessary to review the literature related to production mechanism and acoustic characteristics of misarticulated stops in CLP speech. Misarticulated stops in CLP speech are broadly categorized into two categories, i.e., obligatory (passive) and compensatory (active) errors [120]. During the production of obligatory misarticulations, the speaker will not make any attempt to avoid the consequences of

structural abnormality. Whereas, in case of active misarticulations, the speaker tries to compensate the effects of structural defect by shifting the place of articulation. Obligatory errors include nasalized and weak stops. Compensatory errors include the glottal, pharyngeal velar, and palatal substitutions. In the literature, compensatory errors are classified into two broad categories [121]. The first one is abnormal backing of oral targets to post-uvular place, which includes the pharyngeal and glottal stop substitutions. The second one is abnormal backing of oral targets, but the place remains oral, these errors include the backing of alveolar stops into the velar or palatal region and backing of velar stops into the uvular region. The production mechanism and acoustic characteristics of some important classes of misarticulated stops described in the literature are reviewed here.

Weak stops: In individuals with CLP, presence of VPD leads to the loss of airflow through nasal tract. As a consequence of the loss of airflow, low oral pressure will be developed during the articulatory closure phase of stops. Reduced intraoral pressure results in the production of weak stops [47]. During the production of weak stops, place of articulation remains the same as that of normal. However, stop is perceived to be weak due to the loss of intraoral pressure [47]. Closure region of weak stops contains the noise components due to the nasal air emission. Figures 2.1(a) and (b) represent the waveform and spectrogram of normal unvoiced stop (/t/), respectively. Figures 2.1(c) and (d) depict the waveform and spectrogram of weak stop produced for /t/, respectively. The spectrogram of the weak stop (Figure 2.1(d)) indicates the presence of noise components before the burst onset. The energy of the burst is proportional to intraoral pressure. Hence, the reduced intraoral pressure in weak stops results in burst weak energy [2, 122].

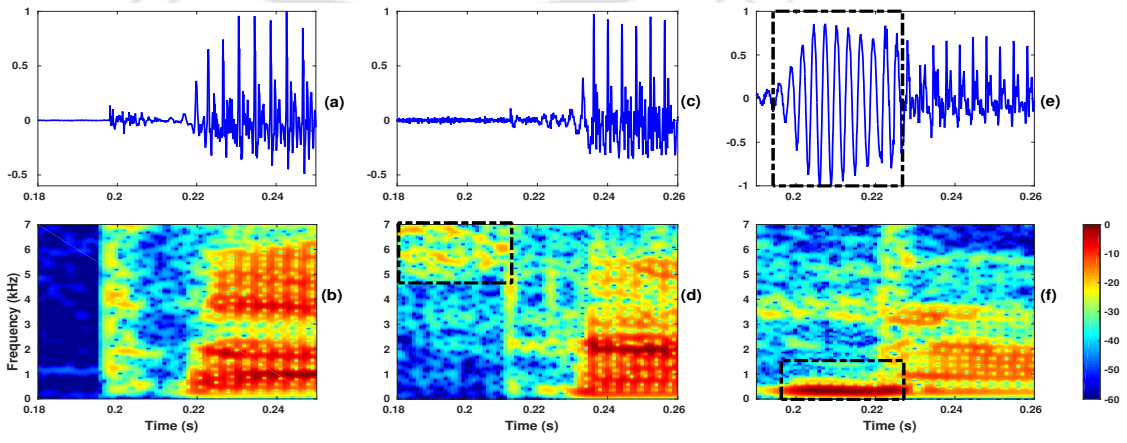

Figure 2.1: (a) Waveform of normal unvoiced stop and (b) its spectrogram, (c) Waveform of weak stop and (d) its spectrogram, (e) Waveform of nasalized stop and (f) its spectrogram. The rectangular box in the spectrogram (d) indicates the presence of high-frequency noise due to nasal air emission. The region corresponding nasal murmur of nasalized stop is indicated on its (e) waveform and (f) spectrogram.

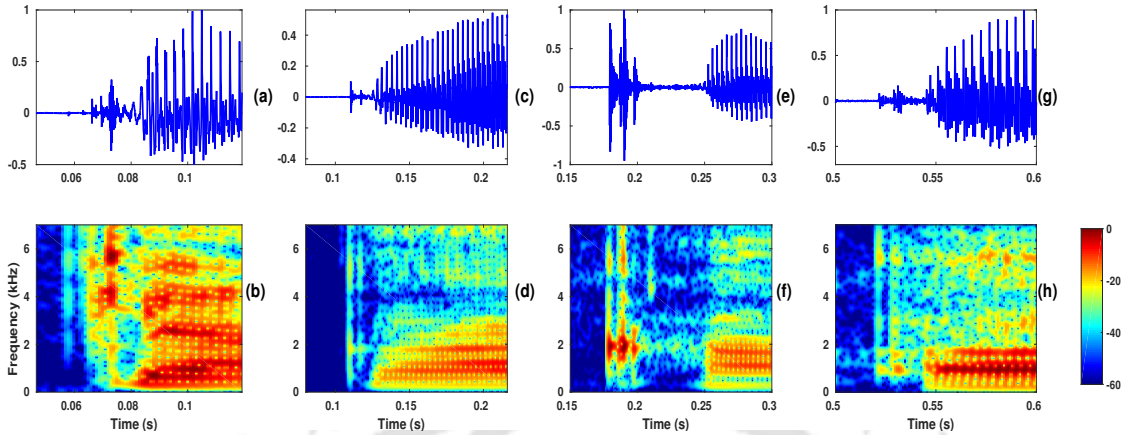

Figure 2.2: (a) Waveform of palatalized stop and (b) its spectrogram, (c) waveform of velar substitution and (d) its spectrogram, (e) waveform of pharyngeal stop and (f) its spectrogram, (g) waveform of glottal stop and its spectrogram.

Nasalized stops: Nasalized stops are produced in presence of moderate-to-severe VPD [7]. During the articulatory closure phase of stops, the presence of VPD leads in the loss of airflow though the nasal tract. Due to the presence of VPD, the stop consonants (/b/, /d/, /g/, /p/, /t/, /k/) perceived like nasal consonants (/m/, /n/, and /ng/). Philips et al., analyzed the spectrogram of nasalized stops and reported that the presence of nasal murmur in the closure region of stops [2]. Figure 2.1(e) and (f) show the waveform and spectrogram of nasalized stop produced for target unvoiced stop (/t/), respectively. Instead of silence in the closure phase, nasal murmur having strong resonance around 250 Hz is observed in the spectrogram shown in Figure 2.1(f). Nasalization is considered as the severe consonant production error, where the manner of articulation of stop is replaced by that of nasal consonant.

Palatalized stops: Shift in the place of articulation of alveolar stops towards palatal region is reported in individuals with CLP [123]. Palatalization of stops is due to the presence of dental occlusions and hard palatal fistula in dental/alveolar region [15, 124]. Production of palatalized stops for the alveolar/dental stops is reported based on the studies using X-ray and ultrasound imaging techniques [15,123]. Studies about the relationship between maxillary arch dimensions and production of mid-dorsum palatal stops revealed that limited anterior oral cavity space due to restricted (or collapsed) maxillary arches result in the palatalization errors [74]. Spectral moments of burst are used to analyze the acoustic characteristics of palatalized stops in [74, 75]. Figures 2.2(a) and (b) represent the waveform and spectrogram of palatalization error produced for dental stop /t/, respectively. Normal dental stops have spectral energy concentrated in the higher frequencies (above 4000 Hz).

But, palatalization of dental stops shifts results in the shift of spectral prominence towards lower frequencies as is indicated in Figure 2.2(c).

Velar substitutions: Like palatalization of dental/alveolar stops, backing of dental/alveolar stops into velar region is also reported in CLP speech [47]. The velar substitution error is caused due to the presence of oro-nasal fistula, dental occlusion. Figures 2.2(c) and (d), represent the waveform and spectrogram of velar substitution error, respectively. From the spectrogram (Figure 2.2(d)), it can be noticed that the spectral prominence of burst gets shifted to the mid-frequency region (1500-2500 Hz). Whereas, the spectral energy is concentrated in the region above 4000 Hz for the target dental/alveolar stops.

Pharyngeal stops: Pharyngeal stops are produced by shifting the place of articulation towards the pharyngeal cavity [15, 47]. Pharyngeal stop is most commonly produced compensatory error for the velar stop (/k/ and /g/) [15]. Philip et al., analyzed the spectrographic characteristics of pharyngeal stops, where they found that pharyngeal stops have strong burst energy, concentrated in the low-frequency region [2]. In addition to burst, the vowel followed by a pharyngeal stop showed closely spaced F_1 and F_2 [125, 126]. Figures 2.2(e) and (f), represent the waveform and spectrogram of pharyngeal stop substituted for a velar stop, respectively. The spectrogram shows that burst of the pharyngeal stop has low-frequency spectral prominence, and the vowel shows closely spaced F_1 and F_2.

Glottal stops: Glottal stops are produced by complete closure and abrupt release of vocal folds. Individuals with CLP produces glottal stops to compensate the effect of VPD. In [2], spectrographic characteristics of glottal stops are analyzed. The analysis showed that the burst of glottal stops has low-frequency spectral prominence [2]. In addition to that, spectral characteristics of glottal stops are almost same as that of the following vowel. Therefore, the slope of formant trajectories of glottal stops is almost zero. Xiao et al. [127] analyzed the spectral moments of glottal stops substituted for alveolar stops in CLP speech. The analysis showed that compared to alveolar stops, the glottal stops have reduced first spectral moment, increased spectral variance, and reduced kurtosis. Figures 2.2(g) and (h) show the waveform and spectrogram of glottal stop, respectively. The spectrogram of glottal stops (Figure 2.2(g)) shows flat formant transitions.

Devoicing errors: Production of unvoiced stops for the target voiced stops is referred to as devoicing error. In order to produce voiced stops, the speaker needs to maintain a positive trans-glottal pressure

(pressure across the glottis) by increasing the volume of the supra-glottal cavity. However, it may be difficult for the speaker with CLP due to the abnormalities associated with the supra-glottal cavity. Due to this unvoiced stops (/k/, /t/, /p/) get substituted for the target voiced stops (/g/, /d/, /b/).

Other errors: The categories mentioned above are the most commonly reported misarticulations in the literature [2, 47, 124]. Apart from these errors, the addition errors and semivowel substitutions for the stop consonant are also reported [47].

Automatic detection of misarticulated stops: With respect to CLP speech, most of the works presented in the literature are focused on the automatic evaluation of hypernasality from vowels [25, 128]. However, the detection of consonant production errors is rarely addressed in the literature [23, 102, 129]. These works include the analysis of stop consonants along with the fricatives and affricates. Automatic detection of consonant omission errors in CLP speech, based on the frame-difference operator is proposed in [129]. Automatic detection of nasalized, pharyngealized, and glottalized misarticulations is carried out in [23]. Authors of [23] segmented for the misarticulated consonant region using force-alignment methods. From the segmented consonant regions, MFCCs, Teager profile, and phoneme-likelihood measures are used to train binary classifiers for the desired category vs. rest of the errors. Various classifiers such as SVM, neural network, random forest, etc. are explored for the detection of consonant production errors in CLP speech [23]. The short-time energy and periodicity measures are used for the segmentation of normal and glottal stops in [102]. Further, MFCCs, wavelet packet energy, and gamma-tone filter-bank energy based features used train K-nearest neighbourhood classifier for the automatic classification of normal and glottal stops.

2.7 Summary and discussion

In the previous sections, different methods employed in the literature for the processing of normal and misarticulated stops are discussed. The review focused on the study of various segmentation methods and acoustic features proposed for the analysis of normal and misarticulated stops. HMM-based force-alignment is the most widely used approach in the literature for the segmentation of misarticulated phonemes. Boundaries estimated using force-alignment method showed around 20 ms deviation from the manual labels [57, 58]. Stop consonants are very short durational sounds with dynamically varying acoustic characteristics. Hence, HMM-based segmentation may not be appropriate for the accurate segmentation of stop consonants. From the study of existing literature

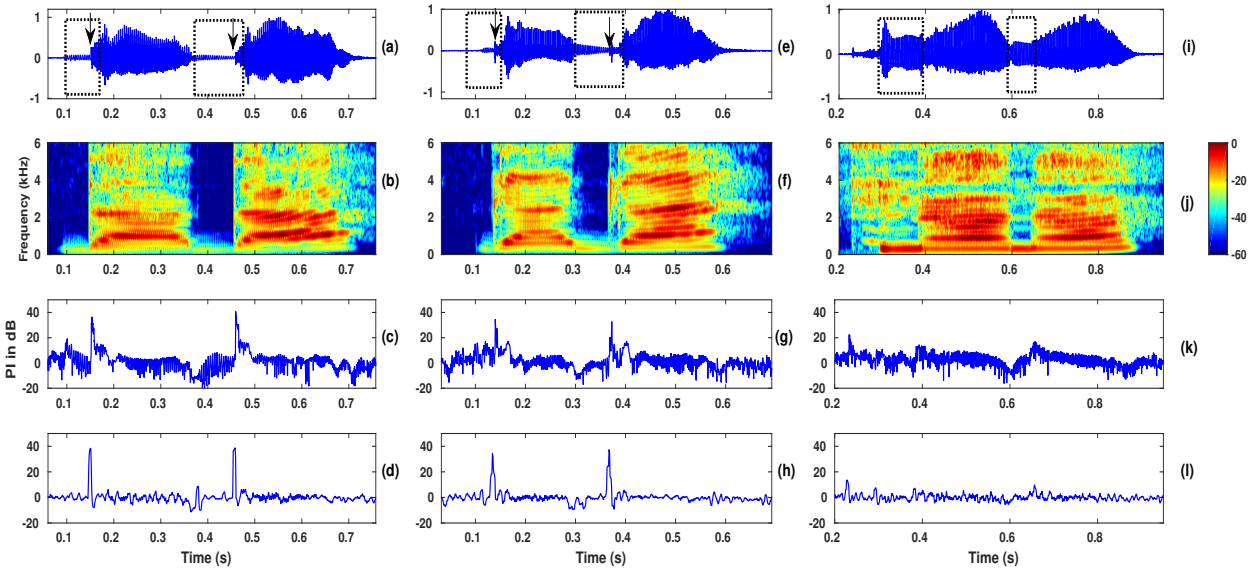

Figure 2.3: (a)-(d), (e)-(h), and (j)-(l), represent the waveform, wide-band spectrogram, plosion index (PI), and RoR for the [CVCV] words containing normal voiced, velar substitution, and nasalized stops, respectively. In subplots (a), (e), and (i), the regions corresponds to the target stop consonants are marked using rectangular box. In subplots (a) and (e), the downwards arrows indicate the locations of burst onset event.

related to normal speech, it can be observed that event-based approaches are more preferred than the statistical methods in the stop consonant recognition experiments [28, 31, 33]. In the event-based methods, acoustic features related to burst onset, VOT, and vowel onset are used for the automatic segmentation of stop consonants. The applications of event-based features are also demonstrated in the automatic detection misarticulated stops in Parkinson's disorder [10, 34]. The advantages and disadvantages of different acoustic cues for the automatic evaluation of misarticulated stops in CLP speech are discussed in this section.

Burst onset is the most widely acoustic event for the processing of normal and disordered speech. Windowed energy-difference and spectral flatness are used as the acoustic measures for the automatic detection of burst onset [28, 68]. The spectral moments of burst and VOT are used to analyze the misarticulated stops in dysarthria, CLP, apraxia, and hearing disorders [130, 131]. However, burst and VOT based analysis is possible only if the burst onset exists in the speech signal. The presence of velopharyngeal dysfunction (VPD) in CLP and dysarthria leads to the loss of airflow through the nasal tract during the articulatory closure phase [2, 48]. Due to the effect of nasalization, the burst may be absent or weakly evident in case of nasalized stops. Therefore, estimation of VOT is unreliable for the misarticulated stops in CLP speech [2]. Figure 2.3 describes the shortcomings of burst evidence

for the spotting of misarticulated stops in CLP speech. Plosion index [29] and RoR [28] features are used to analyze the evidence of burst. Figures 2.3(a)-(d), represent the speech waveform, wideband spectrogram, plosion index, and RoR measure respectively for the [CVCV] words containing voiced stop /d/ produced by a normal speaker. Spectrogram of normal stop shows a vertical striation that corresponds to the impulse-like burst onset event. Due to the presence of burst, RoR and plosion index plots (Figures 2.3(c) and (d)) indicate an increase in their magnitude around the burst onset. The waveform and spectrogram of the velar substitution error are depicted in Figures 2.3(e) and (f), respectively. The velar substitution errors produced in a similar manner of articulation as that of normal stops; however, velar substitution differs in terms of place of articulation. Therefore, the waveform and spectrogram of velar substitution error (Figures 2.3(e) and (f)) indicate the presence of burst event-like normal stops. Consequently, plosion index, and RoR evidence (Figures 2.3(g) and (h)) indicate relatively higher values around the burst onset for the case of velar substitution. The manner of articulation of nasalized stops is completely deviated from the stops. Therefore, ploison index and RoR measures shown in Figures 2.3(k) and (l) indicate the absence of burst event. Therefore, due to the absence of the burst event, it is unreliable to use the burst detection algorithms for the segmentation of misarticulated stop consonants in CLP speech.

Apart from burst, the formant transitions also carry information regarding place and manner of articulation of stops [87]. Transition regions based cues are widely used for the analysis of slow articulatory movements, and place of articulation errors in speech sound disorders [9, 102, 130]. The main advantage of analyzing the transition region is that the segmentation of the transition region does not require the knowledge of burst. Hence, transition regions can be used for the automatic evaluation of misarticulated stops in CLP, where burst is inconsistently present. The knowledge of VOP is essential for the automatic segmentation of transition regions. However, most of the existing VOP detection methods reported around 90% VOP detection rate within the temporal tolerance of ±40 ms [111–113, 115]. Timing error in the detected VOPs affects the accuracy of the segmented transition region. In [116] PRAAT based voicing decision is used for the segmentation of transition regions in unvoiced stops. In case of voiced stops, the voicing is present before VOP. Therefore, voicing features may not be useful in the detection of VOPs for the syllables containing voiced stops. Recently, the epoch-synchronous temporal feature is used for the accurate detection of voice onset event in the normal stop consonants [81]. Hence, there is a scope for the usage of epoch-synchronous features for

the accurate detection VOP.

From the literature review, it is noticed that the transition regions encode the information related to voicing, place, and manner of articulation of stop consonants [1, 42]. The misarticulated stops in CLP speech exhibit deviation in terms of voicing (devoiced stops), place of articulation (glottal, palatal, pharyngeal, and velar substitution) and manner of articulation (weak and nasalized). Therefore, analysis of transition regions will be more appropriate for the detection of misarticulated stops in CLP speech. The transition regions are analyzed using formant trajectories, dynamic spectral shape related cues, and VOP-based approaches. In most of the methods, transition regions are analyzed using the formant locations estimated by LP analysis. However, LP-based approaches may not be suitable for the children speech due to the presence of high-pitch. Moreover, CLP speech is often affected due to nasalization, due to which the speech spectrum contains zeros. The presence of spectral zeros violates the all-pole assumption of LP analysis and makes it unsuitable for the processing of transition regions. Therefore, alternative methods are required for the analysis of transition regions in CLP speech. In [40, 41, 112], MFCC feature is used for the representation of transition information. But, MFCCs may not be suitable for the analysis of transition region as they do not capture temporal dynamics explicitly. Apart from MFCCs, joint spectro-temporal cues proposed in for the analysis of stop consonants [59, 106]. Computation of joint spectro-temporal feature does not require explicit knowledge of formants frequencies [106]. The joint spectro-temporal features are used for the analysis of articulatory precision in speech disorders [52]. Therefore, there is a scope to use joint spectro-temporal features for the analysis of misarticulated stops in CLP speech.

The aforehead mentioned issues are addressed in the subsequent chapters of this thesis. In Chapter 3 an algorithm for the epoch extraction from pathological children speech is presented. The extracted epochs are used for the accurate detection of VOPs in Chapter 4. The 2D-DCT features extracted from CV transition regions anchored around VOP are used to develop normal vs. misarticulated stop and glottal vs. non-glottal stop classification systems. The effect of nasalization on voiced stops is analyzed in Chapter 5. In chapter 6, various modules developed in Chapters 3, 4, and 5, are combined to demonstrate their applications in the assessment of misarticulated stops in CLP speech. Finally, Chapter 7 summarizes the various contributions of the thesis towards the event-synchronous processing of misarticulated stops and mentions the possible future directions.

3

Epoch Extraction using Single Pole Filtering Approach

Contents

Objective

Glottal closure instant (GCI) is referred as the epoch of voiced speech. Existing epoch extraction algorithms reasonably work well for normal speech. But, the effectiveness of these methods has not been carefully analyzed for the pathological children speech. The presence of high pitch, nasality, pitch-perturbation, and glottal noise may affect the performance of state-of-the-art epoch extraction algorithms. The present work proposes an epoch extraction algorithm that considers the vertical striations present in the time-frequency representation (TFR) of the voiced speech as the representative candidates of the epochs. TFR with better localized vertical striations is estimated using single pole filter (SPF) based filter-bank approach. For the epoch extraction, the SPF-based TFR is processed in two different ways. The first approach is based on the time marginal of TFR. The second approach involves the computation of multi-scale product of SPF-based filtered envelopes. The proposed epoch extraction algorithms are evaluated on Saarbruecken Voice Database containing simultaneously recorded speech and electroglottogram (EGG) signals of pathological children. The results of the proposed algorithm are compared with the state-of-the-art algorithms in terms of reliability and accuracy.

3.1 Introduction

Vocal fold vibrations modulate the airflow from lungs and produce quasi-periodic glottal pulses, which will excite the vocal tract system to produce voiced speech. Major excitation of the vocal tract within a pitch period occurs at the GCI. This instant is also referred as the instant of significant excitation or epoch [132, 133]. In the epoch-synchronous analysis of speech, analysis frame is defined around each epoch. A detailed review of the importance of epochs in speech analysis has been presented in [134]. Applications of epoch-synchronous processing include the instantaneous fundamental frequency estimation [135], prosody modification [136], formant estimation [96], speech enhancement [137], multi-speaker separation [32], segmentation of voiced/unvoiced speech regions [138], and acoustic-phonetic segmentation of speech [27]. The knowledge of epochs is used for the detection of burst and voice onset events related to stop consonants [29, 81]. Further, in this thesis, the usefulness of epochs will be demonstrated for the detection of misaticulated stops produced by CLP children. A primary requirement for the epoch-synchronous analysis of speech is the precise detection of epochs. Different methods proposed in the literature for the automatic detection of epochs are briefly reviewed here.

3.1.1 Review of epoch extraction algorithms

Epoch extraction from speech signal is a well-defined area of research. Many of the epoch extraction methods are based on the source-filter model of speech production [139]. Hilbert envelope based methods [140, 141], Dynamic Programming Phase Slope Algorithm (DYPSA) [142], Yet Another GCI / GOI Algorithm (YAGA) [143], and Dynamic Plosion Index (DPI) [144] depend on the residual or voice source signal derived from the LP analysis of speech. Strong peaks present in the inverse filtered signal are considered as the representatives of epochs. Dynamic programming or peak picking algorithm is used to locate the epochs.

Alternative to inverse filter based approaches, algorithms which directly operate on the speech signal are also proposed [133], [35], [145]. In these methods, speech signal is filtered at the frequencies, where the frequency components are rich in excitation source information and least affected by the vocal tract resonances. In these approaches, the excitation at the epoch is assumed to be impulse-like in nature. Though the discontinuity or impulse-like energy present in the source signal at the epochs is reflected across all the frequencies, very low and high-frequency components are relatively free or less affected by the vocal tract resonances. Hilbert envelope of high-pass filtered speech is used for the epoch extraction in [133] and this approach is termed as epoch filter. An approach called Zero Frequency Resonator (ZFR) is proposed in [35] which exploits the presence of impulse-like excitation components around zero Hz. In ZFR, speech signal is passed through a cascade of two zero frequency resonators. The trend of filtered signal is removed by local average subtraction method, and the positive zero crossings of resulting signal are used to locate the epochs. Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) [145] uses a mean based signal computed by the smoothing of speech signal using Blackman-Tukey window. Zero crossings of mean based signal and peaks of the LP residual are used to locate the epochs. ZFR uses epoch information present around zero Hz, and SEDREAMS uses fundamental component. Hence, these methods are considered as smoothing or low-pass filtering methods.

Apart from LP based and smoothing approaches, methods based on the time-frequency representation of speech signal are also proposed in [146], [147], [148], where entire speech spectrum is processed for the epoch extraction. Also, these approaches do not involve inverse filtering or selection of spectral bands, free from vocal tract resonances. The line of maximum amplitude derived from wavelet-based time-frequency representation of speech is used for the epoch extraction [146]. This method fails to

extract the epochs from low energy voiced regions and also, requires *a priori* pitch knowledge and dynamic programming. Cohen's class time-frequency representation is used for the epoch extraction in [147], where the peaks present in the spectral correlation derived from time-frequency representation are claimed as epochs. Epoch extraction based on the peaks derived from the cumulation of sub-band envelopes is proposed in [148]. Further, to improve the performance of sub-band based method, LP residual and dynamic programming techniques are used [149].

3.1.2 Motivation for the present work

The existing epoch extraction methods work reasonably well for the clean speech and evaluated only on the databases containing adult speech [150]. The robustness of epoch extraction methods has been tested against degraded conditions like additive noise [35], reverberant effects [150], distant speech [137], emotional speech [151], and singing voice [152]. However, to the best of our knowledge about the existing literature, issues related to epoch extraction from the pathological children speech are not yet addressed. Children speech is characterized by the presence of high pitch. Most of the epoch extraction algorithms rely on the inverse filtered signal derived using LP analysis. The LP analysis works better for low pitched speech; however, LP analysis may not be suitable for the high pitched cases such as female and children speech [152]. In addition to high pitch, the presence of aperiodicity and glottal noise increases the non-stationarity associated with the pathological children speech. Speech of the children with CLP and motor speech disorders often affected by the hypernasality [48]. Presence of zeros in the nasalized speech spectrum causes another challenge for the epoch extraction using LP analysis [25]. Therefore, high pitch, aperiodicity, glottal noise, and nasality associated with the pathological children speech create challenges for the LP-based epoch extraction methods. Although ZFR method does not involve the inverse-filtering operation, the timing accuracy of this method is not as good as that of LP-based methods [145]. In SEDREAMS, the temporal accuracy of detected epochs is improved using LP residual. But, LP analysis may not be suitable for pathological children speech. Therefore, epoch extraction from pathological children speech is still a challenge; a reliable and accurate method is required for this.

Motivated by the significance of epochs in the speech analysis and need for the epoch extraction from the pathological children speech, the present work proposes an approach for the epoch extraction. The proposed approach has following key features.

- The vertical striations present in the wideband speech spectrogram are considered as the repre-

sentative candidates for the epochs [139].

- Estimation of time-frequency representation (TFR) of speech with better time localized vertical striations using single pole filter (SPF) based filter-bank approach.
- Processing of SPF-based TFR is carried out in two different ways for the extraction of epochs
  - In the first approach, energy present in different frequency bands is integrated by means of time marginal, which is used to locate the epochs.
  - In the second approach, enhancement of impulse-like excitation information present in the SPF-based TFR is carried out using multi-scale product. Enhanced TFR is used for the extraction of epochs.

The performance of time-marginal based method is evaluated on the database containing normal adult speech. Also, the robustness of time marginal method against the telephone quality speech is analyzed. The performance of time marginal and multi-scale product is evaluated on a database containing simultaneously recorded electroglottograph (EGG) and speech signals of pathological children.

The chapter is organized as follows. The basis for the proposed work is described in Section 3.2. Section 3.3 explains the computation of SPF-based TFR. The development of epoch extraction algorithm using time marginal, its evaluation on adult speech database, and limitations are presented in Section 3.4. Epoch extraction algorithm using multi-scale product of SPF-based filtered envelopes is explained in Section 3.5. Also, in Section 3.5, the results obtained for pathological children speech and comparison with the state-of-the-art methods are discussed. Finally, the chapter is summarized in Section 3.6.

3.2 Basis for the proposed approach

According to the source-filter theory of speech production, voiced speech is produced as a result of filtering of glottal pulses by vocal tract system. During the voiced speech production, at the GCI an impulse-like energy is delivered to the vocal tract system. The acoustic pressure signal in a free field can be modeled as the output of vocal tract system excited by the derivative of glottal pulses, the voice source signal. For speech analysis, it is a customary practice to use pre-emphasized signal. The pre-emphasized voiced speech signal can be modeled as the output of vocal tract system excited by the derivative of voice source signal (or the double derivative of glottal pulses). For chest register

type (modal) voice, the derivative of voice source (or the double derivative of glottal pulses) has an impulse-like signal structure at the epoch (at or near the GCI). This impulse-like energy will cause sharp energy transitions in the resulting speech signal [35]. Vocal tract resonances are significantly excited at the epochs, which are causal and exponentially decaying in nature. Hence, the envelope of speech signal attains a maximum value around each epoch and decays exponentially.

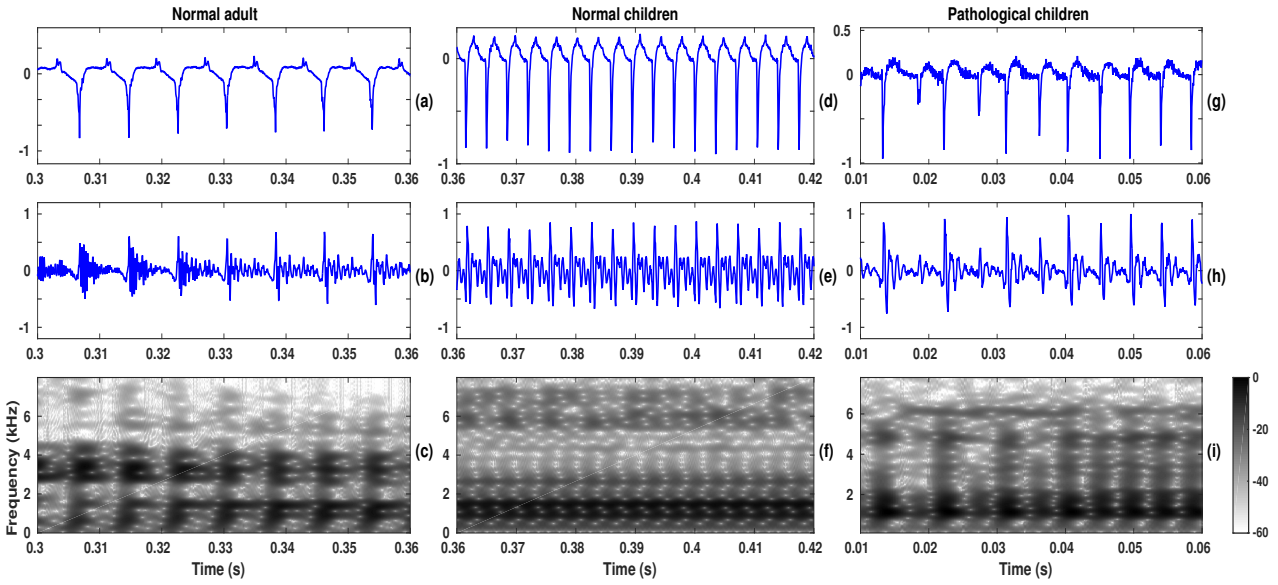

Figure 3.1: Illustration of epochal information in-terms of vertical striations of wide-band spectrogram. (a)-(c), (d)-(f), and (g)-(i) represent the differenced electroglottograph (DEGG), speech signal, and wide-band spectrogram, for normal adult, normal children, and pathological children speech, respectively.

The increase in the energy of speech signal around epochs is observed as the presence of vertical striations in the wideband speech spectrogram [139]. In the acoustic-phonetic analysis of speech, these vertical striations are widely used as the representative candidates for quasi-periodic glottal vibrations. For example, the onset of quasi-periodic vertical striations after the release of stop-burst is used to compute the voice onset time [76]. Figures 3.1(a)-(c), (d)-(f), and (g)-(i) show the differenced electroglottograph (DEGG), speech signal, and wide-band spectrogram, for normal adult, normal children, and pathological children speech, respectively. Here, short-time Fourier transform (STFT) based spectrogram is computed using a window of 5 ms with a sample shift. In the DEGG signals shown in Figures 3.1(a), (d), and (g), the large negative peaks correspond to the epoch locations. Spectrogram depicted in Figures 3.1(c), (f), and (l) indicate the presence of vertical striations, corresponding to the negative peaks of DEGG. In case of children speech, the interval between the adjacent epochs (pitch period) is relatively lesser than that of adult. Hence, due to spectro-temporal resolution issue

of STFT, the epochal information get smeared in spectrograms of children speech (Figures 3.1(f) and (h)). In addition low pitch period, the energy of excitation randomly varies from epoch-to-epoch in case of pathological children speech. Due to variation in the excitation energy, the epochs with dominant energy will mask those with weak energy, which is observed in Figure 3.1(i) at the interval of 0.01-0.03 ms.

Epochs are considered as the instantaneous property of a signal [35]. Even though vertical striations are evident in the spectrogram representation of STFT, the accurate estimation of epoch location may not be possible. Around epochs, speech signal contains sharp energy transitions. Hence, it is difficult to localize such sharp energy transitions using window based STFT. Also, in order to capture such sharp energy transitions, computation of time-frequency representation using a window of shorter duration with a sample shift becomes a difficult task. By addressing these issues, in this work, an infinite impulse response (IIR) filter-bank approach is used to estimate the time-frequency representation with better time localized vertical striations.

The STFT can be interpreted as the filter-bank summation [139], which corresponds to convolution of modulated signal $x[n]$ by $e^{-j\omega_k n}$ with window $h[n]$. That is

$$X(n,\omega_k) = x[n]e^{-j\omega_k n} * h[n] \tag{3.1}$$

For a particular value of ω_k, $|X(n,\omega_k)|$ can be interpreted as the envelope of bandpass filtered signal, where ω_k is the center frequency of the filter. Filtered envelope at ω_k represents the estimate of the amplitude of spectral envelope of the signal at ω_k.

In the literature, SPF based approaches are proposed for the computation of TFR [153], [154]. SPF is an IIR filter. A stable SPF can be implemented by placing a pole in the z-plane with pole radius (r) less than unity. The bandwidth (B) and time constant (τ) of SPF can be easily controlled by varying pole radius. SPF with pole radius r centered at frequency ω_k has a transfer function $H(z)$ is given by

$$H(z) = \frac{1}{1 - re^{j\omega_k} z^{-1}} \tag{3.2}$$

Corresponding impulse response $h[n]$ is given by

$$h[n] = (re^{j\omega_k})^n u[n] \tag{3.3}$$

The envelope of $h[n]$ is exponentially decaying in nature with a time constant (τ) for any value of the

center frequency of the filter (ω_k). The pole radius (r) can be related to bandwidth (B) and time constant (τ) of the time response of the filter as

$$|r| = e^{-\pi BT}, \quad |r| = e^{-\frac{T}{\tau}} \tag{3.4}$$

where T is the sampling interval. First order infinite impulse response filter introduces exponential smoothing and STFT computed using SPF is called as an exponentially forgetting transform (EFT) [155]. In EFT, the analysis window for the STFT is $h[-n]$, which is causal and contains a sharp front end. The sharp leading edge of EFT analysis window allows the accurate detection of signal beginning. SPF has zero phase lag at the center frequency and causal impulse response. Therefore, no delay is introduced in the signal analysis. Since the impulse response has infinite length, window length is characterized in terms of effective window length or time constant. Time-frequency uncertainty i.e., time-bandwidth product for EFT is $\frac{\pi}{2}$, whereas π for the Gaussian window of same length [155].

Apart from the above mentioned advantages, SPF has some specific advantages with respect to epoch extraction. The leading edge present in the time-reversed impulse response of SPF may give better time localization of epochs, as the epochs are characterized by sharp energy transitions. Nature of the impulse response of SPF approximately matches with that of vocal tract response for an impulse, which is also exponentially decaying in nature. Therefore, by the virtue of time localization and matched filter response, SPF-based TFR is expected to result in better localized vertical striations.

3.3 Single pole filter based time-frequency representation

Using SPF, filter-bank can be implemented by varying the center frequency (ω_k) or pole angle. Alternatively, filter-bank can also be implemented by translating the signal in frequency domain to a fixed frequency and filtering of translated signal by a filter centered at that fixed frequency. Recently, the concept of translation of desired frequency components to folding frequency ($f_s/2$) by the multiplication of signal with a complex exponential and filtering of the resultant complex signal by an SPF placed at $f_s/2$ is proposed [156]. This filter is referred as single frequency filter (SFF). The SFF has pole radius equal to 0.99 with a pole angle π corresponding to $f_s/2$, which is highly narrowband in nature. Speech and non-speech classification proposed using SFF showed better performance under different degraded conditions.

Similar to SFF, extraction of spectro-temporal envelopes from a second order highpass filter designed by placing a pair of poles at $f_s/2$ is proposed for speaker verification under reverberant environment [157]. In these studies, it has been mentioned that usage of fixed pole filter avoids the scaling effects due to different frequencies and also, the extracted envelopes are likely to be more accurate as the filtering is carried out at the highest possible frequency of the given signal. For a digital signal sampled at a sampling frequency f_s, possible highest frequency is the folding frequency i.e., $f_s/2$. Modulation of desired frequency components to $f_s/2$ and filtering at $f_s/2$ provides good separation between amplitude modulated and carrier components. Hence, extracted envelopes are likely to be more accurate.

3.3.1 Implementation of SPF-based TFR

In this work, an approach similar to SFF is used to obtain the TFR of speech for the epoch extraction [156]. The filtered envelopes obtained by the translation and filtering of the speech signal at $f_s/2$ are used for the estimation of filter-bank based TFR. SPF used in the current work is different from SFF in terms of bandwidth. In order to make SPF more suitable for the epoch extraction, the filter bandwidth is increased to 160 Hz by decreasing the pole radius. The reason for the increase in the filter bandwidth will be discussed later. SPF used in this work is not a narrowband filter like SFF. Since, SPF is a highpass filter located at $f_s/2$, estimated envelopes for very low or high frequencies will suffer from aliasing effect. In order to overcome the problem of aliasing, the complex analytic signal is computed, where the spectrum contains only positive frequencies [158]. The analytic signal is translated and filtered by SPF. The implementation procedure for SPF based filter-bank is explained as follows.

The complex analytic signal $x_a[n]$ corresponding to a discrete-time real signal $x[n]$ is computed using

$$x_a[n] = x[n] + jx_h[n] \tag{3.5}$$

where $x_h[n]$ is Hilbert transform of $x[n]$. Analytic signal, $x_a[n]$ is multiplied by a complex exponential of frequency $\bar{\omega}_k$ to shift the desired frequency components ω_k to π ($f_s/2$). The resulting operation in time domain is given by

$$x_{ak}[n] = x_a[n]e^{j\bar{\omega}_k n} \tag{3.6}$$

The above operation is equivalent in frequency domain as

$$X_{ak}(\omega) = X_a(\omega - \bar{\omega}_k) \tag{3.7}$$

where $X_{ak}(\omega)$ and $X_a(\omega)$ are spectra of $x_{ak}[n]$ and $x_a[n]$, respectively.

The frequencies ω_k and $\bar{\omega}_k$ are related by

$$\bar{\omega}_k = \pi - \omega_k \tag{3.8}$$

where ω_k is given by

$$\omega_k = \frac{2\pi f_k}{f_s} \tag{3.9}$$

where f_k is the frequency of desired spectral components need to be filtered and f_s is the sampling frequency. Hence, $\bar{\omega}_k$ can be written as

$$\bar{\omega}_k = \frac{2\pi(f_{s/2} - f_k)}{f_s} \tag{3.10}$$

The frequency translated signal $x_{ak}[n]$ is passed through SPF, whose transfer function is computed by substituting $\omega_k = \pi$ in equation 5.6, which is given by

$$H(z)|_{w_k=\pi} = \frac{1}{1 + rz^{-1}} \tag{3.11}$$

Impulse response of SPF at $f_s/2$ or π is given by

$$h[n]|_{w_k=\pi} = (-r)^n u[n] \tag{3.12}$$

where r is the radius of the pole, which is located at $z = -r$ in the z-plane with pole angle equals to π i.e., $f_s/2$. The output of the filter $y_k[n]$ is given by

$$y_k[n] = -ry_k[n-1] + x_k[n] \tag{3.13}$$

The envelope of the filtered signal, $e_k[n]$ is given by

$$e_k[n] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]} \tag{3.14}$$

where $y_{kr}[n]$ and $y_{ki}[n]$ are the real and imaginary components of $y_k[n]$, respectively. Since the filtering of $x_{ak}[n]$ is done at $f_s/2$, the envelope $e_k[n]$ corresponds to the envelope of signal $x_k[n]$ filtered around a desired center frequency f_k. Envelopes $e_k[n]$ for different frequencies can be obtained by varying f_k

from 100 to 7800 Hz in steps of 10 Hz, which will cover the most of the spectral information present in the wideband speech.

3.3.2 Time marginal of TFR

In this work, representative candidates for the epochs are the vertical striations present in the TFR. Time marginal of TFR is computed to represent the 2-D information of vertical striations in the time domain (1-D). Time marginal of TFR $S(t,\Omega)$ corresponding to a continuous time signal $s(t)$ is given by

$$\int_{-\infty}^{\infty} S(t,\Omega)d\Omega = |s(t)|^2 \tag{3.15}$$

The TFR is integrated at every instant to obtain the time marginal, which represents the instantaneous power of the signal [159]. In the current work, filtered envelopes are computed for the range of frequencies from 100 to 7800 Hz, for every 10 Hz. This operation will result in 771 filtered envelopes. Time marginal $\gamma[n]$ is given by

$$\gamma[n] = \frac{1}{N}\sum_{k=0}^{M-1} e_k^2[n] \tag{3.16}$$

where $M = 771$. The nature of time marginal depends on the local energy content of the signal and also, on the impulse response of the filter. In the present work, exponentially decaying impulse response of SPF produces an exponential smoothing effect on the filtered envelopes and hence, reduces the effect of past samples on the estimate of time marginal.

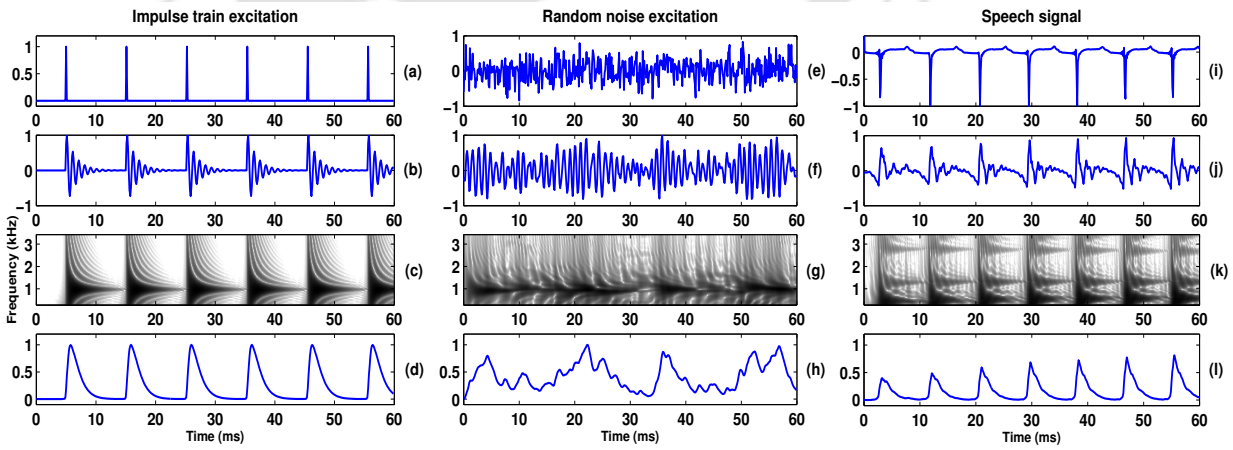

Figure 3.2: Nature of TFR and time marginal for different excitation signals. (a)-(d) impulse sequence, the output of damped sinusoid, spectrogram, and time marginal, respectively. (e)-(h) random noise sequence, the output of damped sinusoid, spectrogram, and time marginal, respectively. (i)-(l) DEGG, speech, spectrogram, and time marginal, respectively. Here, SPF based filter-bank is used to compute the spectrogram.

The nature of SPF based TFR and time marginal are analyzed for the following cases: (1) excitation of a damped sinusoid generator by impulse train, (2) excitation of damped sinusoid by random noise, and (3) speech signal. Consider, a damped sinusoid generator of 1000 Hz with a bandwidth of 100 Hz, excited by the train of impulses separated by 10 ms. Figures 3.2(a)-(d) and (e)-(h) show the excitation signal, the corresponding output of a damped sinusoid generator, spectrographic representation of SPF based filtered envelopes, and time marginal for the impulse train and random noise excitation, respectively. Figures 3.2(i)-(l) represent the DEGG, speech signal, spectrogram by SPF method, and time marginal.

From Figure 3.2 it can be observed that the vertical striations in the spectrogram are well evident in the case of excitation of the damped sinusoid generator by an impulse train. These vertical striations of spectrogram can be projected in the temporal domain effectively by the local peaks of time marginal. Such vertical striations in the spectrogram are absent for the random noise excitation case. During voiced speech production, the glottal excitation delivers an impulse-like energy at the epochs. As a result, vertical striations are well evident in the spectrogram of the speech signal (Figure 3.2(k)). Time marginal computed for the speech signal case (Figure 3.2(l)) also shows local peaks around the negative peaks of DEGG, corresponding to epochs. Another interesting point can be noticed that the vertical striations present in the SPF based spectrogram are highly localized around the epoch locations than that of STFT based approach (Figure 3.1).

3.3.3 Choice of time constant of SPF

Energy transitions at the epochs are fine temporal structures. The time constant of the impulse response of SPF plays an important role in epoch extraction. An experiment is carried out to explain the importance of time constant in epoch extraction. A speech signal segment with average pitch period (T_0) of 8 ms is passed through SPF with different time constants, τ_1=50 ms, τ_2=8 ms, τ_3=2 ms, and τ_4=0.1 ms. For each value of τ, SPF based TFR and time marginal are computed. Four different columns (from left to right) of Figure 3.3 represent the time marginal computation experiment for different time constants τ_1=50 ms, τ_2=8 ms, τ_3=2 ms, and τ_4=0.1 ms, respectively. Each column of Figure 3.3 contains the speech signal segment (Figure 3.3(a)) superimposed with reference epoch locations derived from DEGG, spectrogram representation of filtered envelopes (Figure 3.3(b)), and time marginal (Figure 3.3(c)). In this experiment, the average pitch period (T_0) or inter-epoch interval of the speech segment considered is approximately equal to 8 ms. For the higher values of τ, the impulse

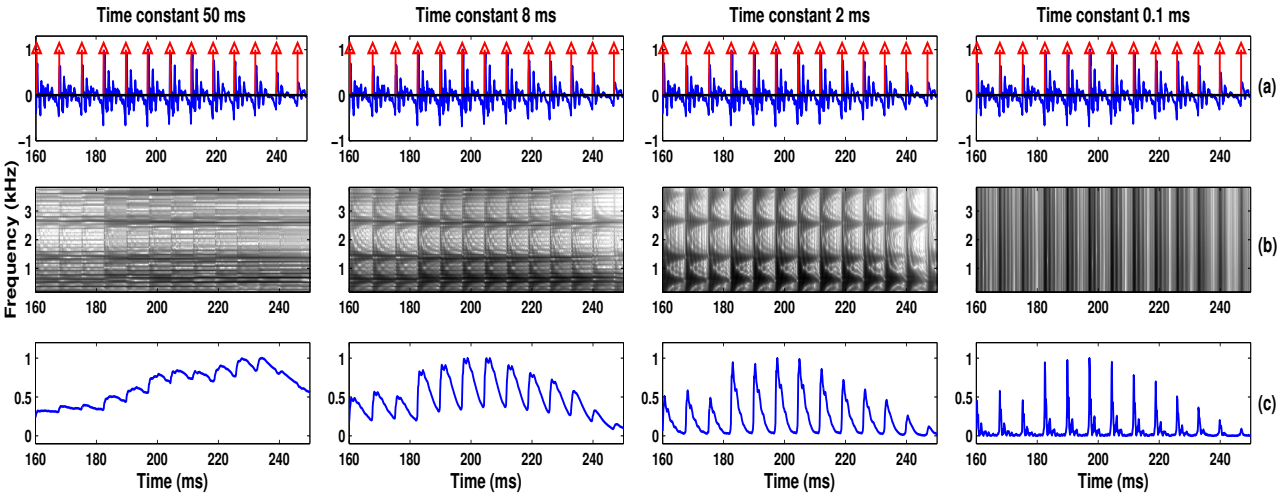

Figure 3.3: Illustration of the effect of pole radius or time constant on the representation of epoch information. In each column, (a) speech signal with reference epochs (indicated by vertical arrows), (b) Spectrogram representation of SPF based TFR, and (c) time marginal computed by the mean of filtered envelopes. First column shows for time marginal for τ=50 ms, second for τ=8 ms, third for τ=2 ms, and fourth for τ=0.1 ms. Vertical striations in the spectrogram and peaks in time marginal are highly evident for the lower values of τ.

response of the filter decays slowly and produces a smoothing effect. The vertical striations in the filter-bank based spectrogram and peaks in the time marginal are not well evident for τ_1. Vertical striations in the spectrogram and peaks in the time marginal are strongly evident for the cases of τ_2 and τ_3, where τ is less than or equal to the inter-epoch interval. The spectrogram for τ_4 contains vertical striations at epochs, but poor in spectral resolution. Time marginal for τ_4 also contains spurious peaks other than epochs and hence, adequate smoothing is required to get the epoch information.

If the decay constant of the filter is too larger than that of an inter-epoch interval, then the energy transitions in the filtered envelopes at the epochs get smoothed out. In the opposite case, where the decay constant is too smaller than that of inter-epoch interval results in TFR with poor frequency resolution. Hence, the time constant, τ plays an important role in epoch extraction and an optimum value of τ needs to be chosen. From the above experiment, it can be observed that if τ is less than that of an inter-epoch interval, then the transitions at the epochs are better resolved. This is because, if the $\tau < T_0$, then the exponential smoothing effect due to SPF emphasizes the energy at the current epoch and suppresses the energy at the previous epochs. Hence, the optimum value of τ is chosen by considering the minimum pitch period of human speech that corresponds to 2 ms (equals to highest pitch frequency of 500 Hz). The pole corresponding to τ =2 ms has bandwidth, $B \simeq 160$ Hz and radius, $r = 0.93$ for $f_s = 8000$ Hz. Similarly, τ =2 ms has radius, $r = 0.96$ for $f_s = 16000$ Hz.The

transfer characteristics of SPF designed for $f_s = 8000\ Hz$ are shown in Figure 3.4. Pole-zero plot, impulse, magnitude and phase responses of SPF are depicted in Figures 3.4(a)-(d), respectively. The gain of SPF designed for $\tau = 2\ ms$ is equal to 23.09 dB. The SPF has roll-off factor corresponding to 20 dB per decade.

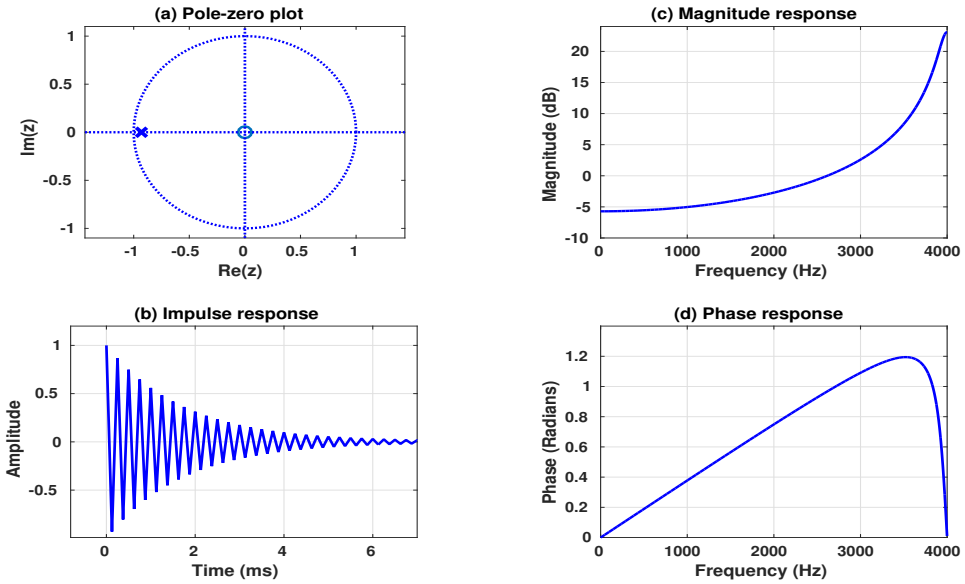

Figure 3.4: Transfer function characteristics of SPF. (a) Pole-zero plot, (b) impulse response, (c) magnitude response, and (d) phase response of the SPF. The SPF designed for the sampling frequency f_s=8000 Hz has pole radius (r)=0.93, bandwidth (B) $\simeq$160 Hz, and time constant (τ)=2 ms.

3.3.4 Other filter-bank approaches

Several bandpass filtering and envelope extraction techniques have been proposed in the literature for amplitude modulation and frequency modulation (AM-FM) analysis of speech. In order to explain the advantages of SPF over AM-FM filter-bank methods for the epoch extraction, a comparative study is carried out using existing filter-bank methods. Gabor, Hamming window based finite impulse response (HWFIR), and gammatone filter-bank methods are considered as the state-of-the-art techniques for comparison. Gabor and gammtone filter-banks are widely used in AM-FM speech analysis [160], [161]. HWFIR is used in [149] for the epoch extraction.

In this work, SPF is implemented for the bandwidth of 160 Hz. Hence, for the comparison purpose, Gabor and HWFIR filters are also implemented for the bandwidth of 160 Hz. Filtered envelopes computed from 300 to 3400 Hz in steps of 50 Hz are used obtain the TFR. Gammatone filter-bank is implemented using Auditory toolbox [161]. The impulse responses of SPF, Gabor, HWFIR filters, and

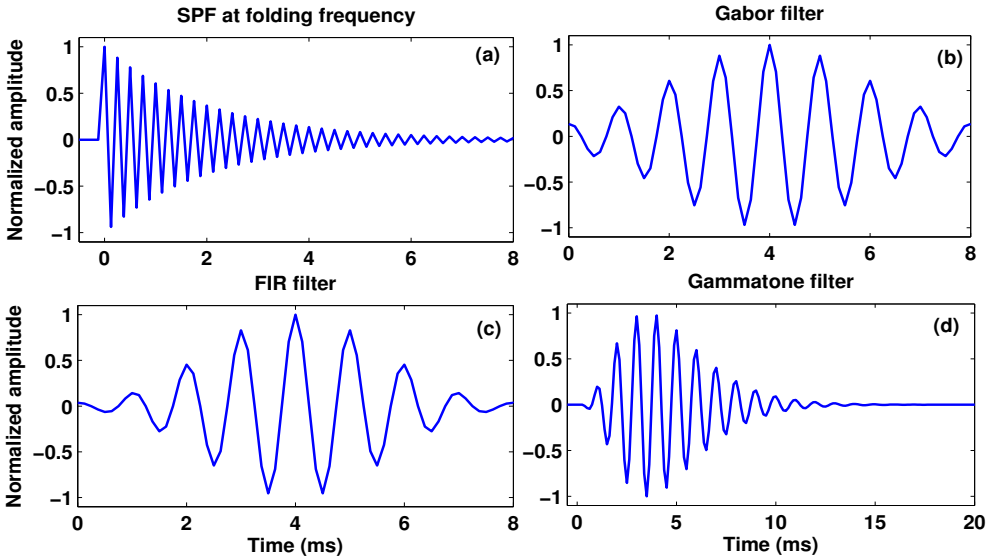

Figure 3.5: Impulse responses of (a) SPF, (b) Gabor, (c) Hamming windowed FIR and (d) Gammatone filters. The impulse response of SPF has an abrupt rise at origin followed by an exponential decay, which is desirable for the epoch extraction.

gammatone filters are shown in Figures 3.5(a), (b), (c), and (d) respectively. Gabor and HWFIR filters have symmetric impulse response whereas, SPF and gammatone have asymmetric impulse response. Figures 3.6(a) and 3.6(b) show the speech signal and spectrogram obtained by different filter-bank methods. The spectrogram obtained by SPF shows better spectro-temporal resolution compared to that of other three methods, specifically, vertical striations are highly localized. As a result, time marginal computed for SPF shows peaks with an instantaneous or abrupt increase in the amplitude as shown in Figure 3.6(c).

Though, time marginal computed using Gabor, HWFIR, and gammatone filter-bank methods show the local peaks at epoch locations, but they are not abrupt as that of SPF. Peaks with an abrupt increase in the amplitude represent the increase in the signal energy at the epochs. This attribute of epochs or significant excitation instants is better reflected in time marginal of SPF method than that of others. In order to quantify this abrupt rise in the instantaneous energy of the signal at epochs, the derivative of time marginal is computed (Figure 3.6(d)). The sharpness of peaks present in the derivative of time marginal represents the degree of abrupt increase in the instantaneous energy of the signal, which is better indicated in the SPF based method. This comparative analysis shows that TFR by SPF gives better temporal resolution than that of other methods. The presence of sharp peaks in the time marginal computed from SPF based method represents the abrupt increase in the

instantaneous energy at the epochs, which makes it more suitable for the epoch extraction.

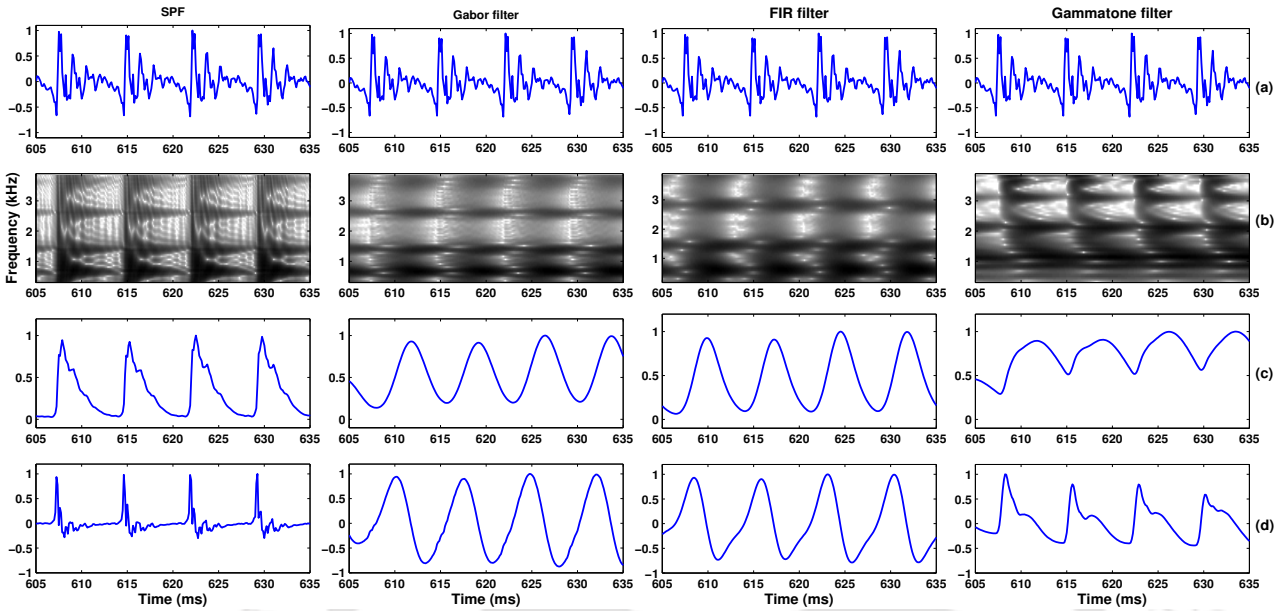

Figure 3.6: Illustration of time marginal computation using SPF, Gabor, Hamming windowed FIR, and gammatone filter-bank methods. In each column (a) speech signal, (b) TFR computed by different filter-bank methods, (c) time marginal, and (d) derivative of time marginal. The figure shows that time marginal computed by SPF method contains highly localized peaks with an abrupt increase in amplitude.

3.4 Epoch extraction using time marginal approach

Time marginal computed using SPF based TFR is considered as the preprocessed signal for the epoch extraction. A reliable method needs to be developed to locate the epochs from time marginal. Figure 3.7 represents different steps involved in the proposed epoch extraction algorithm. Speech segment and time marginal obtained from SPF are as shown in Figures 3.7(a) and (b), respectively. The time marginal obtained from SPF based TFR is further smoothed by a Gaussian window of length equal to 8 ms with a standard deviation (σ) of 1.5 ms (Figure 3.7(c)). The value of σ is chosen by considering a minimum pitch period of 2 ms. The value of σ corresponding to 1.5 ms will not produce the smoothing effect for the peaks separated by the least interval of 2 ms. The smoothed time marginal is differentiated and positive zero crossings are located as shown in the Figure 3.7(d). These positive zero crossings approximately represent the location of valleys present in the time marginal. Hence, the epoch is expected to lie within the interval of successive positive zero crossings. Further, this interval is chosen to search the epoch locations. Apart from the peak corresponding to epoch, there are some spurious peaks also present within the search interval. In order to avoid the ambiguity due to such

spurious peaks, we used the property of time marginal at the epoch i.e., the peak corresponding to epoch, possess largest peak-to-valley difference, because there is a sharp increase in the energy of signal at the epoch. Hence, within each search interval, peaks and valleys are detected by differentiating the time marginal. For each peak, valley present only at the left side is considered and the peak-to-valley difference is computed. The peak with the largest peak-to-valley difference is declared as the epoch. The detected epochs with largest peak-to-valley difference criteria along with DEGG are shown in Figure 3.7(f). This approach is referred as SPF-TM approach. The SPF-TM epoch extraction algorithm is summarized as follows.

1. Extract the TFR from SPF based filter-bank method.
2. Compute the time marginal of the TFR.
3. Smooth the time marginal using a Gaussian window of 8 ms length with σ equal to 1.5 ms.
4. Differentiate the smoothed time marginal and locate the positive zero crossings.
5. Define, the search interval as successive positive zero crossings of the signal, resulted from step (3).
6. Within each search interval locate the peaks.
7. For each peak, locate the valley immediately to the left of it.
8. Choose a peak with the largest peak-to-valley difference as the epoch.
9. Repeat the steps (5)-(8) for the entire signal.

3.4.1 Experimental results

The performance of SPF based epoch extraction method is evaluated on the databases of five speakers namely, BDL, JMK, SLT, KED, and RAB, which provide speech with simultaneously recorded EGG signal. The results are compared with 5 state-of-the-art techniques, namely ZFF, DYPSA, SEDREAMS, DPI, and YAGA. The robustness of the SPF-based algorithm for telephonic channel is evaluated on the telephone quality speech, which simulated using G.191 software [162].

Database: The BDL, JMK, and SLT corpora are from CMU ARCTIC databases. Each of these corpus consists of 1132 phonetically balanced utterances by a single speaker: BDL (U.S. male), JMK

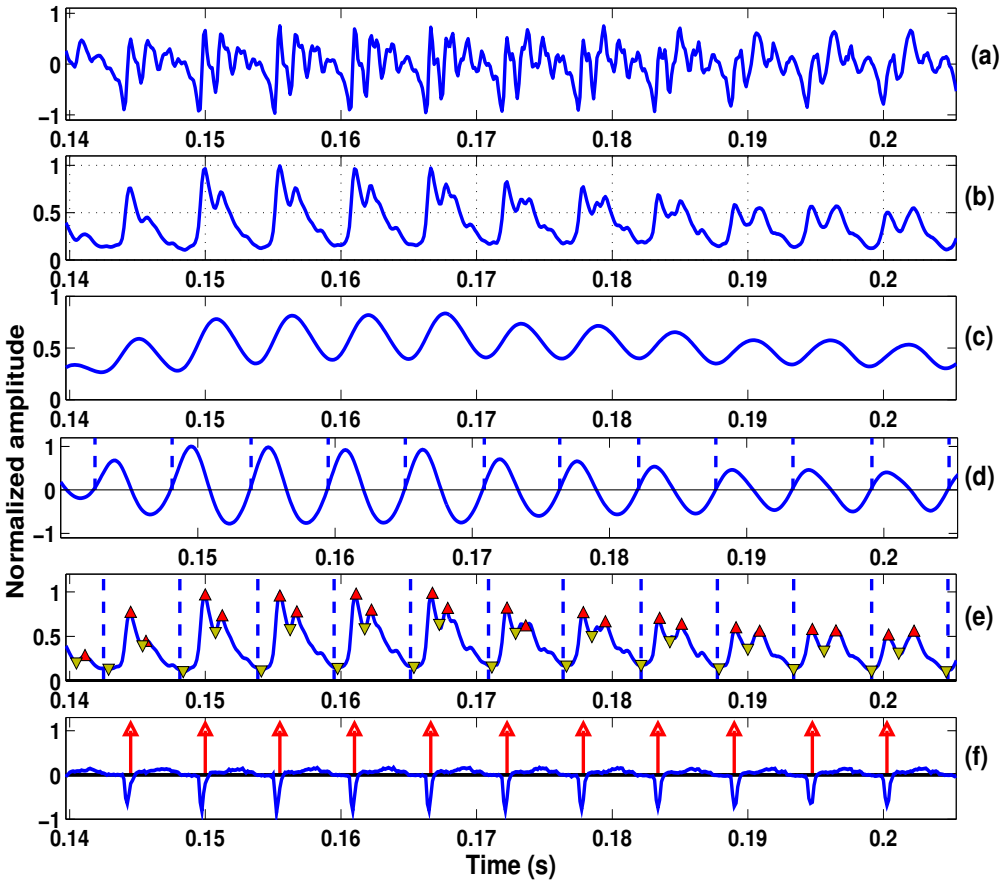

Figure 3.7: Demonstration of time marginal based epoch extration algorithms. (a) Speech signal, (b) time marginal, (c) time marginal smoothed by Gaussian window, (d) differentiated version of smoothed time marginal with positive zero crossings indicated in dotted lines, (e) time marginal indicated with peaks and valleys, and (f) Detected epochs with DEGG. Within each search interval, peak with a largest peak-to-valley difference is chosen as the epoch.

(U.S. male), and SLT (U.S. Female). The fourth speaker is KED (U.S. male), which contains 452 TIMIT sentences. The fifth speaker RAB (U.K.male) consists of a set of nonsense words containing phone-phone transitions in English. All speech samples corresponding to 5 speakers are publicly made available on Festvox web page [163].

Performance measures: In this work, identification rate (IDR), miss rate (MR), false alarm rate (FAR), identification accuracy (IDA), and accuracy to ± 0.25 ms are chosen as the performance measures, which are considered as the standard measures in recent studies [150], [144]. These measures are computed by considering DEGG as ground truth signal. Negative peaks of DEGG represent the GCI locations and considered as ground truth for the evaluation of epoch extraction methods. In this work, for a given utterance, 1/9 times the maximum negative value of DEGG is used as a threshold value to select the ground truth for GCI [144]. Such lower value of threshold considers the epochs at

Table 3.1: Performance of epoch extraction for clean and telephone quality speech

Speaker	Clean speech						Telephone quality speech				
	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	Accuracy to ±0.25 ms (%)	IDR (%)	MR (%)	FAR (%)	IDA (ms)	Accuracy to ±0.25 ms (%)
BDL	ZFR	98.02	0.92	1.06	0.29	82.93	44.95	0.96	54.09	0.88	27.89
	DYPSA	97.34	1.09	1.57	0.35	78.91	90.35	1.34	8.31	0.63	43.89
	SEDREAMS	99.37	0.41	0.22	0.31	88.35	44.81	0.83	54.36	0.71	28.91
	DPI	99.13	0.48	0.39	0.23	91.23	79.89	0.86	19.25	0.51	64.91
	YAGA	99.02	0.32	0.57	0.26	90.42	30.21	1.43	68.36	0.89	21.78
	SPF	98.45	0.83	0.72	0.37	70.71	97.89	0.96	1.15	0.65	53.87
JMK	ZFR	98.74	0.3	0.96	0.42	44.67	29.85	1.31	68.84	0.98	24.19
	DYPSA	98.46	0.42	1.12	0.39	77.54	90.16	1.7	8.14	0.47	49.89
	SEDREAMS	99.12	0.23	0.65	0.29	80.78	35.65	1.64	62.71	1.01	29.06
	DPI	99.32	0.15	0.53	0.31	88.79	86.96	1.12	11.92	0.39	61.88
	YAGA	98.58	0.35	1.07	0.32	85.67	38.63	1.9	59.47	0.89	29.98
	SPF	98.76	0.42	0.82	0.45	69.59	97.78	0.79	1.43	0.65	52.24
SLT	ZFR	99.34	0.29	0.37	0.24	81.65	67.82	0.38	31.8	0.73	40.14
	DYPSA	98.61	0.32	1.07	0.29	83.78	68.91	1.89	29.2	0.64	41.29
	SEDREAMS	95.13	1.29	3.58	0.28	88.98	65.35	2.02	32.63	0.53	50.89
	DPI	99.51	0.06	0.43	0.19	93.51	85.35	2.02	12.60	0.54	60.89
	YAGA	96.15	0.91	2.94	0.37	90.39	87.23	2.38	20.39	0.74	39.12
	SPF	98.39	1.23	0.38	0.43	70.08	97.12	2.19	0.69	0.71	61.79
KED	ZFR	98.34	0.78	0.88	0.35	69.15	41.18	4.72	54.1	1.33	20.12
	DYPSA	98.23	0.31	1.46	0.31	84.59	94.21	1.34	4.45	0.56	69.18
	SEDREAMS	95.27	1.29	3.44	0.28	91.45	23.74	0.14	76.12	0.61	20.81
	DPI	98.41	0.83	0.76	0.18	97.45	73.21	0.61	26.18	0.59	61.23
	YAGA	96.12	0.89	2.99	0.25	94.29	31.89	1.67	66.44	0.81	21.54
	SPF	98.25	1.42	0.33	0.4	74.18	97.87	1.52	0.61	0.69	65.26
RAB	ZFR	89.67	1.28	9.05	0.59	57.15	46.77	1.67	51.56	0.76	29.02
	DYPSA	81.12	0.26	18.62	0.25	77.82	65.12	1.49	33.39	0.58	50.01
	SEDREAMS	99.48	0.12	0.4	0.27	93.29	15.71	0.15	84.14	0.43	12.14
	DPI	99.45	0.14	0.41	0.19	87.56	58.56	0.19	41.25	0.96	40.25
	YAGA	97.31	0.12	2.57	0.36	86.23	7.34	0.45	92.21	0.34	6.98
	SPF	97.02	0.3	2.68	0.56	63.81	84.45	0.98	14.04	0.76	54.09

worst cases like voiced segments of lower energy, voicing onset, and offset [144].

Results on clean speech: The evaluation results of SPF and state-of-the-art techniques are presented in Table 3.1. The reliability measures: IDR, MR, and FAR for BDL, JMK, SLT, KED, and RAB for the proposed method are comparable with that of high performance state-of-the-art techniques. Compared to ZFR and SEDREAMS, accuracy measures, IDA and accuracy to ±0.25 ms are high for LP based methods (DPI and YAGA) [150]. DPI and YAGA extract the epochs from an estimate of voice source signal, whereas ZFR and SEDREAMS operate directly on the speech signal [150].

Results on telephone quality speech: The robustness of time-marginal based epoch extraction algorithm to the telephone quality speech is analyzed. In telephone quality speech, usage of a bandpass filter of bandwidth 300-3400 Hz, significantly attenuates the low frequency components (<300 Hz) rich in the excitation source information or fundamental frequency component (F_0) of speech. In addition to the effects of bandpass filter, distortions caused by channel encoder, decoder, and the addition of external noises will degrade the quality of speech [164]. To study distortions due to BPF and channel, in the current work, the telephone channel is implemented using "G.191: Software tools for speech

and audio coding standardization" [162]. This method for the simulation of telephone quality speech is used in Blizzard challenge-2009 [165], where the effect of telephone channel on the intelligibility of synthetic speech is investigated. The channel implemented in [165] is a relatively low-quality channel with a transmission rate of 16 kbps.

For the telephone quality speech, the time marginal is computed using SPF envelopes of 300 to 3400 Hz. Further, using time marginal, the The results for the telephone quality speech are mentioned in Table 3.1. With reference to the results for clean speech, the performance of state-of-the-art techniques is severely degraded for the telephone quality speech. The performance of LP-based methods, DYPSA, YAGA, and DPI methods is degraded for the telephone quality speech due the distortions caused by telephone channel. ZFR and SEDREAMS rely on the low-frequency components for the epoch extraction. Due to the usage of band-pass filter (300-3400 Hz), the performance of ZFR and SEDREAMS is severely degraded for the telephone quality speech. Whereas, the performance of proposed method remains almost same for both telephone quality and clean speech. This is because of time marginal integrates the epoch information present in the spectral band 300-3400 Hz. Though, epoch information is attenuated for the frequencies less than 300 Hz and greater than 3400 Hz, which will not affect the performance of proposed method. Because, in both clean and telephone quality speech spectrograms, the epoch information present over the frequency range of 300-3400 Hz, which is evident in terms of vertical striations. Therefore, the proposed method results in better identification rate for the telephone quality speech, which is almost equal to that of clean speech, whereas the performance of state-of-the-art methods is severely degraded.

Limitations: During speech production, vocal tract filter introduces some amount of phase delay in the responses of the individual formants. Figure 3.8 shows the SPF based filtered envelopes for 300–1200 Hz frequencies. Peaks present in the filtered envelopes are not exactly aligned with time, because of delay in the response of vocal tract filter. Time marginal is an algebraic mean of filtered envelopes. Phase delay in the vocal tract resonances may introduce some amount of shift in the peak locations of time marginal with reference to epoch locations. Hence, the accuracy measures, i.e., IDA and accuracy to ± 0.25 ms are not as good as that of LP-based methods, but comparable to that of ZFR.

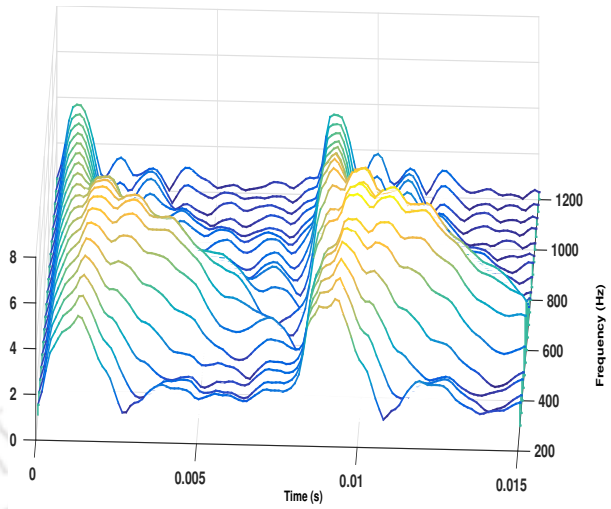

Figure 3.8: 3-D representation of filtered envelopes by SPF (300-1200 Hz). The positions of peaks in the individual envelopes are not exactly aligned with respect to time.

3.5 Multi-scale product based approach

In the previous section, the time marginal based epoch extraction algorithm showed a poor identification accuracy (delay in the epoch estimation). Recently, the combination of multi-scale product (MSP) of SFF-based envelopes and zero-frequency filtered signal is used to locate the epochs [151]. In [151], the application of MSP on SFF-based envelopes showed an enhancement of impulse-like excitation information present in the filtered envelopes. The MSP-based method showed better identification accuracy than the state-of-the-art epoch extraction methods [151]. In this section, MSP computation is carried out on the SPF-based envelopes instead of SFF. The application of MSP-based approach is demonstrated for the epoch extraction from patholgical children speech.

Computation of proposed SPF and SFF-based filtered envelopes involves a similar procedure. However, the key difference between SFF and proposed SPF approaches lies in terms of the choice of pole radius (r). The pole radius of SFF used in [151] is equal 0.99, which is almost equal to 1. For r=0.99 and f_s=16000 Hz, the time constant (τ) of SFF equals to 12.5 ms. Whereas, the values of r and τ are equal to 0.96 and 2 ms, respectively for proposed SPF-based method. The nature of SPF and SFF-based spectrograms are illustrated in Figure 3.9. The speech waveform, DEGG, SFF, and SPF based spectrograms for adult male, normal, and pathological children are depicted in Figures 3.9(a)-(d), (e)-(h), and (i)-(l), respectively. The vertical striations corresponding to epochs are better visualized in SPF based spectrogram than SFF. Especially, the vertical striations are not

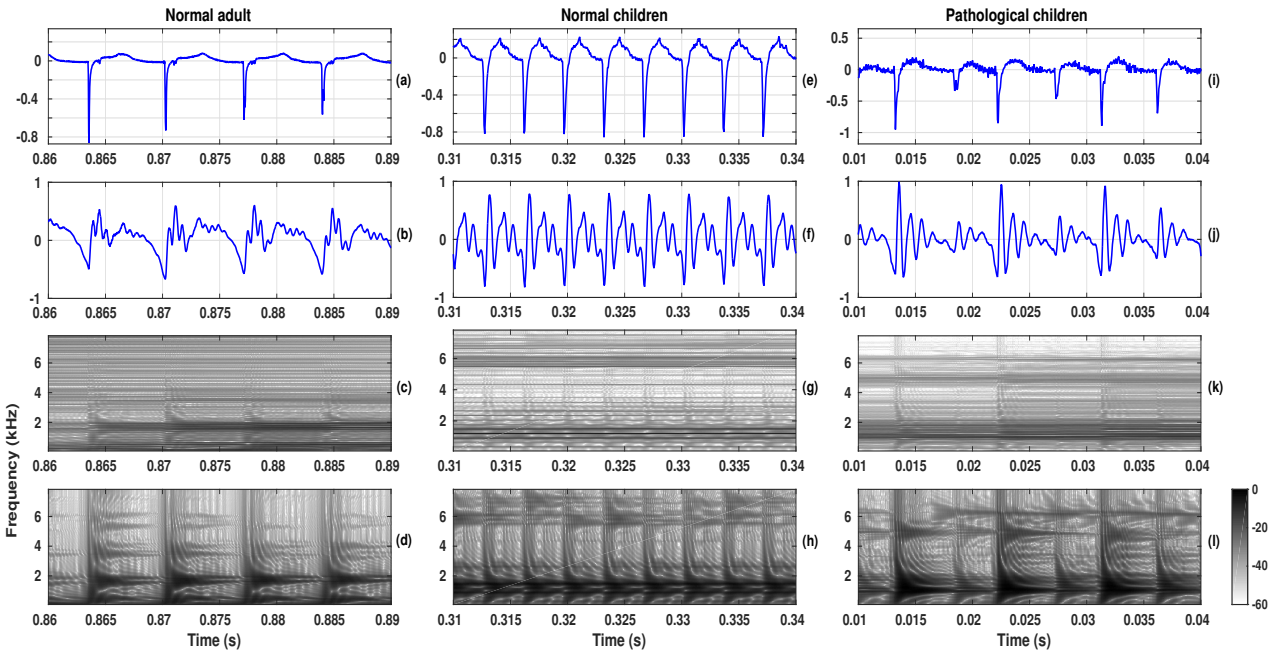

Figure 3.9: (a)-(d), (e)-(h), and (i)-(l) represent DEGG, speech, SFF spectrogram (r=0.99, τ=8 ms, f_s=16 kHz), SPF spectrogram (τ=2 ms, r=0.96 for f_s=16 kHz) for adult male, normal children, and pathological children, respectively.

prominent in SFF based spectrogram for the case of pathological children speech. This is due to the exponential smoothing effect caused by the filter. The degree of exponential smoothing is proportional to τ. Larger τ associated with SFF will smooth the sharp transitions present around the epochs. Therefore, the effect of strong excitation may mask the transitions due to weak excitations in SFF based spectrogram. But, the smoothing effect is minimized in the case of SPF due to the reduction of filter's time constant. Because, time constant of SPF is chosen equal to the minimum pitch period, which can resolve the sharp energy transitions present at consecutive epochs. Hence, impulse-like nature of epochs is better represented by the SPF based approach than SFF.

3.5.1 Computation of multi-scale product

The epoch information is better characterized in SPF based TFR than SFF. Therefore, multi-scale product based approach proposed in [151] is applied on the single pole filtered envelopes instead of single frequency filtering approach. The procedure to compute the multi-scale product of SPF based filtered envelopes is explained as follows.

The multi-scale product of stationary wavelet transform is used enhance the discontinuity infor-

mation present in the signal [143]. SWT of signal $u'[n]$, $1 \leq n \leq M$ at scale j is given by

$$d_j^s[n] = \sum_k g_j[k] a_{j-1}^s[n-k] \tag{3.17}$$

where $j = 1, 2, 3, ..., J-1$ and the maximum scale J is bounded by $log_2 M$. The coefficients a_j^s are given by

$$a_j^s[n] = \sum_k h_j[k] a_{j-1}^s[n-k] \tag{3.18}$$

where $a_0^s = u^{'}[n]$. The $g_j[k]$ and $h_j[k]$ are detail and approximation filters, respectively. The multi-scale product is given by,

$$p[n] = \prod_{j=1}^{j_1} d_j[n] \tag{3.19}$$

where j_1 is chosen equal to 3 [143]. Figures 3.10(a) and (b) represent the waveform of DEGG and corresponding speech signal, respectively. Figure 3.10(c) shows the single pole filtered envelopes for 500 and 2500 Hz. The multi-scale product of these envelopes is plotted in Figure 3.10(d). There are strong negative peaks present in the waveform corresponding to multi-scale product. The peaks of multi-scale product are more sharper than those present in filtered envelopes.

The multi-scale product of filtered envelopes at the different frequencies are computed, which are stacked to form a two-dimensional (2D) representation. Figure 3.10(e) shows the 2D-representation of multi-scale product. The multi-scale product computed for k^{th} filtered envelope $e_k[n]$ is denoted as $e_{pk}[n]$. The $e_{pk}[n]$ computed at different frequencies are averaged to get the evidence $\beta[n]$. That is given by

$$\beta[n] = \frac{1}{N} \sum_{k=0}^{M-1} e_{pk}[n] \tag{3.20}$$

where M is the number of filtered envelopes. The plot of $\beta[n]$ is shown in Figure 3.10(f), which shows close resemblance with DEGG (Figure 3.10(a)). Table. 3.2 shows the normalized cross-correlation coefficients computed for DEGG vs. ILPR, and DEGG vs. $\beta[n]$ for BDL male, SLT female, and children speakers. Samples of BDL and SLT speakers are obtained from CMU arctic database [163], whereas children samples are taken from Saarbruecken database [166]. The normalized cross-correlation coefficient values in Table 3.2 indicates that the correlation values are high for $\beta[n]$, when compared to ILPR. Generally, the inverse filtered signal is believed to have high correlation with DEGG signal. However, the higher correlation values for SPF-based approach indicate that a non-LP based method also better able to characterize the glottal source information.

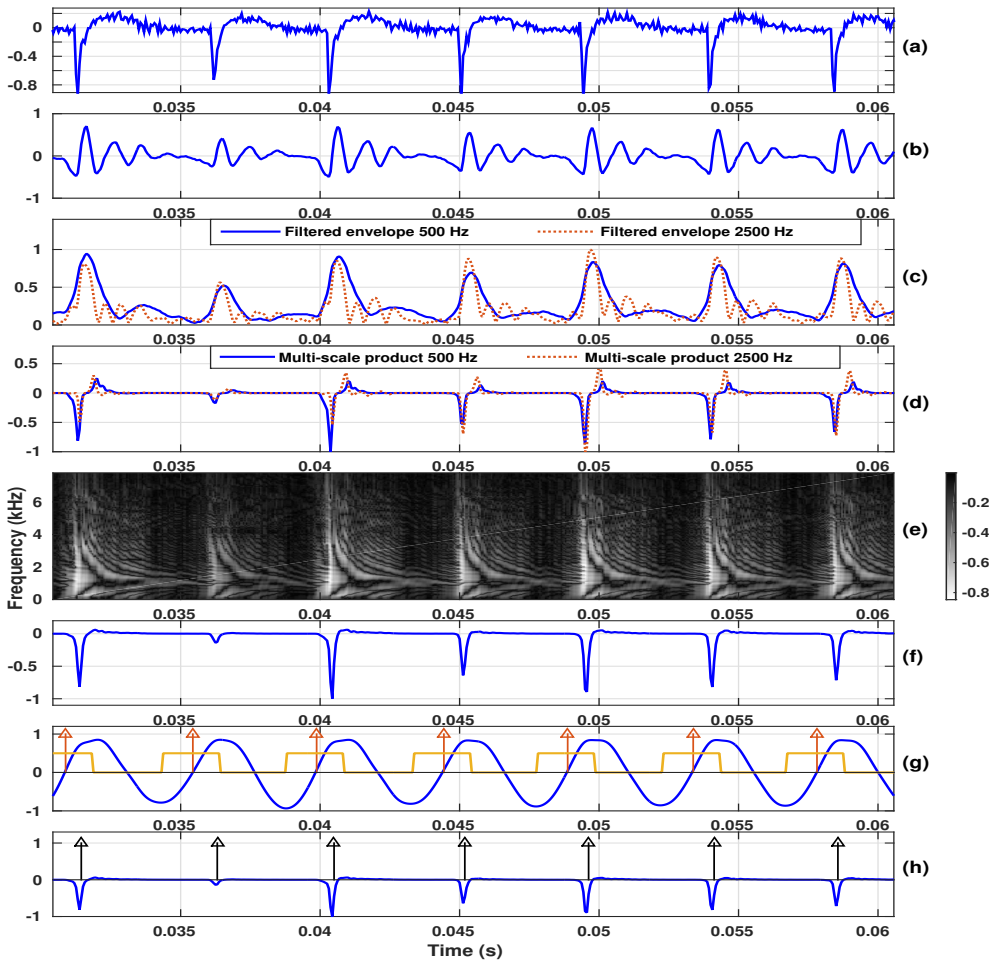

Figure 3.10: Demonstration of multi-scale product based approach for the epoch extraction. (a) DEGG, (b) speech, (c) SPF spectrogram, (c) SPF-based filtered envelopes computed at 500 and 2500 Hz, (d) multi-scale products of filtered envelopes at 500 and 2500 Hz, (e) 2-D representation of multi-scale product, (f) average of multi-scale products computed for the filtered envelopes of 100-7800 Hz, (g) ZFFS (blue color) with search interval (green color) around each positive zero crossings (red colored vertical arrows), and (h) signal of subplot (f) superimposed with the detected epoch locations. In subplot (f) detected epoch locations are indicated by black colored vertical arrows.

3.5.2 Procedure to locate the epochs

Within each glottal cycle, the strongest negative peak of $\beta[n]$ coincides with the negative peak of DEGG. Similar to epoch extraction procedure mentioned in [151], in this work, the combination of zero frequency filtered signal (ZFFS) and SPF based MSP is used to locate the epochs. First, the positive zero crossings of ZFFS are obtained. Around the positive zero crossings, a search interval of 1.2 ms is defined. Within this search interval, the location of the strongest negative peak of $\beta[n]$ is considered as the epoch location. Figure 3.10(g) shows that ZFFS along with the reference interval marked around each positive zero crossings. The detected epochs using $\beta[n]$ are shown in Figure 3.10(h). The

Table 3.2: Normalized cross correlation coefficients between source representative signals and DEGG

Method	Datasets and cross correlation values		
	Adult male (BDL)	Adult female (SLT)	Pathological children
ILPR	0.40	0.39	0.57
MSP-SPF	0.48	0.59	0.61

proposed epoch extraction approach is based on SPF and multi-scale product (MSP), which is referred SPF-MSP method.

The SPF-MSP epoch extraction algorithm is summarized as follows.

(i) Compute the TFR using SPF based filter-bank method.

(ii) Compute the MSP of each filtered envelope.

(iii) Average the MSPs computed for the filtered envelopes of frequencies ranging from 100 to 7800 Hz. The averaged MSPs is denoted as β (refer, Equation 3.20).

(iv) Compute the ZFFS and locate the positive zero crossings.

(v) Define, a search interval of 1.2 ms on either side of positive zero crossings of ZFFS.

(vi) Within the search interval, locate the strong negative peak of β. This negative peak is declared as epoch.

3.5.3 Experimental evaluation

The performance of proposed epoch extraction algorithm is evaluated on Saarbruecken pathological voice database [166]. The database consists of simultaneously recorded speech and EGG signals of adult and children speakers. The database includes the speech and EGG samples of different pathologies such as the cyst, polyp, dysphonia, etc. The recordings consist of vowels /a/, /i/, and /u/ in four different conditions: neutral, increasing, decreasing pitch, and the combination of increasing and decreasing pitch. Also the combination of three vowels (/a/, /i/, and /u/) and sentences are present in the database. In this work, 184 utterances of 13 children with voice pathologies belonging to age group of 5-15 years are considered from Saarbruecken database. The speech and EGG samples present in the database are recorded at 50000 Hz sampling rate. Further, the samples are down-sampled to 16000 Hz.

Ground truth: In the literature, the ground truth epoch locations are derived from DEGG either by peak picking method or SIGMA algorithm Ref. [167]. For pathological children speech, the amplitude of peaks of DEGG corresponding to epochs will vary dynamically. Hence, ground truth epoch locations are derived using threshold-based peak picking from DEGG may not be appropriate for pathological speech case. Automatic detection of ground truth epoch locations from EGG signal using Singularity in Electroglattograph by Multi-scale Analysis (SIGMA) algorithm is proposed in Ref. [167]. SIGMA algorithm uses dynamic programming algorithm to locate the ground truth epochs from the EGG, where the parameters are set for normal adult speakers. These parameters may not be suitable for the EGG of pathological children. In this work, a semi-automatic method is implemented to derive the ground truth epoch locations from DEGG. First, the baseline drift of EGG is removed using a butter worth high pass filter with the cut-off frequency equal 10 Hz. The high-pass filtered EGG is differenced to get DEGG. The ZFFS of DEGG is obtained, and the positive zero crossings are determined. The ground truth epoch locations are obtained by picking largest negative peak of DEGG around the positive zero crossing of ZFFS. Further, the detected epoch locations are visually verified using Wavesurfer tool [168]. Falsely identified epochs are removed, and mis detected epochs are again marked using Wavesurfer tool. The time delay between EGG and speech recording varies across speakers due to the variability in the mouth-to-microphone distance. Time delay between EGG and speech is adjusted for each utterance using the procedure mentioned in Ref. [152].

Results: The LP-based approaches: DYPSA, DPI, and YAGA, smoothing approaches: ZFR and SEDREAMS, SPF-based algorithm using time marginal (SPF-TM) and SFF-based multi-scale product approach (SFF-MSP) [151] are considered for the comparison purpose. The evaluation results are presented in Table 3.3. The LP-based methods (DYPSA, DPI, and YAGA) showed poor performance when compared to non-LP based approaches (ZFR, SEDREAMS, SPF-MSP, and SFF-MSP). This indicates that the presence of high pitch and aperiodicity in the pathological children speech affect the performance of LP-based approaches. The ZFR showed better IDR, but its performance in terms of accuracy parameters, i.e., IDA and IDR within$\pm$0.25 ms are noticed to be poor. The IDR of proposed SPF-MSP is comparable with that of ZFR, whereas IDA and IDR within$\pm$0.25 ms are better than that of ZFR. The peaks of LP residual are used to locate the epochs by SEDREAMS algorithm, where the reference interval derived using mean based signal is used to search the peaks. As the LP analysis is not reliable for high pitched children speech, SEDREAMS results in poor IDA. The

Table 3.3: Results on pathological children speech database (Saarbruecken database)

Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	Accuracy to ± 0.25 ms (%)
DYPSA	93.37	5.4	0.82	0.61	51.16
YAGA	91.69	3.64	4.25	0.9	35.46
DPI	92.64	6.23	0.72	0.57	54.06
ZFR	**96.70**	2.25	**0.63**	0.54	54.81
SEDREAMS	95.12	2.6	1.87	0.62	53.83
SPF-TM	92.49	5.01	2.09	0.74	37.66
SPF-MSP ($\tau = 2\ ms$, $r = 0.96$)	96.47	**2.18**	0.94	**0.27**	**83.17**
SFF-MSP ($\tau = 8\ ms$, $r = 0.99$)	96.42	2.18	0.98	0.42	76.97

SPF-TM algorithm resulted in poor IDA due to the delay introduced by the vocal tract resonances. In SFF-MSP, the filter is implemented with a very large time constant. The proposed SPF-MSP method outperforms the SFF-MSP in terms of accuracy parameters, i.e, IDA and accuracy to 0.25 ms. This result indicates that increase in the temporal resolution by reducing the pole radius value in SPF-MSP approach improves the temporal accuracy of detected epochs.

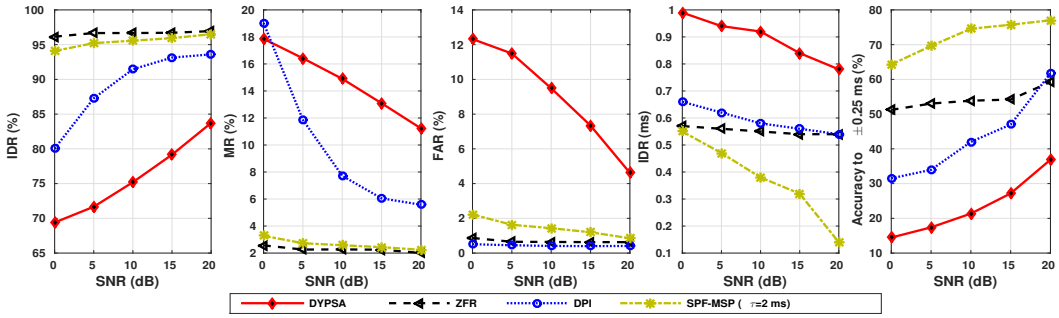

Figure 3.11: Robustness against white noise. Figure shows the performance of four different epoch extraction algorithms on pathological children speech data at different SNRs (0 to 20 dB).

Robustness to noisy conditions: The robustness of DYPSA, DPI, ZFR, and proposed SPF-MSP algorithms against the white and babble noise conditions is evaluated. The noise samples are taken from NOISEX database [169] and added to pathological children speech samples of Saarbruecken pathological voice database. Figures. 3.11 and 3.12 show the IDR, MR, FAR, IDA and IDR within ± 0.25 ms of different for the white and babble noise cases, respectively. The proposed algorithm primarily locates the epochs using ZFR. Further, ZFR based epoch locations are refined using SPF-MSP based impulse-like sequence. Hence, the proposed method shows better robustness to noisy cases, which is almost equal to ZFR. Also, under the noisy conditions, the SPF-MSP shows better

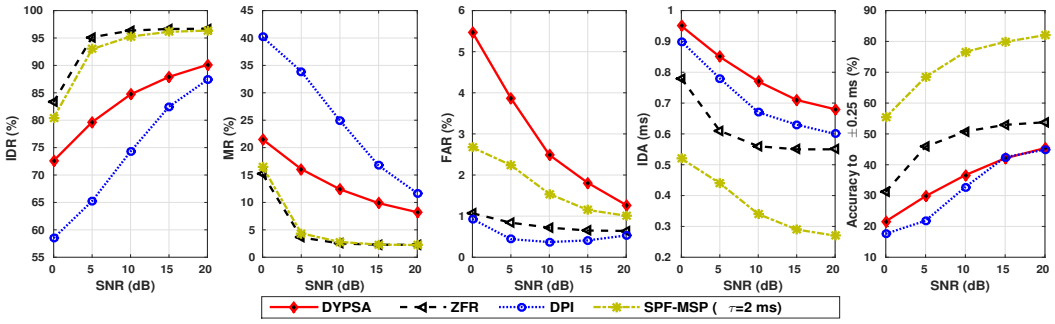

Figure 3.12: Robustness against babble noise. Figure shows the performance of four different epoch extraction algorithms on pathological children speech data at different SNRs (0 to 20 dB).

IDA and IDR within ± 0.25 ms than other methods. Thus, the combination of ZFR and SPF based approaches showed robustness to noisy cases, in-terms of both identification rate and identification accuracy.

3.6 Summary

In this chapter, an epoch extraction algorithm is proposed by considering the vertical striations present in the TFR of voiced speech are considered as the representative candidates for the epochs. To get localized vertical striations, SPF-based TFR is used. For the epoch extraction, SPF-based TFR is processed in two ways: (1) The time marginal of SPF-based TFR is computed by averaging the filtered envelopes. Sharp peaks present in the time marginal are hypothesized as epochs. (2) Multi-scale product of each filtered envelope is computed. Average of the multi-scale product of filtered envelopes at the different frequencies is calculated, where the strong negative peaks are hypothesized as epochs. The robustness of time marginal based method is analyzed on the telephone-quality speech, where the results showed higher performance for the proposed method than state-of-the-art algorithms. Further, both time marginal and multi-scale product based methods are evaluated on the database containing pathological children speech. The temporal accuracy of detected epochs by multi-scale product approach is better than the state-of-the-art methods.

In the subsequent chapters, SPF-based TFR and epoch extraction algorithm are used as the necessary blocks in the development of misarticulated stop detection system. In particular, epoch extraction algorithm is used for the detection of VOP and segmentation voiced consonant regions. The SPF-based TFR is used for the computation of spectral and spectro-temporal features for the detection of misarticulated stops.

4

Vowel onset point based processing of misarticulated stops

Contents

Objective

In this chapter, vowel onset point (VOP) based framework is proposed for the processing of misarticulated stops. The proposed VOP-based approach first detects the VOPs using an epoch-synchronously computed feature called maximum weighted inner product. The detected VOPs are considered as the anchor points to segment the CV transition regions of normal/misarticulated stops. The spectro-temporal dynamics of CV transition region anchored around VOP are analysed using two-dimensional discrete cosine transform (2D-DCT) coefficients, where the SPF-based TFR is used to compute 2D-DCT coefficients. The 2D-DCT feature is used to develop a support vector machine (SVM)-based system for the classification of normal and misarticulated stops. Here, the class of misarticulated stops comprised of weak, nasalized, palatal, velar, pharyngeal, glottal, and devoiced stops, produced by CLP speakers. The performance of the proposed VOP detection algorithm is evaluated on the database containing CV units of normal and misarticulated stops, and the results are compared with the state-of-the-art VOP detection methods. The performance of SVM classifier for the SPF-based 2D-DCT feature is compared with the STFT-based 2D-DCT and MFCC features. Further, the experiments are conducted to demonstrate the significance of VOPs in the classification of glottal vs. non-glottal misarticulations.

4.1 Introduction

Segmentation of the speaker's response corresponding to the target stop is a crucial step during the automatic detection of misarticulated stops. Conventionally, stop consonants in a continuous speech are identified based on the knowledge of burst onset event [68]. However, in case of CLP speech, burst is weakly evident or absent due to the presence of velopharyngeal dysfunction (VPD) [2]. In addition to burst onset, VOP is also considered as an important acoustic event for the detection of stop consonants [40, 41]. VOP refers to the instant at which the onset of vowel takes place in the speech signal. Speech signal anchored around VOP carries the information about the voicing, place, and manner of articulation of stops [1, 42]. In case of nasalized and weak stops, the manner of articulation of stops gets affected. In these categories, the region anchored around VOP contains the information about nasal murmur and noise due to nasal air emission (refer Figure 2.1). Burst and formant transitions anchored around VOP are considered as the important cues for the analysis of the place of articulation [170, 171]. Hence, the misarticulations resulted from the deviant place of articulation (palatal, glottal, velar, and pharyngeal) can be analyzed using the knowledge of VOP. Lack

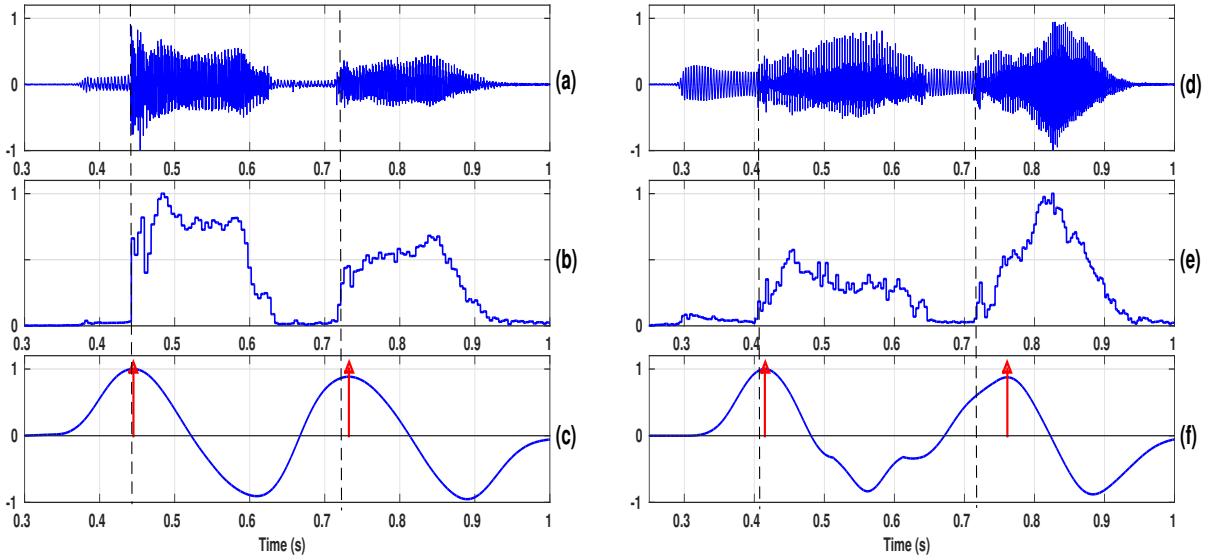

Figure 4.1: Illustration of spectral energy based VOP detection algorithm. (a)-(c) and (d)-(f) represent the speech signal, spectral energy of band 500-2500 Hz, smoothed and differentiated version of spectral energy for the [CVCV] words containing normal voiced and nasalized stops, respectively. In subplots (c) and (f), the vertical arrows indicate the detected VOPs. Black colored dashed lines represent the locations of ground truth VOP.

of pre-voicing and abrupt F_1 transition around VOP encode the information regarding the absence of voicing in devoicing errors [42]. Therefore, the CV transition regions anchored around VOP can be analyzed for the detection of misarticulated stops.

Automatic detection of VOP is the first stage in the event-synchronous processing of CV transition regions [61]. Various VOP detection approaches proposed in the literature are reviewed in Section 2.4.3 of Chapter 2. In most of the VOP detection algorithms [111, 112, 114], the evidence that highlights the vowel characteristics is derived first. Later, the evidence is smoothed and differentiated to locate the VOPs. The smoothing operation smears the abrupt energy transition around VOP and hence, affects the accuracy of the detected VOPs. Figure 4.1 illustrates the VOP detection algorithm proposed in [112]. Figures 4.1(a)-(c) represent the waveform, spectral energy of the band 500-2500 Hz, and the spectral energy resulted after the smoothing and difference operation, respectively for the [CVCV] word containing normal voiced stops. Similarly, speech waveform, spectral energy, and its smoothed and differenced version for the [CVCV] word containing nasalized stop are depicted in Figures 4.1(d)-(f), respectively. The peaks present in the smoothed and differenced version of VOP evidence are hypothesized as the VOPs, which are indicated in Figure 4.1(c) and (f). The speech waveform containing normal voiced stop shows an abrupt transition around VOP. Whereas, the CV

transition corresponding to nasalized stop is very smooth (Figure 4.1(d)). Further, the application of the smoothing operation reduces the accuracy of detected VOPs in case of nasalized stops, which is illustrated in Figure 4.1(f). Therefore, in addition to smoothing process, the presence of smooth CV transitions caused due to nasalization will pose another challenge for the accurate detection of VOP in CLP speech [172].

Apart from the accurate detection of VOP, representation of spectro-temporal dynamics associated with CV transition regions is considered as another challenge. The acoustic characteristics of CV transition regions vary abruptly with respect to time. In the literature [173], MFCCs are used for the analysis of transition regions. In the MFCC feature, the temporal dynamics of speech are represented in the form of Δ and $\Delta\Delta$ variants. Computation of MFCC involves the block-processing of speech, where it is difficult to model the spectro-temporal dynamics of transition regions due to windowing effect [174]. Therefore, (a) *accurate detection of VOP* and (b) *time-frequency trade-off* associated with conventional short-time processing techniques, are considered as the major challenges for processing CV transition regions. By addressing these issues, the current work proposes a framework for the processing of CV transition regions for the detection of misarticulated stops. The proposed approach has the following key features:

- VOP refers to the glottal closure instant (GCI) or epoch at which vowel begins. Based on this fact, an epoch-synchronous algorithm is proposed for the detection of VOP.
- Alternative to STFT, SPF-based TFR is used to derive the features from the CV transition regions. Two-dimensional discrete cosine transform (2D-DCT) coefficients computed using SPF-based TFR are used to capture the spectro-temporal dynamics of CV transitions.
- Using SPF-based 2D-DCT feature, support vector machine (SVM) classifier is developed for the screening of misarticulated stops from normal. The classifier is trained for the normal versus most possible categories of misarticulations in CLP speech, i.e., glottal, pharyngeal, palatal, velar, weak, nasalized stops and devoicing errors.
- An application of VOPs in the classification of glottal vs. non-glottal is demonstrated.

The performance of the proposed VOP detection algorithm is evaluated on the database containing normal and misarticulated stops produced by CLP children. The results are compared with the state-of-the-art VOP detection methods [111, 112, 114]. The performance SPF-based 2D-DCT is compared

with STFT based DCT feature and conventional MFCCs. Further, the chapter is organized as follows: Section 4.2 describes the database. The proposed VOP detection algorithm is presented in Section 4.3. Section 4.4 explains the computation of SPF-based 2D-DCT feature and the development of SVM classifier for the classification of normal and misarticulated stops. The experiments related to the classification of glottal vs. non-glottal misarticulations are explained in Section 4.5. Finally, the chapter is summarized in Section 4.6.

4.2 Database

The database consists of the recordings from 31 children (18 boys and 13 girls) with repaired CLP. Considered CLP children have been diagnosed at All India Institute of Speech and Hearing (AIISH), Mysore, India and reported with the presence of misarticulation. Details about the CLP speakers are described in Table 4.1. All the participants are belonging to the age range of 6-12 years and having Kannada as their mother tongue (Kannada is a Dravidian language, spoken in the southern part of India). Other associated problems like hearing loss, intellectual disability, and nasal pathologies are not found in these children.

Table 4.1: Description of CLP and CN groups

	CLP	CN
Total Number	31	30
Number of boys and girls	18, 13	16, 14
Age	9.06± 2.01	9.70±1.40

Recording material comprised of 8 non-sensical consonant-vowel-consonant-vowel ([CVCV]) words. Unvoiced velar (/k/), unvoiced dental (/t/), unvoiced retroflex (/T/), unvoiced bilabial (/p/), voiced velar (/g/), voiced dental (/d/), voiced retroflex (/D/), and voiced bilabial (/b/) stops along with the combination of low vowel (/a/) are used form [CVCV] words, i.e., [kaka], [tata], [TaTa], [papa], [gaga], [dada], [DaDa], and [baba]. During recording, participants are asked to repeat the target words spoken by the instructor. The responses of children for the target words are recorded using Bruel & Kjaer sound level meter (type 2250-s hand-held analyzer) in a sound-treated room. The recordings are obtained at a sampling frequency of 48,000 Hz and digitized at 16 bits per sample.

From each child, for each target word, 10 sessions of recordings are obtained. In total $31 \times 8 \times 10 = 2480$ words are recorded from 31 CLP children. Each [CVCV] word has two occurrences of target phoneme.

Table 4.2: Database of misarticulated stops

Target	Weak	Nasalized	Palatal	Velar	Devoiced	Pharyngeal	Glottal	Other	Total
/k/	216	26	0	0	x	46	282	0	**570**
/t/	140	64	56	192	x	0	76	18	**546**
/T/	206	60	60	134	x	0	76	36	**572**
/p/	398	44	0	16	x	0	54	14	**526**
/g/	252	116	0	0	50	0	108	0	**526**
/d/	70	140	88	92	66	0	48	22	**526**
/D/	66	150	48	96	100	0	36	20	**516**
/b/	182	194	0	14	80	0	10	24	**504**

Note: x-not applicable. For velar targets, velar substituion is not applicable. Also, for unvoiced targets, devoicing error is not applicable

Hence, 2480× 2 =4960 tokens of CV units are present in the database. The recorded words are presented to three experienced SLPs of AIISH, Mysore, India, for the perceptual evaluation. Each SLP is asked to categorize the response of CLP speaker into the various types of misarticulations mentioned in [7, 47, 175], i.e., (1) stop with weak oral pressure, (2) nasalization error, (3) glottal stop, (4) pharyngeal stop, (5) palatal stop, (6) velar substitutions, (7) devoicing error, and (8) other misarticulations (semivowel substitution, omission, and addition errors). After evaluation, only the tokens for which at least two SLPs are reported the same category are considered. Therefore, out of 4960 tokens, 4286 are retained for the study. Table 4.2 gives details about the database of misarticulated stops. In Table 4.2, the number of different categories of misarticulations produced for each target is mentioned. The table indicates that the production of a particular type of misarticulation depends on the place and voicing nature of the target stop. For example, glottal stops are more frequency produced for the target unvoiced stops than voiced stops. Also, among the different unvoiced targets, glottal stops often substituted for the velar stop (/k/).

4.2.1 Controlled normal group

There are 30 typically developed children (16 boys and 14 girls) having Kannada as their mother tongue, belonging to the same age group as that of CLP children are considered as the controlled normal (CN) group. Details about the CN speakers are provided in Table 4.1. From each child, ten repetitions of each target word are recorded. In total 30×8×10=2400 [CVCV] words, i.e., 4800 tokens of CV units are obtained from the CN group. Therefore, in the database of CN group, 600 examples of normal stops are present for each target.

4.2.2 Data labeling

The current work aims to analyze the CV transition regions around VOP. For this purpose, VOPs are manually labeled by the careful visualization of speech waveform and spectrogram using PRAAT software [117]. Labeling is carried out by a person having the knowledge of acoustic-phonetics. The onset of the first glottal cycle followed by burst is marked as VOP for the normal stops. Similarly, for the misarticulated stops, if the burst is present, then the onset of the first glottal cycle is labeled as VOP. If the burst is absent, then the onset of F_1 transition is considered as the cue for VOP. These manually labeled VOPs are considered as the ground truth for the evaluation of the performance of automatic VOP detection algorithms. To validate the reliability of VOP marking, 30% of the samples are given another person for the labeling. There is a difference of 0.8±0.01 ms obtained between two annotators.

4.3 Vowel onset point detection algorithm

The proposed VOP detection algorithm first detects the syllable nuclei using band-pass filtered speech. Next, an epoch-synchronous temporal measure called maximum weighted inner product (MWIP) is computed. The concept of MWIP is proposed originally in [81]. In this work, MWIP is computed with some modifications. Based on the position of syllable nucleus and MWIP evidence, a rule-based procedure is developed for the detection of VOP.

4.3.1 Syllable nuclei detection

In a CV unit, vowel corresponds to the syllable nucleus. Vowel is a voiced sound unit, produced due to the excitation of vocal tract system by quasi-periodic glottal vibrations. In this work, segmentation of glottal activity regions is carried out using the zero-frequency filtering approach [138]. The positive going zero-crossings of zero-frequency filtered signal (ZFFS) gives an approximate locations of epochs. First order slope of ZFFS computed at each epoch attributes to the strength of excitation (SoE) [138]. Figure 4.2(a) shows the speech signal corresponding to [CVCV] word containing nasalized stops produced by a CLP speaker. Figure 4.2(b) shows the SoE contour, where SoE values are high at voiced regions as compared to silence or unvoiced regions. Using SoE (η), glottal activity regions

are detected. The glottal activity decision (d_g) is computed as

$$d_g[n] = \begin{cases} 1 \ if \ \eta[n] > T_1 \\ 0 \ \ otherwise, \end{cases} \tag{4.1}$$

where threshold T_1 is computed as the 0.3 times of average SoE, estimated over the entire utterance. Figure 4.2(b) shows the evidence d_g superimposed over SoE contour.

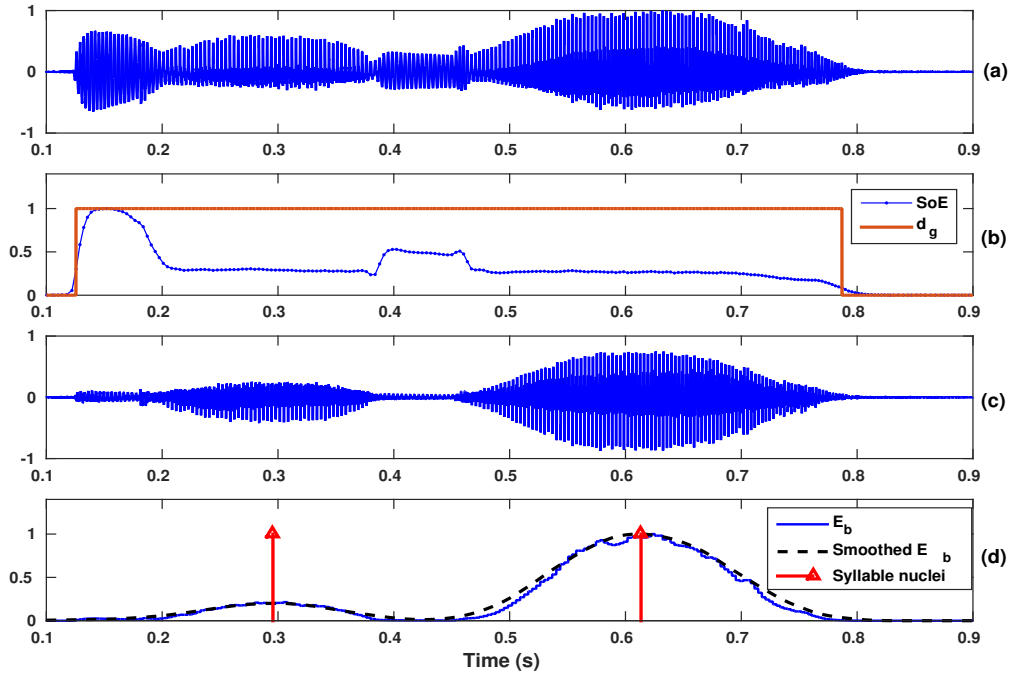

Figure 4.2: Illustration of procedure for the detection of syllable nuclei. (a) Speech waveform corresponding to [CVCV] word containing nasalized voiced stop, (b) strength of excitation (SoE) superimposed with glottal activity decision (d_g), (c) band-pass filtered signal (BPFS), and (d) short-term energy contour of BPFS (E_b), its smoothed version, and detected syllable nuclei.

To enhance the contrast between consonant and vowels, the speech signal is passed through fourth-order Butter worth band-pass filter (BPF). The BPF has pass-band frequencies in the range of 500-4000 Hz. Figure 4.2(c) shows the band-pass filtered signal (BPFS), where the energy of vowel regions get enhanced with respect to the consonant regions. Epoch-synchronous short-term energy of BPFS is computed using 20 ms durational speech frames anchored around epochs. The short-term energy contour is smoothed using Hamming window of 100 ms duration. Further, within the glottal activity regions, peaks of the smoothed energy contour are detected to get the positions of syllable nuclei. Figure 4.2(d) shows the short-time energy contour of BPFS, its smoothed version, and detected syllable nuclei.

4.3.2 Computation of maximum weighted inner product

Computation of MWIP requires the segmentation of inter-epoch intervals or glottal cycles, for which the epochs need to be located accurately. However, the epochs detected by ZFFS showed a large temporal deviation with respect to ground truth [150]. Therefore, SPF-based multi-scale product approach proposed in Chapter 3 is used to get accurate epoch locations. MWIP evidence is computed as follows: let $s_1[n]$ and $s_2[n]$ be the speech segments corresponding to adjacent inter-epoch intervals and $s_{b1}[n]$ and $s_{b2}[n]$ be their bandpass filtered versions. Lengths of $s_1[n]$ and $s_2[n]$ made equal by zero-padding. Then the weighted inner product (W) is given by

$$W_{(s_1,s_2)} = \left\langle \frac{\|s_{b1}\|_2}{\|s_1\|_2} s_1, \frac{\|s_{b2}\|_2}{\|s_2\|_2} s_2 \right\rangle \tag{4.2}$$

where $||s_1||_2$, $||s_2||_2$, $||s_{b1}||_2$, and $||s_{b2}||_2$ are the l_2-norms of s_1, s_2, s_{b1}, and s_{b2}, respectively. The BPFS required for the computation of weighted inner product is obtained using a fourth order Butterworth BPF having passband frequencies ranging from 500 to 4000 Hz. A detailed discussion about the choice of the BPF is going to be presented in Section 4.3.5. Weighted inner product values are highly sensitive to the temporal alignment of speech segments. Therefore, the weighted inner product is computed for all lags up to 0.25 ms, and the maximum value among all the lags is considered. This maximum value is referred as MWIP. The MWIP value at each epoch is padded over the entire inter-epoch interval (duration from the current epoch to the next epoch) to make the length of the evidence equal to signal length. Further, the MWIP evidence is normalized between 0 and 1. The MWIP evidence highlights two aspects of vowel, namely, periodicity and spectral energy present in the band of 500-4000 Hz. Figures 4.3(a) and (b) demonstrate the nature of MWIP for CV units containing normal unvoiced stop and nasalization error, respectively. In both cases, it is observed that the MWIP evidence have relatively higher values in the vowel regions than consonants. Unvoiced stops are the aperiodic consonants. Hence, MWIP values are low in the unvoiced stop region (Figure 4.3(a)). Even though, nasalized stops possess periodic structure, the energy of nasal murmur is relatively lower in the band of 500-4000 Hz than vowels. Therefore, MWIP shows relatively lower value for nasalized stop than vowel (Figure 4.3(a)).

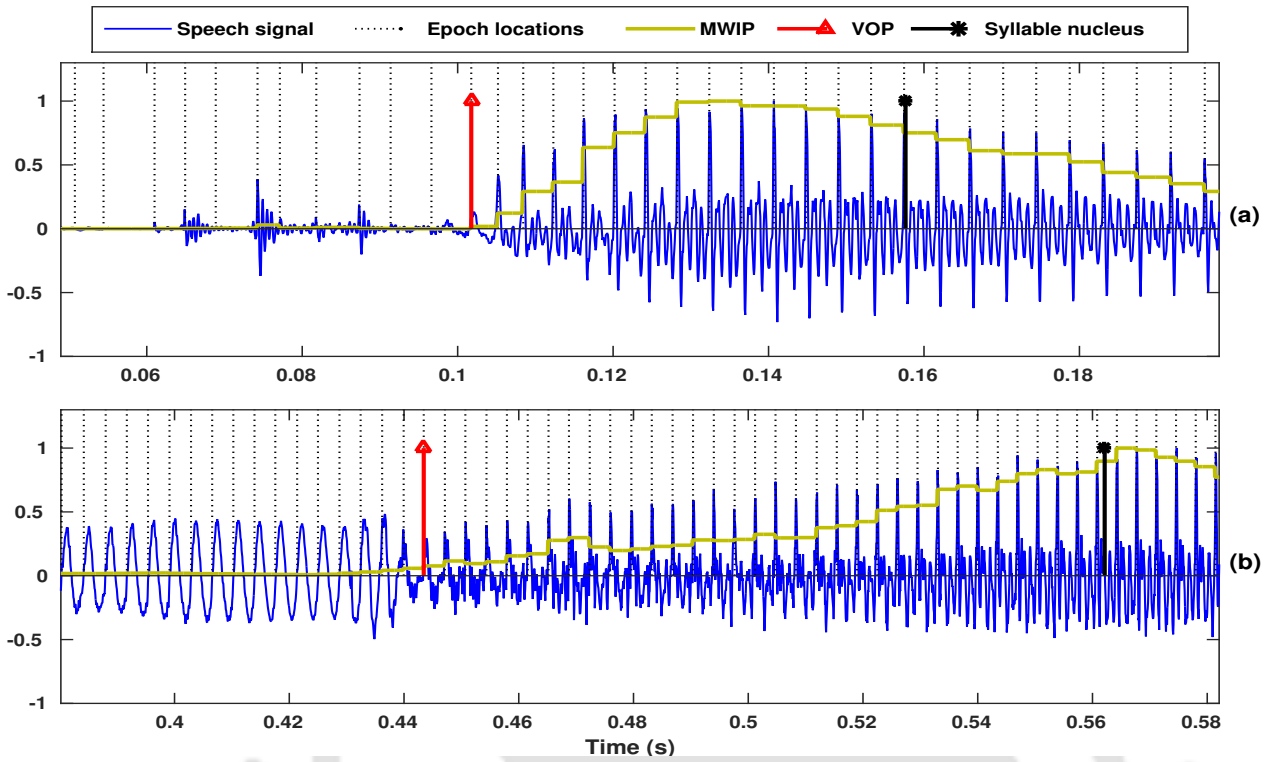

Figure 4.3: Illustration of VOP detection algorithm. (a) Waveform of CV units containing unvoiced stop produced by normal speaker superimposed with MWIP evidence. (b) Waveform of syllable containing nasalized stop produced by CLP speaker superimposed with MWIP evidence. In each case, the locations of detected epochs, syllable nucleus, and VOP are indicated over the speech waveform.

4.3.3 Detection of VOP

The MWIP evidence highlights the vowel regions present in the speech signal. Also, as indicated in Figures 4.3(a) and (b), the evidence shows a low-to-high transition around VOP location. To locate the low-to-high transition in the MWIP evidence, a rule-based algorithm developed. The algorithm involves the following steps:

1. Let g_i be the closest epoch to the detected syllable nucleus.
2. Check, whether MWIP values over two successive inter-epoch intervals at the $g_{(i)}^{th}$ and $g_{(i-1)}^{th}$ epochs are less than the T_2.
3. If the condition in step 2 becomes true, then the i^{th} epoch is chosen as the VOP.
4. If the condition in step 2 is not satisfied, then update g_i to $g_{(i-1)}$ and repeat the steps (2) and (3) until VOP is detected.

The comparison is continued up to a predefined search interval, which is defined towards the left side of syllable nucleus. In CLP speech the vowels have a longer duration than normal [172]; hence, a quite longer search interval, i.e., 200 ms is defined.

Estimation of threshold T_2: The threshold T_2 is computed as a function of MWIP value at the syllable nucleus. Let $V = \{v_1, v_2, ...v_N\}$ be the set of syllable nuclei present in an utterance. Here, N corresponds to the total number of syllable nuclei in the considered utterance. For a syllable v_i, threshold T_2 is given by

$$T_2(v_i) = \phi \times M_w(v_i),\ i \in \{1, 2, .., N\} \tag{4.3}$$

where ϕ is a constant and M_w is the MWIP feature. The value ϕ is estimated using a developmental set comprised of 50 [CVCV] words containing normal stops (100 CV units) and 50 words containing misarticulated stops (100 CV units). For all the words, vowel regions are manually marked. Within the manually marked vowel region, the position of the maximum value of BPFS energy is determined as the syllable nucleus. For each syllable of the developmental set, the ratio between the values of MWIP at the VOP and syllable nucleus is computed. Further, the distribution of these ratio values is obtained. The mode of the distribution is considered as ϕ, which is found to be equal to 0.15. Therefore, threshold (T_2) equals to 0.15 times of the MWIP value present at the syllable nucleus. Figure 4.3(a) and (b), demonstrate the detected VOPs using the syllable nucleus and MWIP evidence for the cases of CV units containing unvoiced and nasalized stops, respectively.

In [81], computation of MWIP is carried out for the estimation of voice onset time. However, the key difference between the current work and the approach proposed in [81] is regarding the choice of BPF. The current work aims to detect the VOP and computes MWIP using the BPF, which has passband frequencies in the range of 500-4000 Hz. Whereas the work proposed in [81], aims to detect the voice onset time and emphasis is given on detection of the onset of fundamental-frequency components. Hence, the BPF whose pass-band frequencies in the range of fundamental frequency is used for the estimation voice onset time [81]. To compare the approach presented in [81] with present work, MWIP is computed using the BPF of 50-500 Hz. The BPF of 50-500 Hz passes the speech components present around fundamental frequency. Figure 4.4 illustrates the nature of MWIP computed for the BPF of 50-500 Hz (MWIP_{50-500}) and 500-4000 Hz ($\text{MWIP}_{500-4000}$). For the unvoiced stop case (Figure 4.4(a) and (b)), MWIP_{50-500} shows abrupt low-to-high transition around VOP, because of the onset of voicing. However, due to the presence of low-frequency dominant voiced components,

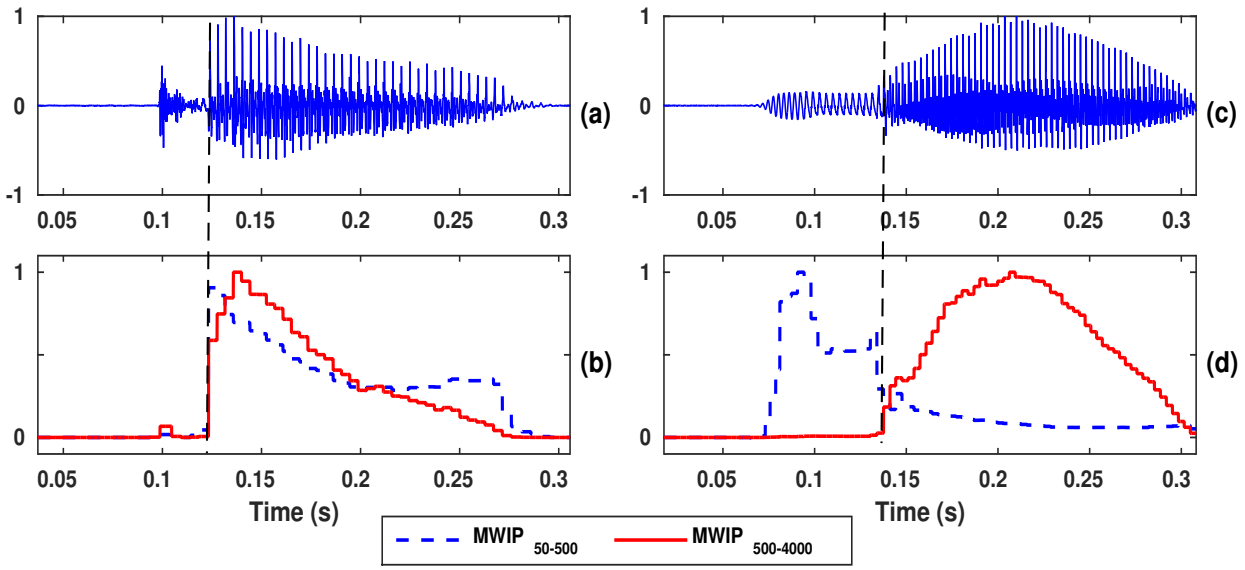

Figure 4.4: Illustration of MWIP computed using the BPFs of 50-500 Hz and 500-4000 Hz (MWIP_{50-500} and $\text{MWIP}_{500-4000}$). Speech waveforms of syllables containing (a) normal unvoiced stop and (b) nasalized stop, corresponding MWIP_{50-500} and $\text{MWIP}_{500-4000}$ are shown in sub-figures (b) and (f), respectively. In each sub-figure, dashed line (black color) indicates the ground truth VOP location.

MWIP_{50-500} does not show low-to-high transition for nasalized stop case (Figure 4.4(c) and (d)). Therefore, MWIP_{50-500} is not suitable for the VOP detection for the nasalized stop case. Vowels are periodic in nature and have strong energy in the range of 500-4000 Hz. Therefore, $\text{MWIP}_{500-4000}$ consistently shows low-to-high transition around VOP in both unvoiced and nasalized stop cases (Figure 4.4(b) and (d)). Therefore, BPF of 500-4000 Hz is more suitable for the VOP detection than BPF of 50-500 Hz.

4.3.4 Effect of BPF on VOP detection

The proposed VOP detection algorithm is based on the MWIP evidence, whose computation involves the bandpass filtering of the speech signal using a BPF of 500-4000 Hz. To verify the choice of BPF, experiments are carried out by varying the pass-band frequencies of BPF in the range of 50-500 Hz, 500-4000 Hz, 1000-4000 Hz, 500-1700 Hz, and 500-2500 Hz. Each BPF captures different attributes of the vowel. In a CV unit containing unvoiced stop, VOP is characterized by the onset of the fundamental frequency. Hence, BPF of 50-500 Hz captures the variation of the speech signal components around the fundamental frequency. Spectral band 500-4000 Hz contains the first four formants (F_1-F_4) of most of the vowels. Spectral band of 1000-4000 Hz do not contain the information about the F_1 of the vowel. Hence, BPF of 1000-4000 Hz is considered to analyze the effect of F_1 on VOP detection. Vowels are highly sonorous in nature due to the widely opened vocal tract. Spectral band

500-1700 Hz captures the sonority aspect of vowels [176]. The spectral energy computed for the band 500-2500 Hz is used for the detection of VOP in [112]. This choice of band is based on the assumption that the energy of voiced consonants lies below 500 Hz, whereas the energy of unvoiced consonants is concentrated above 2500 Hz. MWIPs computed for BPFs of 50-500 Hz, 500-4000 Hz, 1000-4000 Hz, 500-1700 Hz, and 500-2500 Hz are denoted as MWIP_{50-500}, $\text{MWIP}_{500-4000}$, $\text{MWIP}_{1000-4000}$, $\text{MWIP}_{500-1700}$, and $\text{MWIP}_{500-2500}$, respectively. For each BPF case, the threshold (T_2) is estimated using the procedure mentioned in Section 4.3.3.

Table 4.3: VOP detection rates within temporal deviation of ± 10 ms (DR_{10}) and ± 40 ms (DR_{40}) for the normal and misarticulated stops produced for the unvoiced and voiced stops.

		Normal		Misarticulated	
Target	Band (Hz)	DR_{10} (%)	DR_{40} (%)	DR_{10}%)	DR_{40}%)
Unvoiced stop	0-500	86.63	99.49	57.40	89.38
	500-4000	**95.55**	**99.87**	**79.07**	**97.44**
	1000-4000	88.31	98.78	54.036	95.90
	500-1700	94.98	98.9	64.74	96.73
	500-2500	94.27	98.98	65.07	96.98
Voiced stop	0-500	39.98	45.07	50.41	61.85
	500-4000	**87.78**	**98.33**	**77.01**	**94.48**
	1000-4000	80.48	95.67	60.06	85.68
	500-1700	92.01	97.5	76.33	90.74
	500-2500	91.97	97.62	78.09	91.68

The VOP detection rate (DR) computed within the temporal deviations of ± 10 ms and $\pm 40ms$ (DR_{10} and DR_{40}) are used to evaluate the performance of MWIP computed for different BPFs. For the evaluation, 2214 tokens of misarticulated and 2400 tokens of normal stops produced for the unvoiced targets (/k/, /t/, /T/, /p/) are considered. Similarly, for the voiced targets (/g/, /d/, /D/, /p/), 2072 tokens of misarticulated and 2400 number of normal stops are used. The performance of different BPFs for VOP detection is reported in Table 4.3. For the normal unvoiced stop case, performance of the BPF of 50-500 Hz is almost comparable with respect to others. Whereas, the presence of voicing components in normal and misarticulated voiced stops severely degrades the VOP detection performance for the band 50-500 Hz. Due to the suppression of voiced consonant regions, BPFs having lower cut-off frequency greater than 500 Hz, i.e., 500-4000, 500-1700, 500-2500, and 1000-4000 Hz are able to locate the VOPs in a better way than that of 50-500 Hz. However, the performance of 1000-4000 Hz is comparatively less due to the suppression of F_1. The DR_{40} is almost equal for the BPFs of 500-4000, 500-1700, 500-2500 Hz, but the DR_{10} is high for 500-4000 Hz. The

DR_{10} parameter indicates that the detected VOPs by BPF of 500-4000 Hz are very much closer to the ground truth VOPs than others.

4.3.5 Performance of the VOP detection algorithm

The performance of the proposed VOP detection algorithm is evaluated using the parameters: Detection rate (DR), miss rate (MR), and spurious rate (SR) [112]. These measures are computed within a predefined temporal deviation from the ground truth VOP [112]. In this work, manually labeled VOPs using PRAAT software are considered as the ground truth or reference VOPs. DR is computed as the ratio of the number of VOPs identified within a predefined temporal deviation to the total number of reference of VOPs. MR is computed as the ratio of the number of undetected VOPs within a specific temporal deviation to the total number of reference VOPs. The VOPs detected within a predefined temporal deviation from the ground truth VOP are considered as the genuine VOPs. The VOPs detected other than genuine ones are referred as spurious VOPs. The ratio of the spurious VOPs to the number of reference VOPs is termed as SR.

The VOP detection algorithms proposed in [111, 112, 114] are considered as the state-of-the-art methods. Three state-of-the-art-methods are (a) **COMB**-uses the combination of source, spectrum, and modulation spectral energies [111], (b) **SPEC**-detects VOPs using spectral energy present within the band of 500-2500 Hz [112], and (c) **ZFF-HE**-uses zero-frequency filtered signal and Hilbert envelope of LP residual [114]. The proposed and state-of-the-art VOP detection algorithms are evaluated on the CV units containing normal and misarticulated stops, and the results are reported in Table 4.4. The performance measures mentioned in Table 4.4 are calculated for the temporal deviation of 40 ms.

Table 4.4: Performance of VOP detection algorithms within temporal deviation of ± 40 *ms*

		Normal			Misarticulated		
Target	Methods	DR (%)	MR (%)	SR (%)	DR (%)	MR (%)	SR (%)
Unvoiced stop	COMB	97.1	2.9	12.9	92.46	7.53	25.51
	SPEC	97.18	2.82	5.49	85.8	14.20	24.82
	ZFF-HE	99.76	0.24	2.27	95.42	4.58	17.92
	Proposed	**99.87**	**0.13**	**1.5**	**97.44**	**2.55**	**2.62**
Voiced stop	COMB	93.4	6.6	24.13	85.37	14.62	34.76
	SPEC	92.21	7.79	12.48	79.05	20.94	29.86
	ZFF-HE	97.66	2.34	26.71	89.21	10.78	30.40
	Proposed	**98.33**	**1.67**	**2.73**	**94.48**	**5.52**	**4.56**

The detected VOPs by different algorithms for the CV units containing nasalized, and pharyngeal misarticulations are demonstrated in Figures 4.5(a) and (b), respectively. Due to the presence of

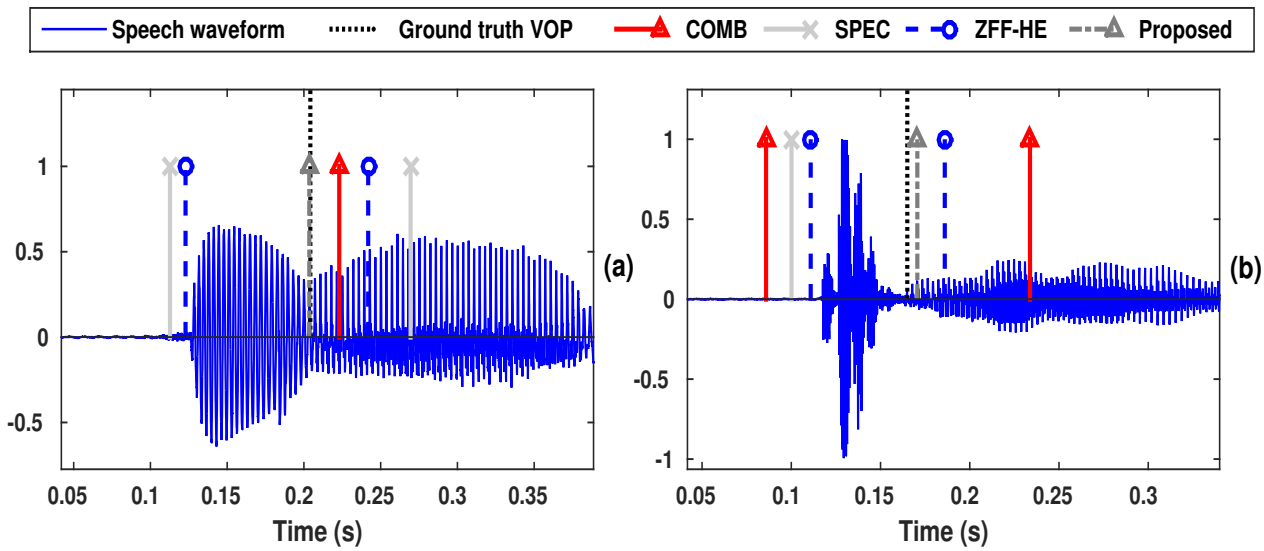

Figure 4.5: Speech waveform of syllables containing (a) nasalized and (b) pharyngeal stops. The waveforms are superimposed with ground truth VOP, detected VOPs by COMB, SPEC, ZFF-HE, and proposed algorithms.

smooth CV transition in nasalized stops, the state-of-the-art algorithms result in the falsely detected VOPs. Despite the smooth CV transition, the proposed method identifies VOPs accurately for the case of nasalized stop. The energy of burst component of pharyngeal stop is concentrated in the low-frequency region [2]. Also, the burst of pharyngeal stops has relatively high energy than normal stops. Hence, as shown in Figure 4.5(b), burst onset is falsely detected as VOP by COMB, SPEC, and ZFF-HE methods. The proposed algorithm considers the signal periodicity and band-energy (500-4000 Hz) aspects to locate the VOP. Hence, the proposed algorithm avoids the false detection of the onset of high energy burst as VOP. Also, the proposed algorithm first detects the high energy syllable nuclei and then locates the VOP. Due to this also, false detections get minimized.

Computation of MWIP does not involve smoothing operation, like state-of-the-art methods. Also, MWIP values are analyzed at every epoch to locate VOP. Hence, as shown in Figure 4.5, the VOPs detected by the proposed algorithm are very much closer to ground truth than COMB, SPEC, and ZFF-HE methods. To evaluate the temporal accuracy of the proposed VOP detection algorithm, DR is computed by varying the temporal deviation parameter from 5 to 40 ms, in steps of 5 ms. Variation of DR with respect to different temporal deviations for the normal and misarticulated stops is plotted in Figures 4.6 (a) and (b), respectively. DR is computed for the CV units containing normal and misarticulated stops produced for the target voiced and unvoiced stops. Figures 4.6(a) and (b)

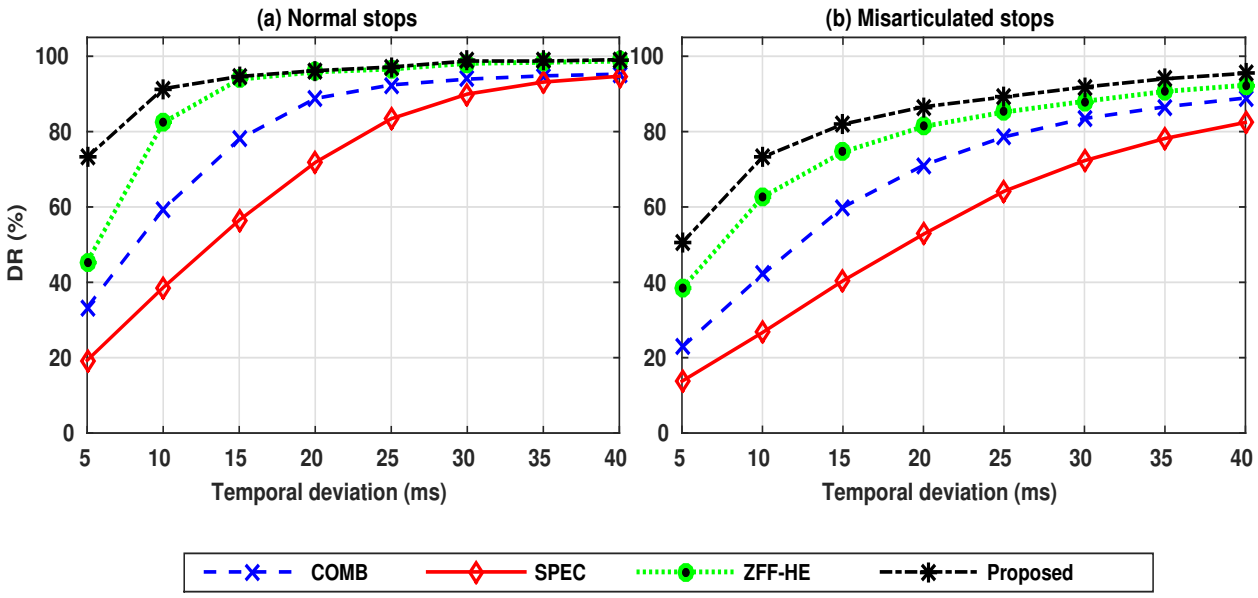

Figure 4.6: Variation of VOP detection rate (DR) with respect to temporal deviation parameter (timing difference between detected and ground truth VOP) for (a) normal and (b) misarticulated stops.

indicate that the proposed method detects around 90% and 75% of VOPs within ±10 ms temporal deviation for the normal and misarticulated cases, respectively. COMB, SPEC, and ZFF-HE methods involve the smoothing operation, due to which the identified VOPs highly deviate from the ground truth.

4.4 Classification of normal and misarticulated stops

In this section, the application of VOPs in the screening of misarticulated stops from the normal is demonstrated. The detected VOPs are used to segment the CV transition regions. From the segmented CV transition regions, joint spectro-temporal features are extracted to develop a system for the classification of normal and misarticulated stops.

4.4.1 Feature extraction

The CV transition is segmented by considering the regions on either sides of VOP. The optimum duration of the transition region is determined experimentally, which will be discussed later in Section 4.4.3. Segmentation of transition region around VOP is illustrated in Figure 4.7(a) and (d), for the CV unit containing normal unvoiced and glottal stop, respectively. Segmented transition region is analyzed using 2D-DCT based joint spectro temporal features. Conventionally, 2D-DCT coefficients

are computed using STFT [59, 106]. As discussed in the previous Chapter 3, SPF-based TFR has higher spectro-temporal resolution than conventional STFT. Figures 4.7(a)-(c) represent the speech waveform, STFT and SPF spectrograms, respectively for a CV transition region containing normal unvoiced stop. Similarly, Figure 4.7(d)-(f) represent speech waveform, STFT and SPF spectrograms, for the glottal stop case. Due to the higher time-frequency resolution associated with SPF-based TFR, spectral characteristics of burst and temporal dynamics of formants are better visualized in SPF-based spectrogram than STFT. Therefore, SPF-based TFR is used for the computation of 2D-DCT feature.

Computation of 2D-DCT features using SPF-based TFR is explained as follows. Let us consider CV transition region anchored around VOP, having a duration equal to T ms ($T/2$ ms on the left and $T/2$ ms on the right sides of VOP). The TFR is passed through 40 Mel filter-banks, spaced equally in Mel-domain to obtain Mel-TFR. Further, the logarithm is taken on Mel TFR to get log-Mel-TFR, and it is denoted as X_{Mel}. The size of X_{Mel} is equal to $M \times N$, where M and N represent the number of Mel filter-banks and number of samples corresponding to T ms, respectively. The 2D-DCT coefficients $c(k,l)$ of X_{Mel} are computed using,

$$c(k,l) = \sqrt{\frac{2}{N}}\sqrt{\frac{2}{M}}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} w(k)w(l)X_{Mel}(i,j)cos\frac{\pi k(2i+1)}{2M}cos\frac{\pi l(2j+1)}{2N} \tag{4.4}$$

where $k = 0, 1, 2, ..., K-1$, $l = 0, 1, 2, ..., L-1$, K and L are the number of 2D-DCT coefficients. $w(k)$ is given by $w(k) = \frac{1}{\sqrt{2}}$ if $k = 0$, $w(k) = 1$ if $k \neq 0$. Similarly $w(l)$ is given by $w(l) = \frac{1}{\sqrt{2}}$ if $l = 0$, $w(l) = 1$ if $l \neq 0$. In this work, K and L values are chosen equal to 9 and 3, respectively. Further, 9×3 dimensional 2D-DCT coefficients are concatenated to form 27-dimensional feature vector, which is referred as SPF-based 2D-DCT feature.

4.4.2 Classification system

In clinical practices, an SLP compares the subject's response against the target and classifiers the response as normal or misarticulated. Hence, the information about the target is necessary in screening of misarticulated stops from normal. Analogous to clinical method, the present work also uses the knowledge of target information to develop normal vs. misarticulated stop classification system. The block representation of the proposed misarticulated stop detection system is shown in Figure 4.8. As shown in Figure 4.8, there are eight SVMs are trained for normal and misarticulated stops produced for the considered eight target stops. $SVM_{/k/}$, $SVM_{/t/}$, $SVM_{/T/}$, $SVM_{/p/}$, $SVM_{/g/}$, $SVM_{/d/}$, $SVM_{/D/}$, and $SVM_{/b/}$ are the classifiers developed for targets /k/, /t/, /T/, /p/, /g/, /d/,

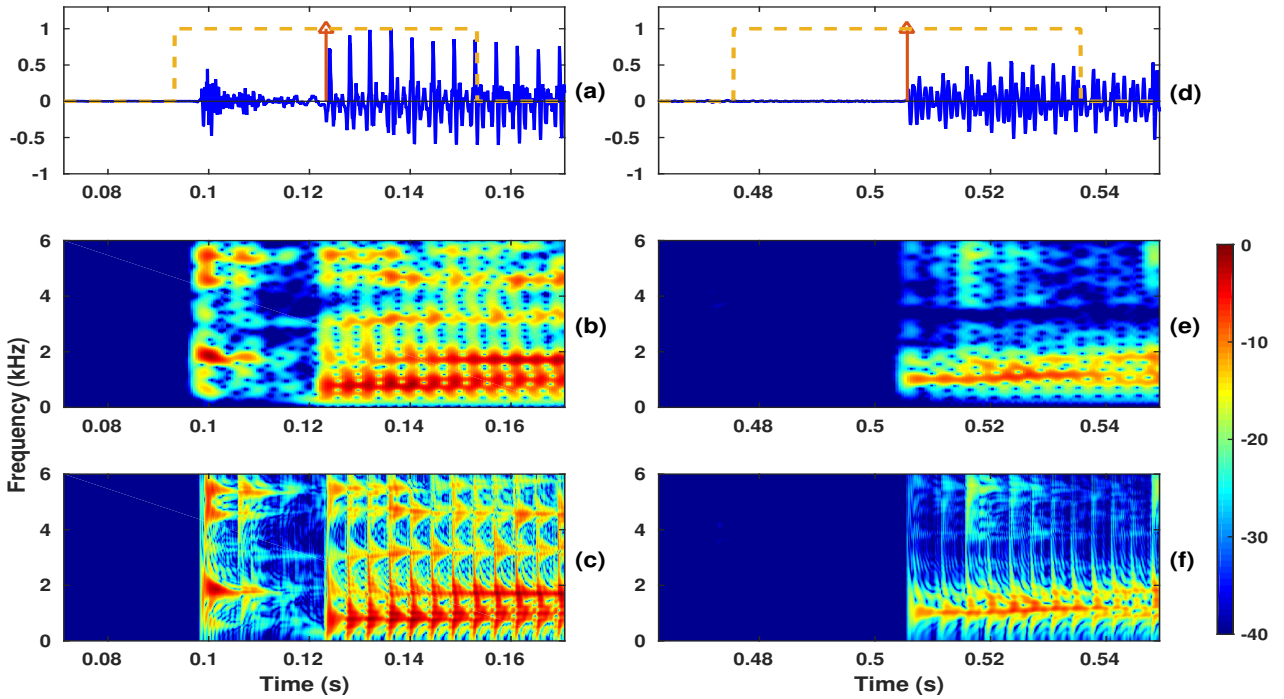

Figure 4.7: Comparison of STFT and SPF spectrograms. (a)-(c) and (d)-(f) represent the waveform, STFT, and SPF spectrograms of the CV unit containing normal unvoiced velar stop (/t/) and glottal stop, respectively. In subplots (a) and (d), the vertical arrow (red color) indicate the location of VOPs and dashed line (green color) indicate the segmented CV transition region around VOP.

/D/, and /b/, respectively.

Training phase: To develop SVM classifier for each target, the normal stops produced by CN group and misarticulated stops produced by CLP group are considered. For each target, SVM is trained using radial basis function (RBF) kernel. During the training phase of the system, manually labeled VOPs are used for the segmentation of transition region.

Testing phase: In the testing phase, automatically detected VOPs are used for the segmentation of transition regions. Further, SPF-based 2D-DCT coefficients are computed from the transition region. Based on *a priori* knowledge of target information of the test utterance, SVM corresponding to the particular target is chosen. Based on the SVM's output, the segmented transition region is decided as normal or misaritculated.

4.4.3 Experimental results

Database described in Section 4.2 is considered for the evaluation of proposed misarticulated stop detection system. The performance of SVM classifier is evaluated using the parameters: detection

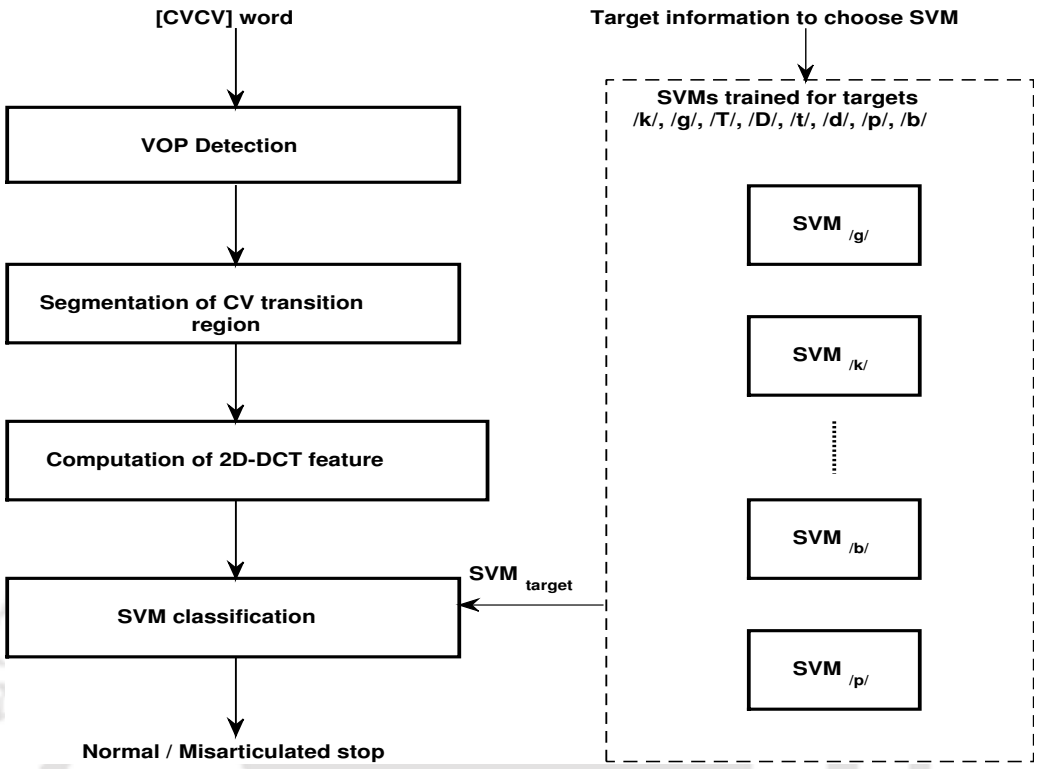

Figure 4.8: Block digram of the proposed system for the classification normal and misarticulated stops. The system contains eight SVMs trained for the normal and misarticulated stops produced for /k/, /g/, /T/, D/, /t/, d/, /p/, and /b/. During testing phase, *a priori* target information is used to select an appropriate SVM.

rate of misarticulated stops, detection rate of normal stops, and overall accuracy. These parameters as computed only for the transition regions, whose VOPs are detected within ±40 ms from the ground truth. The performance of SVM classifiers developed for each target stop is evaluated using leave one subject out (LOSO) cross-validation technique. At each fold, the optimum values of RBF kernel parameters (σ, γ) are determined using grid search over the sets $\sigma = [2^{-15}, 2^{-13}, ..., 2^{15}]$ and $\gamma = [2^{-15}, 2^{-13}, ..., 2^{3}]$ [10].

Optimum duration of transition region: Experiments are carried out to determine the optimum duration of transition region required for the classification of normal and misarticulated stops. In this experiment, 2D-DCT feature is computed using manually labeled VOPs. Duration of the transition region anchored around VOP is varied from 20 to 100 ms, in steps of 10 ms. For example, 20 ms transition region refers to the speech segment of 10 ms on the left and 10 ms on the right sides of VOP. Variation of the cross-validated accuracy of the classifier as a function of transition duration is demonstrated in Figure 4.9. From 20 to 60 ms, the classification accuracy increases as a function of the transition duration. After 60 ms the classification accuracy decreases as the duration of the

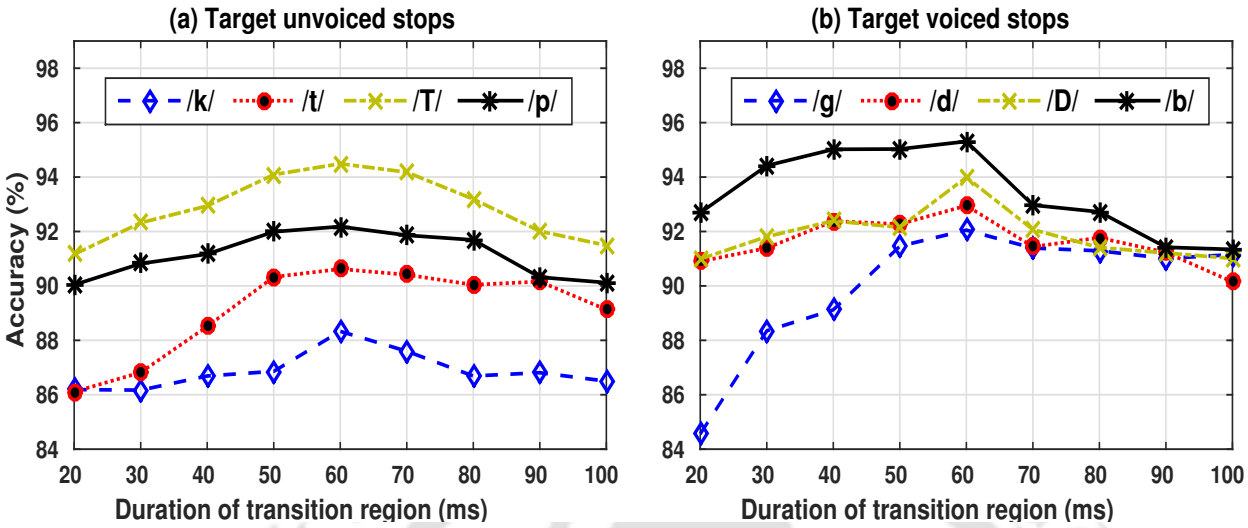

Figure 4.9: Variation of performance of SVM with respect to duration of the transition region for (a) target unvoiced stops and (b) target voiced stops.

transition region increases. Therefore, from the experimental results, 60 ms is chosen as the optimum duration of the transition region.

Performance of the misarticulated stop detection system: SVM for the automatic detection of misarticulated stops is trained by the features extracted using the manually marked VOPs. Whereas, the automatically detected VOPs are used for the testing purpose. The performance of SVM classifier for the manually labeled VOPs is considered as the reference. With respect to this reference, performance of the classifier for the automatically detected VOPs is compared. The performance of SVM for the manually labeled and automatically detected VOPs for different target stop cases are mentioned in Table 4.5. The results presented in Table 4.5 indicate that the for the automatically detected VOPs, the detection rate of misarticulated stops is almost comparable with that obtained for manual markings. Whereas, the detection rate of normal stops is significantly reduced for the automatically detected VOPs. The reason behind this behavior of classifier is explained as follows.

- With respect to normal stops, the misarticulated stops exhibit deviant CV transition characteristics around VOP. SVM-based normal vs. misarticulated stop classifier is trained in a such way that any deviation in the acoustic characteristics of transition region from the normal is categorized as misarticulated stop.

- Timing error associated with the automatically detected VOPs will alter the characteristics of segmented transition regions from the training examples of normal stops. This results in the

Table 4.5: Performance of SVM-based normal and misarticulated stop classification system developed using SPF-based 2D-DCT features. In the table misart. refers to misarticulated stops.

Target stop	Manually labeled VOP			Automatically detected VOP		
	Misart. (%)	Normal (%)	Accuracy (%)	Misart. (%)	Normal (%)	Accuracy (%)
/k/	91.26±15.33	89.51±15.27	87.54±17.39	87.44±11.26	88.79±17.91	87.53±18.39
/t/	89.61±12.44	90.84±13.99	90.05±14.19	90.6±13.95	87.99±19.63	89.36±16.77
/T/	94.07±9.98	93.28±11.88	94.09±10.55	93.92±11.32	88.89±14.09	91.5±12.86
/p/	93.46±10.60	93.23±10.59	93.53±10.44	92.25±10.26	91.33±10.56	91.78±10.33
/g/	90.49±14.26	94.68±11.11	92.69±14.11	90.79±15.84	88.58±13.89	89.97±16.34
/d/	92.36±11.71	91.33±13.99	92.11±14.81	91.52±14.08	89.86±10.63	91.16±15.44
/D/	93.08±11.68	93.91±10.89	93.14±11.98	93.36±14.43	87.71±14.06	90.89±13.95
/b/	94.09±12.92	94.77±8.97	92.89±12.53	92.61±16.03	88.96±15.21	90.85±15.06

false classification of normal stops as misarticulated and hence, reduces detection rate of normal stops.

- Similar to normal stops, temporal deviation in the automatically detected VOPs alters the acoustic characteristics of misarticulated stops also. The acoustic characteristics of misarticulated stops are already deviant from the normal due to the involvement of abnormal production mechanism. Hence, deviation induced by VOP detection error does not have a significant impact on the detection rate of misarticulated stops.

- Therefore, in development of normal vs. misarticulated stop classification system, accurate detection of VOPs is more important for the correct classification of normal stops than misarticulated ones.

Comparison with STFT and MFCCs: As discussed in Chapter 3, SPF-based TFR has better time-frequency resolution than STFT. To analyze the significance of time-frequency resolution of SPF-based TFR, the performance of normal vs. misarticulated stop classifier is evaluated for the 2D-DCT features derived using conventional STFT. The STFT is computed using the Hamming windowed speech frames of 5 ms duration with a frameshift of 1 ms. Further, 9×3 dimensional 2D-DCT coefficients are extracted from the CV transition region of 60 ms duration. The 2D-DCT feature computed using STFT is referred as STFT-based 2D-DCT feature.

Most of the misarticulated phoneme detection systems proposed in the literature are based on the MFCC features [23, 102]. Also, MFCCs are used to capture the spectro-temporal dynamics of CV transition regions [40, 112]. Therefore, the performance of SVM developed for SPF-based 2D-DCT is compared with the MFCC feature. The MFCCs are computed using speech frames of 20 ms duration with a frameshift of 5 ms. From each frame, 39-dimensional MFCC feature (13-dimensional MFCCs, and their Δ, $\Delta\Delta$ variants) is computed. From the 60 ms durational CV transition region, twelve frames of 39-dimensional MFCCs are obtained. Further, the MFCCs of 12 frames are concatenated to form a 468-dimensional feature vector. Further, the principal component analysis (PCA) based dimensionality reduction is carried out. 150-dimensional feature vector resulted after the application of PCA is used to train and test the SVM.

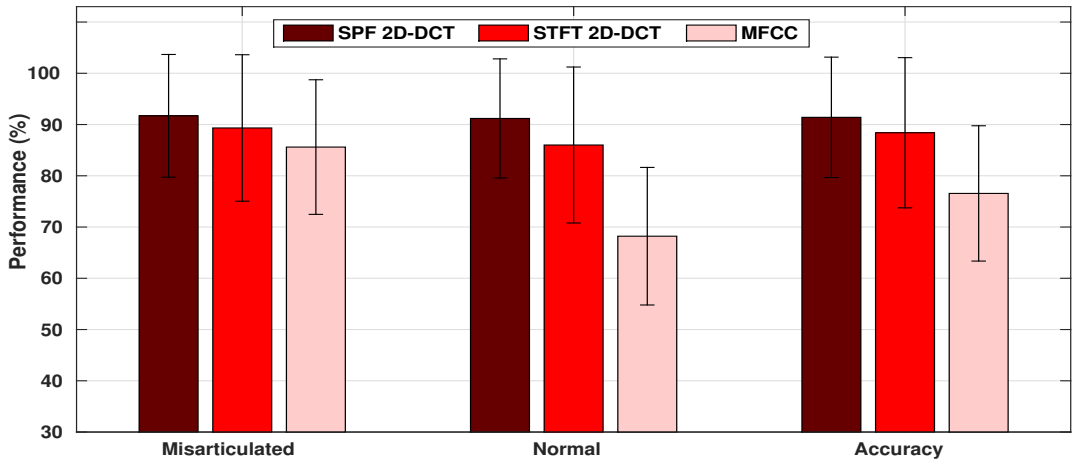

Figure 4.10: Comparison of performance of SVM for SPF-based 2D-DCT, STFT-based 2D-DCT, and MFCC features.

The performance of classifier is sensitive to both VOP locations and features. Therefore, for the comparison of different features, SVM is trained and tested using manually marked VOPs. The usage of manual labels reduces the influence of segmentation error on the SVM performance and helps to analyze only the significance of features. The bar graph shown in Figure 4.10 represents the performance of SVM for SPF-based 2D-DCT, STFT-based 2D-DCT, and MFCC features, where the performance is averaged over eight target stops. Figure 4.10 indicates that 2D-DCT-based joint spectro-temporal features give better performance than MFCCs. This result indicates that MFCCs poorly capture the spectro-temporal dynamics of transition regions. The SPF-based TFR has better spectro-temporal resolution than STFT. Hence, between SPF and STFT-based 2D-DCT features, the

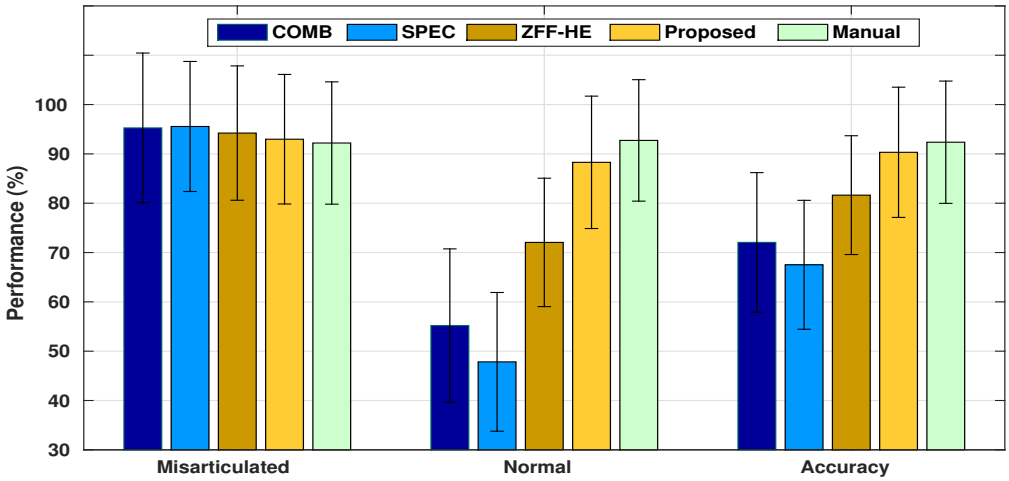

Figure 4.11: The performance of the normal vs. misarticulated stop classifier for different VOP detection algorithms. SVM is trained for the features extracted using manually labeled VOPs, whereas automatically detected VOPs are used during testing.

SPF shows better performance (Figure 4.10). This improvement signifies the importance of spectro-temporal resolution of TFR in the analysis of transition regions.

Comparison of different VOP detection algorithms: VOP plays a crucial role in the segmentation of CV transitions. Due to epoch-synchronous processing, the VOPs detected by the proposed algorithm is very much closer to ground truth VOPs than the state-of-the-art algorithms (refer the graphs shown in Figures 4.6(a) and (b)). To analyze the importance of temporal accuracy of VOP detection algorithm, the performance of SVM classifier is evaluated for the CV transition regions segmented using different VOP detection algorithms. Bar graph shown in Figure 4.11 represents the performance of SVM classifier for the different VOP detection algorithms. For the current database, the timing accuracy of VOP detection for the different algorithms is in the order of SPEC<COMB<ZFF-HE<Proposed. Subsequently, Figure 4.11 shows that the average accuracy of SVM varies in proportion to the temporal accuracy of VOP detection algorithms. Especially, as seen in Figure 4.11, the detection rate of normal stops is severely degraded for the state-of-the-art methods. The VOPs detected by the proposed algorithm are very much closer to ground truth than SPEC, COMB, and ZFF-HE methods. Hence, the performance of SVM classifier for the proposed VOP detection algorithm is closer to that for the manually labeled VOPs.

4.5 Classification of glottal and non-glottal misarticulations

In the previous section, the potential of CV transition anchored around VOP is demonstrated in the screening of misarticulated stops from normal. The basis for the choice of CV transition is that formant transitions of misarticulated stops are significantly different from the target stops. Among the different classes of misarticulations, compensatory articulation errors, i.e., glottal, pharyngeal, palatal, and velar substitutions are produced due to the shift in the place of articulation. Each of these compensatory errors produced with a different place of articulation, which can be differentiated based on the formant transitions. In this section, CV transitions are used to analyze the discrimination between two different categories of misarticulations. For this analysis, glottal and non-glottal misarticulations are considered.

Glottal stop is one of the most frequently occurring compensatory error in CLP speech [102]. For supporting this statement, the current database (refer, Table. 4.2) contains a larger number of glottal stops than pharyngeal, palatal, and velar substitutions. As described in the Section 2.6 of Chapter 2, CV transition region glottal stop exhibit flat formant transitions. Also, voicing is absent at the articulatory closure phase of glottal stops. Hence, the temporal structure of glottal stops is almost similar to that of unvoiced stops. In contrast to glottal stops, the weak, palatal, and velar stops are produced by making constriction within the oral cavity and hence, exhibit raising/falling formant transitions. In this thesis, weak, palatal, and velar stops produced for the target unvoiced stops are considered as the non-glottal stops. The non-glottal stops have a similar temporal structure like glottal stops and differ only in terms of place of articulation. Whereas, the weak, palatal, and velar stops produced for the target voiced stops differentiable from glottal stops in two aspects, namely, voicing and place of articulation. The details of glottal and non-glottal misarticulations used in the current experiment are provided in Table 4.6. In the CLP group of the current database, the production of glottal stop substitution varies with respect to the target place of articulation. For example children 'A' has produced glottal stop for target /k/, whereas non-glottal stop (velar substitution) for the target /t/. Due to this variability, database (Table 4.6) indicates that out of 1418 non-glottal stops, 498 are contributed by the children who produced glottal stops.

For the classification of glottal and non-glottal misarticulations, SVM classifier is trained for the SPF-based 2D-DCT coefficients computed from 60 ms of the transition region. Similar to earlier experiments, CV transition regions segmented using manually marked VOPs are used for the training purpose. The SVM classifier is trained using RBF kernel for the classification of glottal and non-glottal

Table 4.6: Database for glottal vs non-glottal misarticulation classification experiment. The class of non-glottal misarticulation comprised of weak, velar, and palatal stops produced for the unvoiced targets (details of symbols used in the table: B-boys, G-girls, μ-mean and σ-stranded deviation).

No. of children	Age ($\mu \pm \sigma$)	Glottal	Non-glottal
14 (9B+5G)	9.06±2.01	690	498
15 (9B+6G)	9.07±2.09	0	920

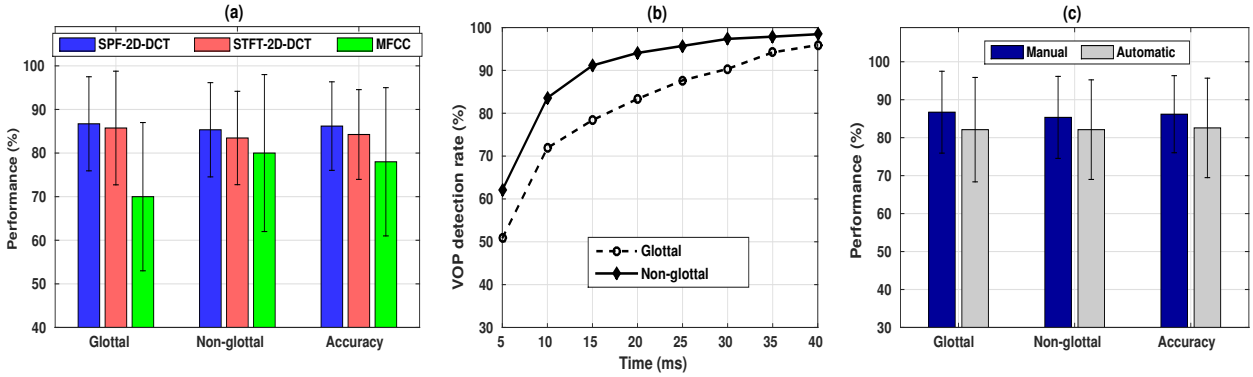

Figure 4.12: (a) Performance of glottal vs. non-glottal classification system for SPF-2D-DCT, STFT-2D-DCT, and MFCC features, (b) variation of VOP detection rate with respect to temporal deviation parameter, and (c) performance of glottal vs. non-glottal classification system for SPF-2D-DCT feature, computed using manually labeled and automatically detected VOPs.

misarticulations. The performance of SVM classifier is evaluated using the LOSO cross-validation technique. In the current database, some of the speakers produced both glottal and non-glottal misarticulations for the target unvoiced stops. Therefore, samples of a single speaker may be present in both glottal and non-glottal classes during the training of SVM classifier. For such speakers, only glottal realizations are considered during the training phase. Whereas, both glottal and non-glottal tokens are considered for testing. This avoids the repetition of the same speaker's data between glottal and non-glottal classes. Finally, by excluding the non-glottal realizations of the speaker who produces glottal stops, there are 690 glottal and 920 non-glottal misarticulations considered for the training purpose (Table 4.6).

LOSO cross-validated performance of SVM for SPF-based 2D-DCT, STFT-based 2D-DCT, and MFCCs features are represented in the form of bar graph, which is shown in Figure 4.12(a). To compare the features, training and testing of SVM classifier is carried out using manually marked VOPs. The usage of manual labels avoids the effect of segmentation error on the performance of classifier and hence, helps to analyze the significance of features. Figure 4.12(a) shows that proposed

SPF-based 2D-DCT performs better than the STFT-based 2D-DCT and MFCC features. The VOPs detected using the proposed algorithm are used for the automatic classification of glottal and non-glottal misarticulations. Figure 4.12(b) shows the variation of VOP detection rate with respect to different temporal deviations, for the syllables containing glottal and non-glottal misarticulations. Proposed algorithm assumes the presence of periodicity in vowel region. However, vowel followed by the glottal stop contains aperiodic components caused due to creaky phonation. Due to the presence of aperiodicity, the performance of proposed VOP detection algorithm is degraded for the glottal misarticulations. The performance of glottal vs. non-glottal classifier for the manually marked and automatically detected VOPs are compared in Figure 4.12(c). As shown in Figure 4.12(c), classification accuracy of glottal vs. non-glottal misarticulations is 86% for the manually marked VOPs, whereas it gets reduced to 82% for the automatically detected VOPs.

4.6 Summary

In this chapter, the significance of CV transition regions anchored around VOPs is analysed for the detection of misarticulated stops. Initially, to segment the CV transition regions, VOP detection algorithm is developed using the knowledge of syllable nucleus, and epoch-synchronously computed MWIP feature. Due to the epoch-synchronous processing, the proposed VOP detection algorithm detects the VOPs more accurately than state-of-the-art methods. The spectro-temporal dynamics of CV transition regions anchored around VOP are parameterized by 2D-DCT coefficients, which are computed using SPF-based TFR. Using SPF-based 2D-DCT feature, SVM classifier is trained for the classification of normal and misarticulated stops. Since SPF gives better spectro-temporal resolution than STFT, the classifier showed better performance for the proposed SPF-based 2D-DCT feature than STFT-based 2D-DCT and MFCC features. The performance of the proposed misarticulated stop detection system is tested using both manually marked and automatically detected VOPs. The performance of the normal vs. misarticulated stop classification system for the proposed VOP detection algorithm is near to the manually marked VOPs and better than the state-of-the-art VOP detection methods. Application of the proposed VOP-based framework is extended for the classification of glottal vs. non-glottal misarticulations.

In the present chapter, significance of VOPs and joint spectro-temporal features in the screening of various categories of misarticulated stops from normal is analyzed. In addition to that discrimination

of glottal stops from the non-glottal misarticulations is carried out using VOPs. In the next chapter, the detection of nasalization error is addressed. In particular, next chapter analyzes the effect of nasalization on voice bar and CV transition regions using epoch-synchronously computed spectral, spectro-temporal, excitation source, and periodicity features.

5

Detection of Nasalized Voiced Stops

Contents

Objective

The production of nasalized misarticulations for the target voiced stops indicates the presence of moderate-to-severe VPD. In this chapter, an epoch-synchronous framework is proposed for the detection of nasalized voiced stops in CLP speech. First, the speech regions corresponding to consonant and CV transitions are segmented using the knowledge of glottal activity, syllable nucleus, low-frequency spectral dominance, and VOPs. The segmented regions are epoch-synchronously processed to analyze the spectral, spectro-temporal, excitation source, and periodicity characteristics of normal and nasalized voiced stops. The spectral and spectro-temporal features are computed from SPF-based time-frequency representation. The amplitude of Hilbert envelope of linear prediction residual, measured around the epoch is used to analyze the effect of nasalization on excitation source. Comparison of speech frames of successive inter-epoch intervals is carried out to analyze the periodicity characteristics. The proposed features are used to develop an SVM classifier for the classification of normal and nasalized voiced stops. Performance of the proposed segmentation algorithm is compared with HMM-based force alignment approach. Also, the performance of epoch-synchronous features for the detection of nasalized voiced stops is compared with the conventional MFCCs. Further, the significance of proposed features in the discrimination of nasalized voiced stops from other voiced misarticulations, i.e., weak, palatal, and velar substitutions is analyzed.

5.1 Introduction

In the previous chapters, algorithms for the detection epochs and VOPs are developed. The detection events are used to demonstrate the systems for the screening of misarticulated stops from normal and detection of glottal stops in CLP speech. In this chapter, analysis of a particular category of misarticulation is carried out. Categorization of the misarticulated stops provides the details about structural/functional deviations associated with the speech production system. The present work focuses on the analysis and detection of nasalized stop, which indicates the presence of moderate-to-severe VPD [7, 46]. The production of stop consonants requires the complete closure of the nasal tract by velopharyngeal mechanism [1]. In the presence of moderate-to-severe VPD, vocal tract configuration for the stops will tend to reach as that of nasal consonants. As a result, stop consonants (/b/, /d/, /g/, /p/, /t/, /k/) produced by CLP speaker tend to be perceive like nasal consonants (/m/, /n/, /ng/). This category of misarticulation is referred as nasalized stops [7, 47, 175]. The

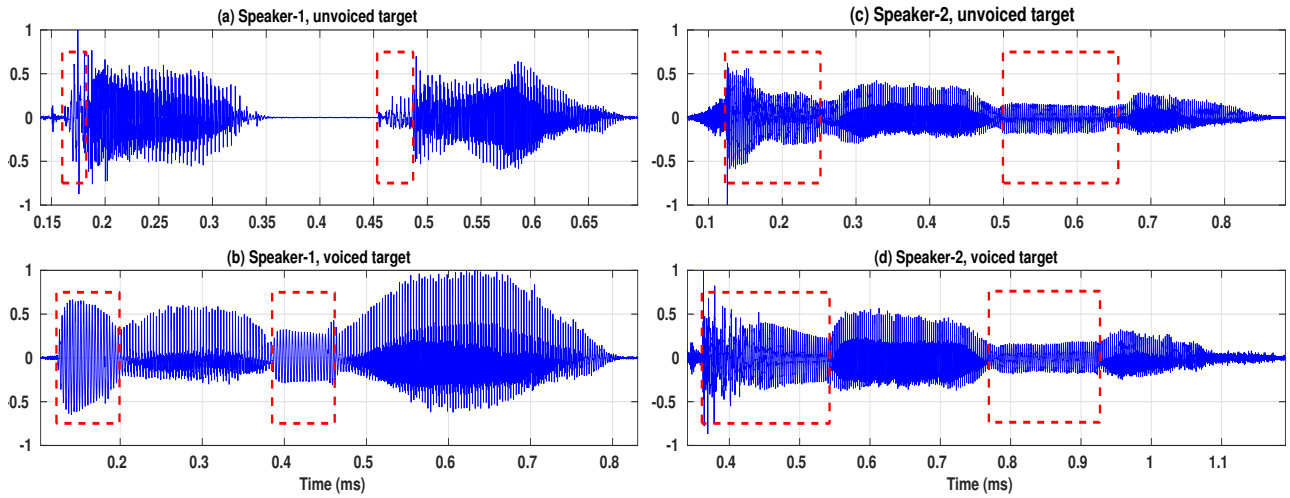

Figure 5.1: Illustration of nasalized misarticulations produced for the target unvoiced and voiced stops. (a) and (b) represent speech waveforms containing nasalized stops produced speaker-1 for the target unvoiced and voiced stops, respectively. (c) and (d) represent speech waveforms containing nasalized stops produced by speaker-2 for target unvoiced and voiced stops, respectively. In all subplots, the region corresponding to nasal murmur regions is indicated by a red colored rectangular box.

production of nasalized stops is considered as a primary indication of the presence of significant VPD [7]. In [46], nasal realization of stops is categorized as a severe consonant production error, where both surgery and therapy are recommended to correct it. Also, identification of nasalized stop consonants is helpful in the severity rating of hypernasality [47]. Therefore, the evaluation of nasalized stops is considered to be important during the diagnosis and therapy of individuals with CLP. The VPD is most commonly found in individuals with CLP, but it can also be present in non-cleft cases [48]. Mechanical obstructions like tonsils, irregularly shaped adenoid pad, and motor speech disorders, such as dysarthria (Parkinson's disease, cerebral palsy and Huntington disorder) and apraxia are the non-cleft causes of VPD [48, 177]. Therefore, in addition to CLP, detection of nasalized stops can also be helpful in the objective assessment of VPD caused by non-cleft conditions.

In presence of VPD, both voiced and unvoiced stops perceive like nasal consonants; however, voiced stops are more affected by nasalization. The nasalized stops are characterized by the presence of nasal murmur at the articulatory closure phase of stops [2]. Figure 5.1 demonstrates the speech waveforms of [CVCV] words containing nasalized stops produced for the unvoiced and voiced targets by two different speakers. In case of speaker-1, the closure phase of unvoiced stops is incompletely nasalized, where the nasal murmur is present only for a small interval (Figure 5.1(a)). Whereas, in case of speaker-2, the silence interval of unvoiced stop is completely replaced by nasal murmur(Figure 5.1(c)). Both speakers

demonstrated the presence of nasal murmur throughout the articulatory closure phase of voiced stops (Figure 5.1(b) and (d)). Therefore, the effect of nasalization is observed more consistently in case of voiced targets, than unvoiced ones. Also, in the current database, the number of nasalized stops for the unvoiced targets are very less as compared to voiced targets. The database (refer, Table 4.2 of Chapter 4) contains 600 number of voiced stops affected by nasalization, whereas only 194 samples for the unvoiced targets. Also, out of 31 CLP speakers in the current database, 12 have demonstrated the nasalization error for the voiced targets, whereas only 3 for the unvoiced targets. Due to the availability of less number of examples and the presence of inconsistent acoustic characteristics of nasalized stops for the unvoiced targets, the thesis focuses only on the analysis of nasalization effect on voiced targets.

5.1.1 Challenges

Voiced stops affected by the nasalization are referred as the nasalized voiced stops [47]. Automatic detection of nasalized voiced stops requires the segmentation and analysis of nasal cues from the words containing target voiced stops. Figures 5.2 (a) and (b) represent the waveform and spectrogram of normal voiced stops. From Figure 5.2(a) it can be observed that the voiced stop consonant comprised of multiple sub-phonetic segments, i.e., voice bar (signal corresponding to articulatory closure phase of voiced stop), burst, and formant transitions anchored around VOP. The waveform and spectrogram of nasalized voiced stop are depicted in Figures 5.2 (c) and (d), respectively. In the consonant region, the effect of nasalization can be observed in terms of increase in the energy of voice bar and absence of burst [2]. The CV transition of nasalized voiced stops is relatively smoother than the normal voiced stops. Therefore, the consonant and CV transition regions need to be segmented accurately for the analysis of different aspects of nasalization on voiced stops. In the literature, HMM-based force-alignment, short-time energy, and periodicity based approaches are proposed for the segmentation of various articulation errors [23, 102]. These methods may not be appropriate for the purpose of segmentation of nasalized voiced stops due to the reduced energy difference between consonant and vowel regions, and dynamically varying transition characteristics.

During the production of nasalized voiced stops, the velopharyngeal port is not completely open like nasal consonants. Hence, nasal resonances may be weakly evident in the spectrum of nasalized voiced stops. The computation of conventional block-based spectral features like MFCCs uses a fixed window of duration about 20-30 ms, where the position and size of analysis window affect the spectral

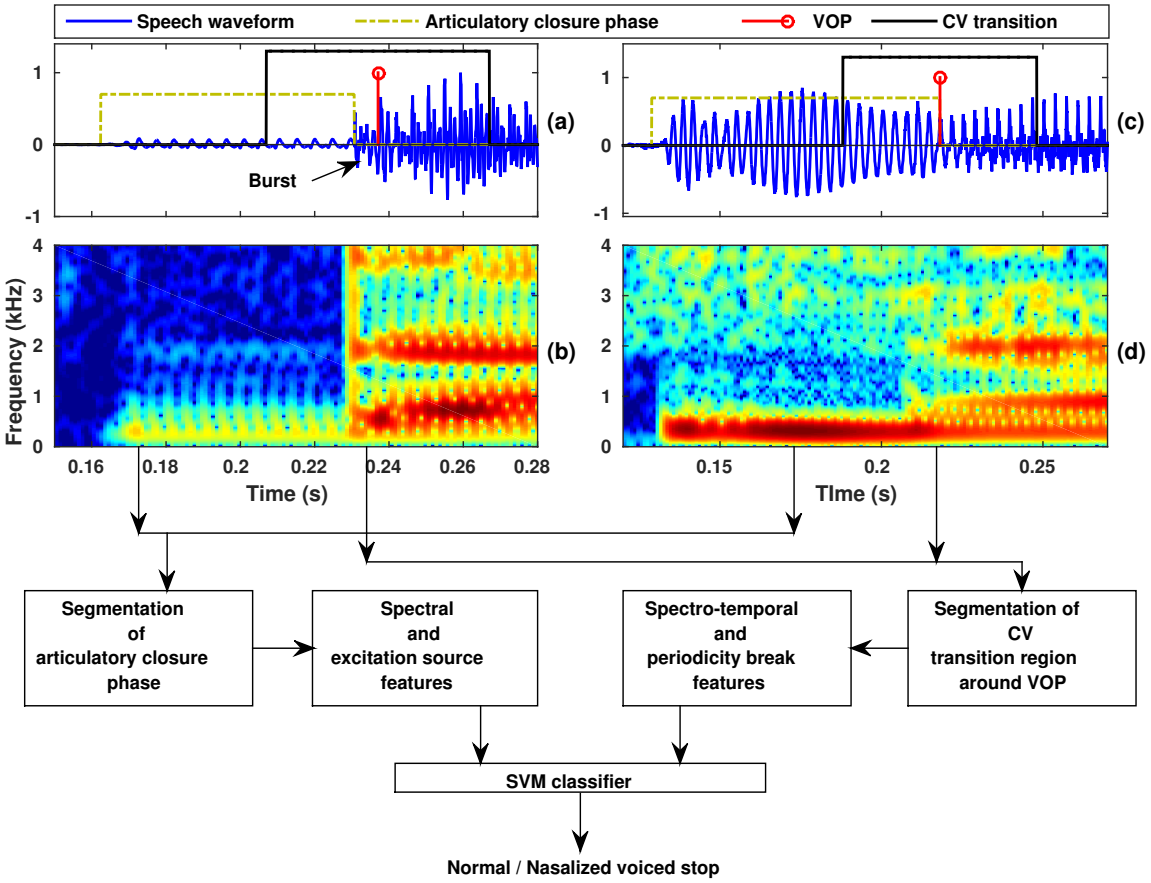

Figure 5.2: Proposed approach for the detection of nasalized voiced stops. (a) CV unit containing normal voiced stop and (b) its spectrogram. (c) CV unit containing nasalized voiced stop and (d) its spectrogram.

estimation [95, 96]. Hence, block-based spectral features may not be suitable for the characterization of weak nasal resonances present in the nasalized voiced stops. In addition this, the time-frequency resolution issues related to block-based features will create challenges for the analysis of CV transition regions [174]. Therefore, issues related to segmentation of consonant and CV transition regions, and shortcomings of block-based spectral features are considered as the major challenges for the detection of nasalized voiced stops.

5.1.2 The present work

The present work proposes an epoch-synchronous framework for the processing of nasalized voiced stops by addressing the issues involved in the segmentation and analysis of nasalized voiced stops. Block diagram depicted in Figure 5.2 gives an overview of the proposed approach for the detection of nasalized voiced stops. The proposed approach has the following key features.

(i) Effect of nasalization is analyzed from the consonant and CV transition region anchored around VOP. To segment these regions, a knowledge-based algorithm is proposed.

(ii) Spectral and spectro-temporal characteristics of nasalized voiced stops are analyzed using the spectrum derived from the epoch-synchronous processing of SPF-based TFR.

(iii) Hilbert envelope of linear prediction residual and similarity of speech segments measured between the successive inter-epoch intervals (glottal cycles) are used to analyze the effect of VPD on excitation source and periodic structure of the nasalized voiced stops, respectively.

(iv) Combination of spectral, spectro-temporal, excitation source, and periodicity break features are used to develop an SVM-based classification system for the classification of normal and nasalized voiced stops. Further, the application of proposed features in the classification of nasalized vs. non-nasalized misarticulations is demonstrated.

Performance of the proposed knowledge-based segmentation approach is compared with HMM-based force-alignment method. Whereas, the performance of proposed epoch-synchronous features is compared with the conventional MFCCs. Further, the significance of proposed epoch-synchronous features in the classification of nasalized and non-nasalized voiced misarticulations produced for the target voiced stops is analyzed.

The chapter is organized as follows: Section 5.2 describes the database. Section 5.3 explains the knowledge-based segmentation algorithm. Methods for the extraction of spectral, spectro-temporal features, excitation source, and periodicity break features are described in Section 5.4. Development of system for the classification of normal and nasalized voiced stops, and experimental results are presented in Section 5.5. The application of proposed features in the classification of nasalized and non-nasalized is discussed in Section 5.6. Finally, the chapter is summarized in Section 5.7.

5.2 Database

The database used in this thesis comprised of 31 children with CLP and 30 CN children. Different categories of misarticulations produced by 31 CLP speakers for the target voiced stops (/b/, /d/, /D/, /g/) are summarized in Table. 5.1. From Table. 5.1, it can be noticed that among the different consonant production errors, nasalized voiced stop is the most frequently occurring misarticulation for the target voiced stops. In the current database, nasalized voiced stop group is formed by 12 CLP

children (7 boys and 5 girls) of age range 6-12 years (9.08±2.46 years). A subset of CN group, i.e., 12 children (7 boys and 5 girls) belonging to the age range of 6-12 years (9.5±1.5 years) are considered. Further, 300 [CVCV] words containing voiced stops (300×2=600 tokens) are recorded from CN group.

Table 5.1: Different categories of misarticulations produced for target voiced stops

Categories→	Weak	Nasalized	Palatal	Velar	Devoiced	Glottal	Other
Number of tokens→	570	600	136	202	296	202	66

Voice bar, burst, and vowel regions of the recorded [CVCV] words are manually labeled by the careful visualization of waveform and spectrograms using PRAAT software. As shown in Figure 5.2(d), nasalized voiced stops do not contain burst component due to low intraoral pressure. Therefore, burst is marked only if there exists an impulse-like vertical striation corresponding to burst onset in the wide-band spectrogram (Figure 5.2(b)). Energy and low-formant frequency cues are considered for the marking of voice bar and nasalized voice bars. Vowel boundaries are marked based on the onset and offset of formants at vowel onset and offset points, respectively. Based on these rules, labeling is carried out by a person having knowledge of acoustic-phonetics. Further, 30% of the samples are given to another person for labeling to validate the reliability of manual makings. There is a small difference of 0.9±0.1 ms obtained between the labellings of two annotators.

5.3 Knowledge-based segmentation algorithm

In this work, a knowledge-based approach is proposed for the segmentation of consonant and CV transitions. The developmental set comprised of 50 [CVCV] words (25 from CN and 25 from CLP groups) is used to compute the thresholds required for the segmentation. Different stages involved in the proposed knowledge-based segmentation algorithm are explained in this section.

5.3.1 Segmentation of voiced consonant regions

Voiced stop consonant contains voice and burst components. Burst onset is an important event used for the identification of stop consonants in continuous speech. However, the burst may be absence or weakly evident in nasalized voiced stops. By addressing this issue, acoustic cues related to voice bar are considered for the identification of normal and nasalized voiced stops. Proposed algorithm for the segmentation of voiced consonant region is based on the knowledge of the glottal activity,

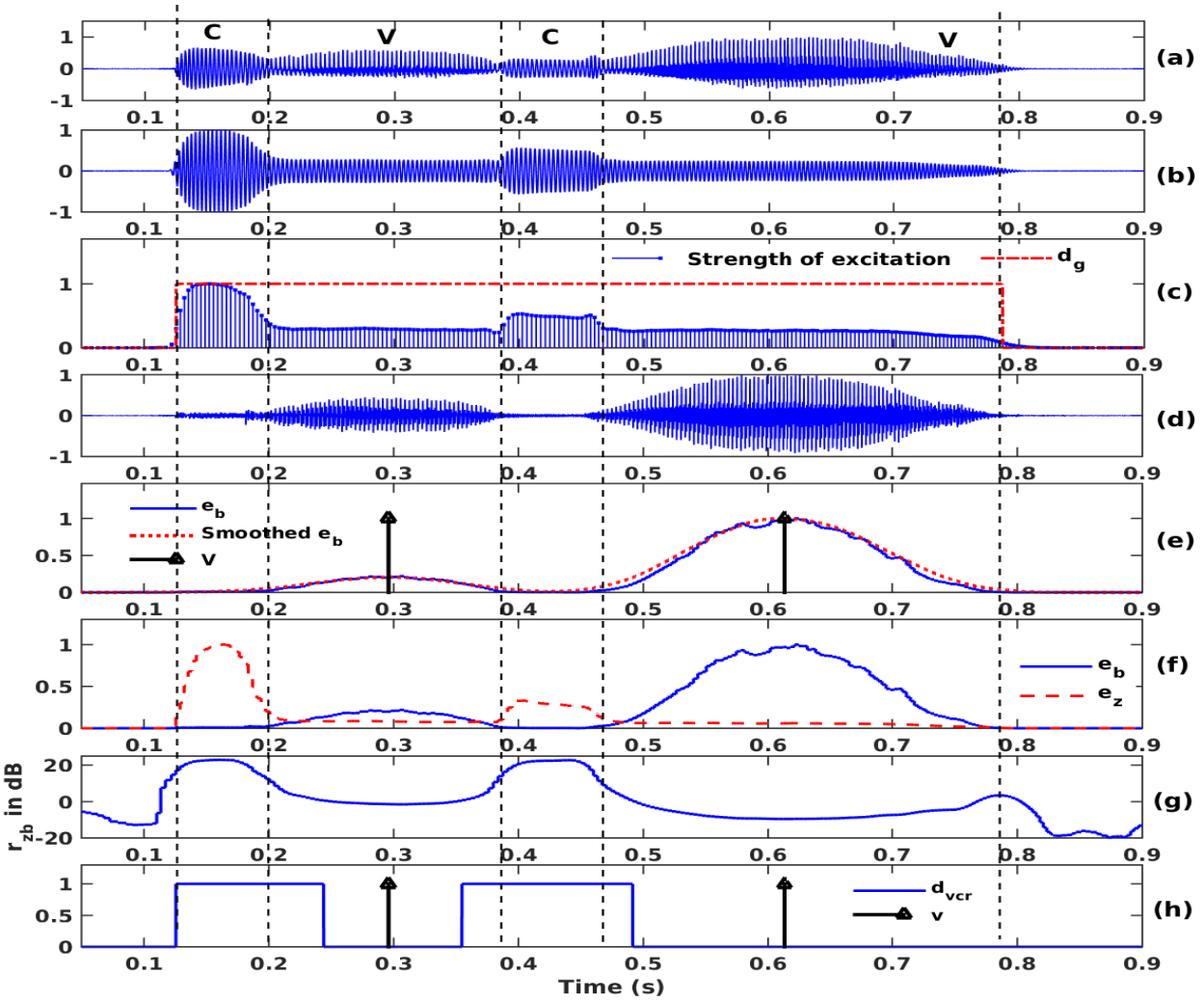

Figure 5.3: Segmentation of voiced consonant regions from [CVCV] word. (a) speech waveform, (b) ZFFS, (c) SoE along with the glottal activity decision (d_g), (d) BPFS (500-4000 Hz), (e) BPFS energy (e_b), smoothed-e_b, with detected syllable nuclei (V), (f) short-time energy contours of BPFS (e_b) and ZFFS (e_z), (g) ratio of e_z to e_b in dB, (h) decision of voiced consonant regions (d_{vcr}).

syllable nucleus, and low-frequency spectral dominance. Different steps involved in the segmentation of consonant regions are described in Figure 5.3 for the [CVCV] word containing nasalized voiced stops. These steps are explained as follows.

(i) **Glottal activity detection:** The speech signal is passed through zero-frequency filtering operation [35]. Speech signal and corresponding zero frequency filtered signal (ZFFS) are shown in Figures 5.3(a) and (b), respectively. Using ZFFS, SoE is computed. Based on SoE, the glottal activity regions are detected by implementing the procedure mentioned in Chapter 4. Figure 5.3(c) shows the detected glottal activity regions superimposed over SoE contour.

(ii) **Syllable nuclei detection:** The detected glottal activity region contains vowel and voiced consonants. Hence, to separate consonant part from vowel, the identification of vowel locations (syllable nuclei) is carried out first. Due to the effect of nasalization, speech waveform shown in Figure 5.3(a) indicates a reduced difference between the amplitudes of vowel and consonant regions. To suppress the effect of nasalization and enhance the discrimination between consonant and vowel regions, a fourth order Butter-worth band-pass filter having pass-band frequencies of 500-4000 Hz is used. Figure 5.3(e) shows the band-pass filtered signal (BPFS), which indicates the enhanced difference between vowel and consonant regions when compared to the original speech signal. The epoch synchronous short-time energy of BPFS ($e_b[n]$) is computed using Hamming windowed speech frames of 20 ms duration, anchored around the epochs. Further, $e_b[n]$ is smoothed using a Hamming window of 100 ms. Finally, within the glottal activity regions, peaks of smoothed-$e_b[n]$ are detected to locate the syllable nuclei. Figure 5.3(f) depicts $e_b[n]$, smoothed-$e_b[n]$, and detected syllable nuclei.

(iii) **Detection of low-frequency dominant voiced consonant regions:** After the identification of vowel or syllable nucleus of a CV unit, segmentation of consonant is carried out. Due to the absent or weakly evident burst in nasalized voiced stops, cues related to voice bar are considered for the segmentation of consonant regions. Both normal and nasalized voice bars are characterized by the presence of predominant low-frequency components below 500 Hz. To enhance the low-frequency dominant regions, ratio between the energies of ZFFS and BPFS is calculated as $r_{zb}[n] = e_z[n]/e_b[n]$. Here, $e_z[n]$ is epoch synchronous short-time energy of ZFFS. As shown in Figure 5.3(g), the evidence $r_{zb}[n]$ is relatively high at nasalized voice bar regions than vowels because of the presence of low-frequency dominant components. Using r_{zb} and glottal activity decision d_g, binary evidence for the voiced consonant d_{vcr} is computed as

$$d_{vcr}[n] = \begin{cases} 1 \; if \; (r_{zb}[n] > T_3) \; \& \; (d_g[n] = 1) \\ 0 \; otherwise, \end{cases} \tag{5.1}$$

where the threshold T_3 is given by,

$$T_3 = \frac{1}{M} \sum_{i=1}^{M} r_{zb}[v_i] + \epsilon \tag{5.2}$$

where v_i is the i^{th} syllable nucleus, M is the total number of detected syllable nuclei present in the given utterance. The parameter ϵ is determined as follows. For each syllable in the developmental set, difference between the average values of r_{zb} at the vowel and voice bar regions is computed using the manual labels. The mean and standard deviation of these difference values, across 100 syllables of the developmental set are found to be equal to 20 dB and 10 dB, respectively. The lower bound of the distribution, i.e., 10 dB is considered as the ϵ value. Further, using ϵ, threshold T_3 is computed to derive the binary evidence d_{vcr}. The d_{vcr} is plotted in Figure 5.3(h).

5.3.2 Refinement of boundaries

Segmentation of voiced consonants using d_{vcr} results in the inaccurate boundaries by including portions of adjacent vowels. Figures 5.4(a) and (e) represent the speech waveforms of [CVCV] words containing normal and nasalized voiced stops, respectively. Segmented consonant regions are shown in Figures 5.4(b) and (f), respectively, where the estimated boundaries are not accurate with respect to the ground truth (in Figure 5.4, ground truth is indicated by dashed lines). Therefore, a boundary refinement procedure is developed using r_{zb} for the accurate segmentation of consonants. The r_{zb} values are computed epoch synchronously, whose length is equal to the number of epochs present in the utterance. The value of r_{zb} at current epoch is appended over the entire inter-epoch interval (duration from current epoch to next epoch), to make the length of evidence equal to that of the utterance. The appended version is denoted as r_{zb_1}. The differenced r_{zb} i.e., r_{dzb} at an arbitrary epoch location g_0 epoch is computed using

$$r_{dzb}[g_0] = 10log_{10}\left(\frac{1}{D}\sum_{i=g_0}^{g_0+D} r_{zb_1}[i]\right) - 10log_{10}\left(\frac{1}{D}\sum_{i=g_0-D-1}^{g_0-1} r_{zb_1}[i]\right) \tag{5.3}$$

where D is chosen equal to the number of samples corresponding to 15 ms. r_{dzb} is computed at every epoch present in the entire utterance. The r_{dzb} for normal and nasalized voiced stops are shown in Figures 5.4(c) and (g), respectively. Only for the word-initial syllable, the onset of glottal activity is considered as the beginning instant of the consonant. Within the detected consonant region, the location of the deepest valley is determined as the end point of the consonant. Whereas, for the word medial consonants, locations of the largest peak and the deepest valley of r_{dzb} present within the detected consonant regions are determined. These peaks and valleys are used to refine the medial consonant boundaries. The boundary refinement procedure results in the precise boundaries of

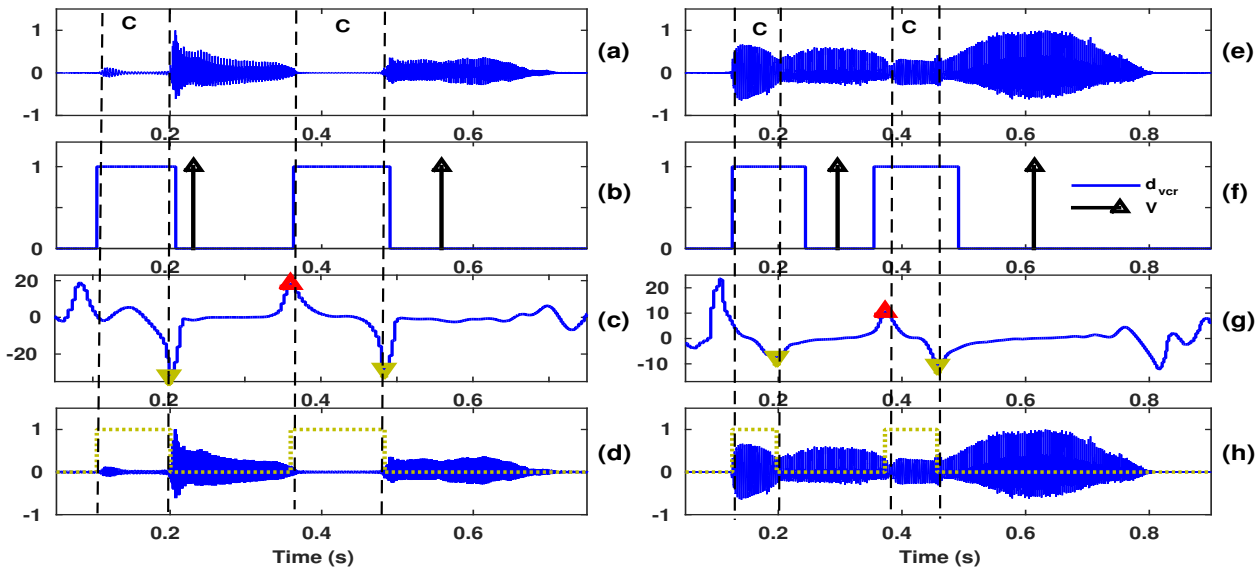

Figure 5.4: Refinement of segmented boundaries of normal and nasalized voice stops. (a)-(d) and (e)-(h) represent the speech signal, segmented consonant regions, differenced r_{zb}, and refined boundaries for normal and nasalized voice bars, respectively. The dashed lines indicate the ground truth boundaries of normal and nasalized voice bars.

normal/nasalized voiced bar regions of normal/nasalized voiced stops. The normal and nasalized voice bar regions resulted after the application of boundary refinement procedure are shown in Figures 5.4(d) and (h).

5.3.3 Segmentation of CV transition regions

The knowledge of VOP helps to segment the CV transition regions. In Chapter 4, the VOP detection was carried out using the knowledge of syllable nuclei and MWIP evidence. The algorithm showed a reduction in its performance for the misarticulated stops. The presence of smooth CV transition in nasalization error is one of the important reason for the reduced performance. To improve the VOP detection accuracy, modified VOP detection algorithm is proposed in this chapter. In the modified algorithm compares MWIP values in a forward direction from the consonant region. Whereas, the algorithm developed in Chapter 4 compares the MWIP values in a backward direction. In the modified algorithm, the end point of voice bar resulted after the boundary refinement procedure (refer Section 5.3.2) is considered as the reference point. From the reference, MWIP values are compared in a forward direction to locate VOP. The modified VOP detection algorithm using MWIP evidence and end point of voice bar involves the following stages:

(i) Let g_i be the epoch corresponding to the end point of voice bar region.

(ii) Check the MWIP values at $g_{(i)}^{th}$ and $g_{(i+1)}^{th}$ epochs, whether they are above threshold T_2 or not. T_2 is computed using the procedure mentioned in section 4.3.3.

(iii) If the condition mentioned in step (ii) get satisfied, then the i^{th} epoch is chosen as the VOP.

(iv) If condition in step (iii) is not satisfied, then update g_i to $g_{(i+1)}$ and repeat (ii) and (iii) till 20 ms from the reference.

(v) Within the search interval (20 ms), if no epoch satisfied step (iii), then the endpoint of the voice bar is claimed as the VOP.

Between the end point of voice bar and VOP, the region corresponding to burst is present. Hence, the search interval needs to cover the lead VOT (interval between burst onset to VOP) of the voiced stops. To determine the search interval, distribution of VOT values of normal voiced stops present in the database is obtained (since the burst is not present in the nasalized voiced stops, only normal voiced stops are considered). The mean and standard deviations of the distribution of VOT values are equal to 12 and 8 ms respectively. The upper bound of the distribution, i.e., 20 ms is chosen as the search interval.

Figure 5.5 illustrates the modified VOP detection algorithm for the syllables containing normal and nasalized voiced stops. Figure 5.5(a) represents the syllable containing normal voiced stop. The endpoint of the voice bar region detected using the r_{zb} feature is shown in Figure 5.5(b). Generally, voiced stop in normal speech contains voice bar and burst components. Figure 5.5(d) shows an increase in the MWIP evidence at soon after VOP due to the onset of glottal vibrations and increase in the band energy (500-4000 Hz). Figures 5.5(d)-(f) depict the syllable containing nasalized voiced stop, binary evidence for the detection of nasalized voice bar region, and MWIP evidence. In case of nasalized voiced stops, the burst is not evident. Due to this, the end point of nasalized voice bar coincides with VOP. As shown in Figure 5.5(f), MWIP values will be greater than T_2 at the end point of nasalized voice bar. Hence, the VOP searching process ended up at the reference point itself. Therefore, the reference point (endpoint of the nasalized voice bar region) is considered as the VOP. After the detection of VOP, 60 ms region around VOP (30 ms on the left and 30 ms on the right sides of VOP) is segmented as the CV transition region.

The proposed knowledge-based approach for the segmentation of voiced consonant and CV transition regions is summarized as follows:

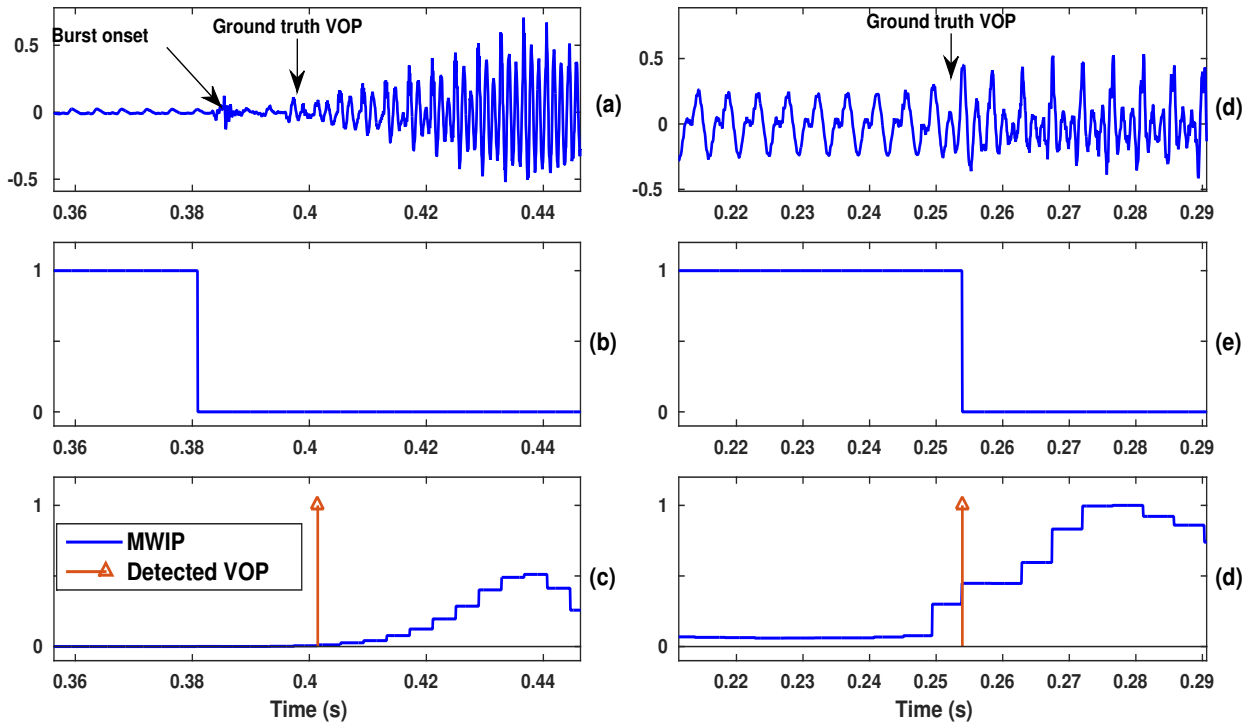

Figure 5.5: Detection of vowel onset point (VOP). (a)-(c) and (d)-(f) speech segment representing consonant-vowel boundary with manually marked VOP, end point of voice bar region, and MWIP evidence with detected VOP, for normal and nasalized voiced stops, respectively.

(i) The epoch locations and SoE are computed using ZFFS.

(ii) Using SoE, glottal activity regions are segmented using equation 4.1 given in chapter 4.

(iii) Within the glottal activity regions, syllable nuclei are detected by processing the BPFS computed using a band-pass filter having passband frequencies in the range of 500-4000 Hz.

(iv) Ratio of ZFFS to BPFS energy (r_{zb}) is computed to emphasize the voiced consonant regions.

(v) Based on the knowledge of glottal activity, syllable nucleus and r_{zb}, consonant regions are segmented using equation 5.1.

(vi) Boundary refinement of detected consonant regions is carried out using the peaks and valleys present in the differenced r_{zb}, i.e., r_{dzb} computed using equation 5.3.

(vii) Using r_{dzb} and MWIP, the VOPs are detected for the segmentation of CV transitions.

5.4 Computation of epoch-synchronous features

Segmented consonant and CV transition regions are epoch-synchronously processed to extract (a) spectral, (b) spectro-temporal features, (c) excitation source, and (d) periodicity break features. Computation of these features is explained in this section.

5.4.1 Spectral and spectro-temporal features

Around the epochs, impulse-like energy is delivered to the vocal tract system. Due to impulse-like excitation, the region around epochs has relatively high signal-to-noise ratio (SNR). The estimation of vocal tract characteristics around the epochs is considered to be better than the conventional block-processing approaches like STFT [96]. In this work, SPF-based TFR is epoch-synchronously processed to estimate the spectrum around each epoch. The SPF-based TFR is computed using the procedure explained in Chapter 3. Figures 5.6(a) and (b) depict the waveform and SPF-based spectrograms corresponding to a syllable containing normal, respectively. The SPF-based TFR shown in Figure 5.6(b) indicates the concentration of spectral energy around the vertical striations, corresponding to epochs. To estimate the spectrum around these vertical striations, the epoch locations need to be identified accurately. For this purpose, the precise epoch locations are obtained using the multi-scale product-based method proposed in Chapter 3.

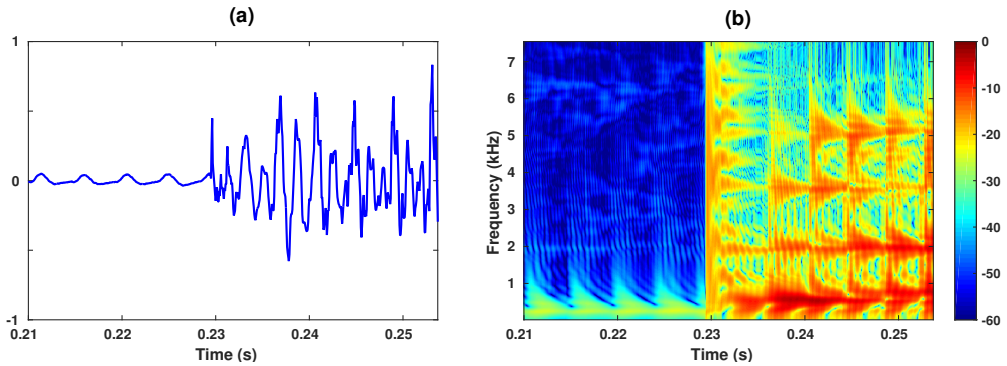

Figure 5.6: (a) Speech segment corresponding to voiced stop-to-vowel transition, (and (b) corresponding SPF-based spectrogram.

Let $V[n, f]$ denotes the SPF based time-frequency representation and $G = \{g_1, g_2, g_2, ..., g_N\}$ represents the set of detected epochs, where N corresponds to total number of epochs present in the entire utterance. After each epoch g_i, a small region $V[g_i + \delta_n, f]$ in time-frequency domain is considered. Further, $V[g_i + \delta_n, f]$ is averaged over δ_n number of samples to get an estimate of spectrum around

the epoch. The estimated spectrum $X_{g_i}[f]$ around g_i^{th} epoch is given by

$$X_{g_i}[f] = \frac{1}{\delta_n} \sum_{n=g_i}^{g_i+\delta_n} V[n, f] \tag{5.4}$$

where $i \in \{1, 2, ..., N\}$ and δ_n equals to the number of samples corresponding to 1 ms ($\delta_n = 16$ samples for $f_s = 16000$ Hz), which is less than the minimum pitch period (2 ms) of human.

Spectral features: Epoch-synchronously estimated spectrum is used for the computation of spectral features from the consonant regions. The segment corresponding to normal voiced stop-vowel transition, superimposed with the detected epochs is shown in Figure 5.7(a), and the corresponding SPF-based TFR is depicted in Figure 5.7(b). Epoch-synchronously estimated spectra are shown in Figure 5.7(c). Similarly, Figures 5.7(e)-(g), represent the CV transition region, SPF based time-frequency representation, and estimated spectrum around each epoch, respectively for the syllable containing nasalized voiced stop. The epoch-synchronously estimated spectra shown in Figures 5.7(c) and (g) indicate a significant difference between the normal and nasalized voice bar regions. The degree of vocal tract constriction is high for normal voiced stops than nasalized voiced stops. During the production of nasalized voice bars, the acoustic energy flows through the nasal tract. Whereas, during the production of voice bars, the acoustic energy propogates through the pharyngeal wall, which offers greater resistance than that of nasal tract. Hence, the spectral energy of voice bars is less than that of nasalized voice bars. Due to the reduced vocal tract constriction, F_1 of nasalized voice bar region is higher than that of voice bar. In Figure 5.7(c), voice bar shows the presence of strong resonance around 200 Hz, corresponding to closed vocal tract resonance [1]. Whereas, in Figure 5.7(g) nasalized voice bar shows the presence of high-frequency resonances introduced due to the propagation of air-flow through the nasal tract.

To capture the effect of nasalization from voice bar spectrum, discrete cosine transform (DCT) is used. DCT coefficients ($c_{g_i}[l]$) computed on the SPF-based log-magnitude spectrum at ${g_i}^{th}$ epoch is given by

$$c_{g_i}[l] = \sqrt{\frac{2}{M}} \sum_{f=0}^{M-1} w[l] log(X_{g_i}[f]) cos\left(\frac{\pi l[2f+1]}{2M}\right) \tag{5.5}$$

where $w[l] = 1/\sqrt{2}$, for $l = 0$ and $w[l] = 1$ for $l \neq 0$. M corresponds to the number of SPF-based filtered envelopes, and $l = 0, 1, ..., L-1$, where L is the number of DCT coefficients. In this work, the value of L is chosen equal to 13. These 13-DCT coefficients are referred as DCT feature. The 0^{th} cepstral coefficient represents the mean value of the log spectral envelope. Distribution of $c[0]$ is

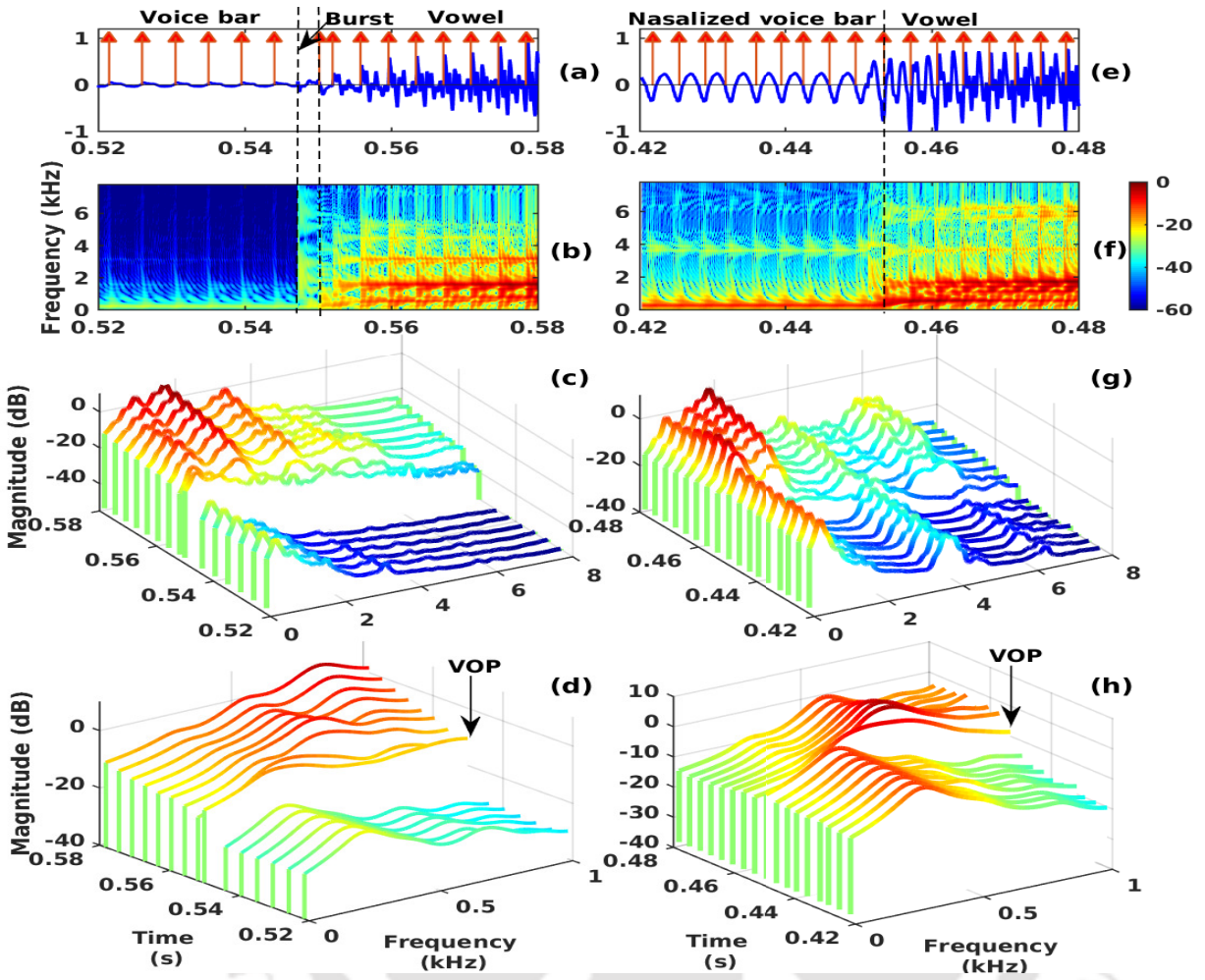

Figure 5.7: Spectral and spectro-temporal characteristics of normal and nasalized voiced stops. (a)-(c) and (e)-(g) represent the segment of CV transition superimposed with epoch locations (vertical arrows), SPF-spectrogram, estimated spectrum around each epoch, for normal and nasalized voiced stops, respectively. (d) and (h) show the variation of spectral components within 100-1000 Hz as a function of time for normal and nasalized voiced stops, respectively.

shown in Figure 5.10(a). The distribution indicates that the nasalized voice bars have higher energy than normal voice bars. Coefficient $c[1]$ captures the spectral slope, whose distribution shown in Figure 5.10(b) indicates that spectral slope of the nasalized voice bars is less than that of normal voice bars. Other higher dimensional DCT coefficients capture the higher order spectral moments of normal and nasalized voice bars.

Spectro-temporal features First formant (F_1) transition is considered as an important feature for the discrimination of voiced stops from nasals [44]. To analyze the F_1 dynamics at CV boundary, frequency components from 100 to 1000 Hz are considered. Figures 5.7(d) and (h) represent the epoch-to-epoch variation of spectral characteristics in the vicinity of first formant (100-1000 Hz). Due

to the effect of VPD, F_1 transition is relatively smooth for nasalized voiced stops than normal voiced stops (Figures 5.7(d) and (h)).

To analyze the effect of nasalization on CV transition region, 2D-DCT based joint spectro-temporal features are computed [59]. Consider a time-frequency region $Z[f, n]$ of size $K_1 \times N_1$, where N_1 is the number of epochs present within the transition region and K_1 is the number of SPF-based filtered envelopes present with in the range of 100-1000 Hz. The 2D-DCT coefficients $\alpha_{(p,q)}$ of $log(Z[n, f])$ are computed using,

$$\alpha[p,q] = \sqrt{\frac{2}{N_1}}\sqrt{\frac{2}{K_1}}\sum_{n=0}^{N_1-1}\sum_{f=0}^{K_1-1} w[p]w[q]log(Z[n,f])cos\left(\frac{\pi p[2n+1]}{2N_1}\right) \times cos\left(\frac{\pi q[2f+1]}{2K_1}\right) \quad (5.6)$$

where $w[p] = 1/\sqrt{2}$, for $p = 0$ and $w[p] = 1$ for $p \neq 0$. Similarly, $w[q] = 1/\sqrt{2}$, for $q = 0$ and $w(q) = 1$ for $q \neq 0$. Here, p and q are given by $p = 0, 1, 2..., P-1$ and $q = 0, 1, 2..., Q-1$, where P and Q represent the number of rows and columns of 2D-DCT coefficient matrix, respectively. In this work, $P = 3$ and $Q = 3$ are chosen. Distribution of $\alpha[(1, 0]$ is shown in Figure 5.10(c). The 2D-DCT coefficient $\alpha[1, 0]$ encodes the temporal slope of spectral energy over the transition duration. Figures 5.7(d) and (h) indicate an abrupt CV transition for the normal voiced stop, whereas nasalized voiced stop shows relatively smooth transition. Hence, the distribution of $\alpha[1, 0]$ shown in Figure 5.10(c) indicates a significant discrimination between the normal and nasalized voiced stops. Further, the 2D-DCT matrix of size 3×3 is concatenated to form 9-dimensional feature vector.

The key difference between the 2D-DCT feature computed in this chapter and the previous chapter (Chapter 4) lies in the choice of frequency band. In the previous chapter, the frequency band ranging from 100 to 7800 Hz are considered, whereas this chapter considers only the range of 100-1000 Hz. Therefore, the spectro-temporal feature computed in this chapter is referred as sub-band 2D-DCT. The 2D-DCT feature captures the dynamics of F_1-F_4, while sub-band 2D-DCT encodes only F_1 transition.

5.4.2 Excitation source feature

During the articulatory closure phase of nasalized voiced stops, the presence of VPD allows the air-flow to escape through the nasal tract. Due to the loss of air-flow, intra-oral pressure developed for the nasalized voiced stops is less than the normal voiced stops. This reduced intra-oral pressure increases the transglottal pressure (pressure across the glottis) [1]. The abruptness of glottal closure is proportional to transglottal pressure. The presence of VPD in nasalized voiced stops increases the

transglottal pressure and hence, increases the abruptness of glottal closure [118, 178].

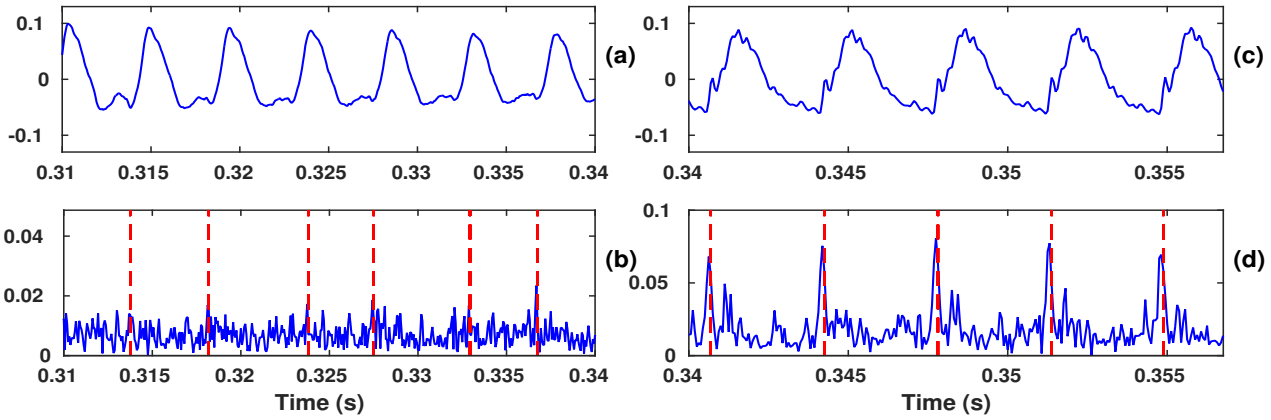

Figure 5.8: Nature of Hilbert envelope of linear prediction residual (HELPR) for normal and nasalized voiced stops. (a) and (c) show the segments of normal and nasalized voice bars, respectively, their Hilbert envelope of LP residuals are plotted in (b) and (d). Dashed lines in (b) and (d) indicate the detected peaks of HELPR around the epochs.

The strength of impulse-like peaks present in Hilbert envelope of LP residual (HELPR) attributes to the abruptness of glottal closure [118]. In this work, LP analysis is carried out for the hamming windowed speech frames of size equal to 20 ms with a shift of 5 ms. The order of LP analysis is chosen equal to $f_s/1000 + 4$, where f_s is the sampling frequency. Figures 5.8(a) and (c) represent the waveforms of normal and nasalized voice bars, corresponding HELPR are shown in Figures 5.8(b) and (d). Due to the effect of increase in the transglottal pressure, HELPR of nasalized voice bar shows sharp impulse-like peaks. Amplitude of the peaks present in HELPR around the epochs (1.5 ms on either sides of epoch) is computed and it is referred as HELPR feature. Figure 5.10(d) shows the distribution of HELPR feature. The distribution indicates higher values for nasalized voiced stops than normal voiced stops.

5.4.3 Periodicity break feature

The presence of noise-like burst breaks the periodic structure of CV syllable containing normal voiced stop. Whereas, in case of nasalized voiced stops, burst is weakly evident or absent due to inadequate intra-oral pressure [2]. In this work, speech segments of adjacent inter-epoch intervals are compared using cosine distance metric to analyze the periodic structure of the signal. Let s_i and s_{i+1} correspond to the speech segments of i^{th} and $(i+1)^{th}$ inter-epoch intervals, respectively. Cosine

similarity (ρ_i) between s_i and s_{i+1} is computed using

$$\rho_i = \frac{\langle s_i . s_{i+1} \rangle}{\|s_i\|_2 \, \|s_{i+1}\|_2} \tag{5.7}$$

where $\|s_i\|_2$ and $\|s_{i+1}\|_2$ are the l_2-norms of the segments s_i and s_{i+1}, respectively, $i = 1, 2, 3..., N-1$, and N is the total number of epochs present in the utterance. In order to compute ρ_i, segments s_i and s_{i+1} should have equal length, which is achieved by zero-padding. The value of ρ_i is appended over entire i^{th} inter-epoch interval to make the evidence equal to the length of the signal.

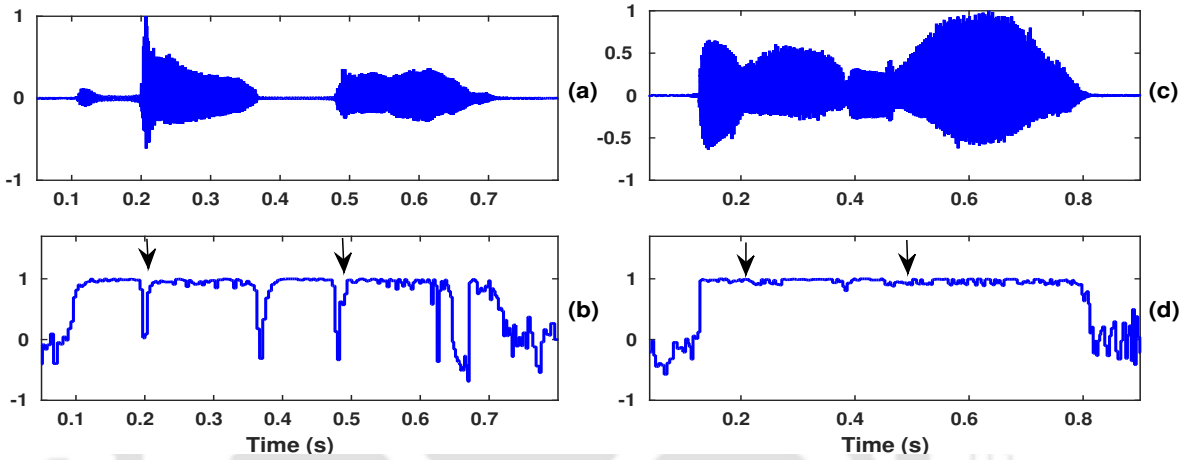

Figure 5.9: Periodicity break feature. (a)-(b) and (c)-(d), represent speech signal, Cosine similarity evidence, for [CVCV] word containing normal and nasalized voiced stops, respectively. The downward arrows in (b) and (d) indicate the location of the minimum value of cosine similarity evidence measured from the beginning of the consonant to syllable nucleus.

Figures 5.9(b) and (d) show the cosine similarity measure (ρ) computed for the [CVCV] words containing normal and nasalized voiced stops, respectively. The cosine similarity values are high for voice bar and vowel regions because of the presence of periodicity. Since voiced stop contains the aperiodic burst component, a sudden drop in the cosine similarity evidence is observed in Figure 5.9(b). Due to the absent or weakly evident burst, cosine similarity measure does not show a sudden drop for the nasalized voiced stop case (Figure 5.9(d)). To quantify the effect of nasalization, a measure called periodicity break (Δ_ρ) is computed as

$$\Delta_\rho = min(\rho[n]), n \in (n_c, n_v) \tag{5.8}$$

where n_c and n_v are the sample indices corresponding to beginning of normal/nasalized voice bar and syllable nucleus, respectively. Figure 5.10(e) shows the distribution of Δ_ρ for normal and nasalized voiced stops, where the evidence is high for the nasalized voiced stops than voiced stops.

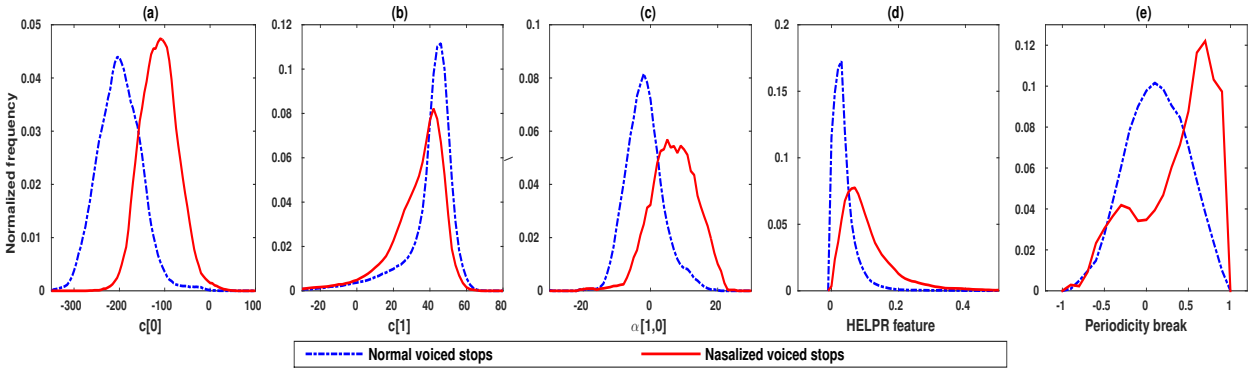

Figure 5.10: Distribution of features for normal and nasalized voiced stops: (a) the 0^{th} DCT coefficient ($c[0]$), (b) the 1^{st} DCT coefficient ($c[1]$), (c) $(1,0)^{th}$ element of sub-band 2D-DCT matrix ($\alpha[1,0]$), (d) HELPR feature, and (e) periodicity break.

5.5 System for detection of nasalized voiced stops

The features proposed for the analysis of nasalized voiced stops are summarized in Table 5.2. Combination of knowledge-based segmentation algorithm and epoch-synchronous features is used to develop a system for the automatic detection of nasalized voiced stops. The DCT and HELPR features computed from the voice bar region of voiced stop consonant. Hence, the lengths of DCT and HELPR features are equal to the number of epochs present within the consonant region. Hence, DCT and HELPR features are grouped as *consonant level features*. The sub-band 2D-DCT and periodicity break features extracted from each CV unit present in an [CVCV] utterance, and these features grouped as *syllable level features*. Using DCT and HELPR features, SVM-consonant is developed, whereas sub-band 2D-DCT and periodicity break features are used to develop SVM-syllable.

Table 5.2: Summary of features

Sl. No	Feature	Dimension	Computed from
1	DCT	13	Voice bar
2	sub-band 2D-DCT	9	CV transition region
3	HELPR	1	Voice bar
4	Periodicity break	1	Beginning of consonant to syllable nucleus

The features extracted from manually labeled sound units are used to train SVM-consonant and SVM-syllable systems. During the testing phase, the features are computed from the automatically segmented regions using knowledge-based approach. The SVM scores are normalized between 0 and 1 using logistic regression technique. The score level fusion of SVM-consonant and SVM-syllable is performed to analyze the significance of the combination of consonant and syllable level features. The

length of SVM scores obtained from SVM-consonant equals to the number of epochs present in the consonant region. These scores are averaged to get a single value per consonant region. Further, the averaged scores from SVM-consonant are combined with the scores of SVM-syllable. The class for which the combined score is found to be highest is claimed as the class corresponding to the test phoneme.

The performances of proposed knowledge-based segmentation algorithm and SVM classification system are evaluated using the database containing normal and nasalized voiced stops. In the database, there are 600 tokens present for each case of normal and nasalized voiced stops cases. Excluding the samples of developmental set (50 normal and 50 nasalized voiced stops), 1100 tokens (550 normal and 550 nasalized voiced stops) are considered for the evaluation of the performance of segmentation algorithm. However, due to the availability lesser number of samples, all 1200 tokens are considered for evaluating SVM classifier.

5.5.1 Evaluation of segmentation algorithm

Performance of the proposed knowledge-based algorithm for the segmentation of consonant regions is validated in terms of correct detection and false alarm rates. A voiced consonant is said to be detected correctly if it covers at least 10 ms of the manually labeled normal/nasalized voice bar. The correct detection rate is computed as $P_c = N_c/N_{vcr}$, where N_c and N_{vcr} are the number of correctly identified voiced consonant regions and total number of voiced consonants, respectively. Similarly, if more than 10 ms of non-voiced consonants such as silence or vowel is detected as voiced consonant, then it is considered as a false detection. The false alarm rate is computed as $P_f = N_f/N_{nvcr}$, where N_f and N_{nvcr} are the number of non-voiced consonants detected as voiced consonant and the total number of non-voiced consonants, respectively. The considered [CVCV] utterances contain silence, voiced consonant, and vowels. Hence, with reference to this chapter, the term 'non-voiced consonant' refers to silence and vowel regions. The difference between the temporal boundaries for detected voiced consonant (voice bar region) and ground truth (manually marked labels) is computed to evaluate the temporal accuracy of voiced consonant segmentation algorithm. Performance of knowledge-based segmentation algorithm is presented in Table 5.3.

Performance of VOP detection algorithm: Manually marked VOPs are considered as the ground truth, using which the performance of proposed VOP detection algorithm is evaluated. The modified VOP detection algorithm proposed in this chapter is based on comparison of MWIP values in a forward

Table 5.3: Performance of proposed and HMM-based methods for the segmentation of voice bar region of normal and nasalized voiced stops. P_c-Correct detection rate, P_f-False detection rate, and t_e-timing error.

	P_c(%)		P_f (%)		t_e (ms)	
Voiced consonant	Proposed	HMM	Proposed	HMM	Proposed	HMM
Normal voiced stop	97.46	**97.67**	**1.98**	45.83	**7.29±3.52**	39.03± 24.99
Nasalized voiced stop	97.25	**97.83**	**3.43**	39.99	**9.90±4.28**	42.31±18.00

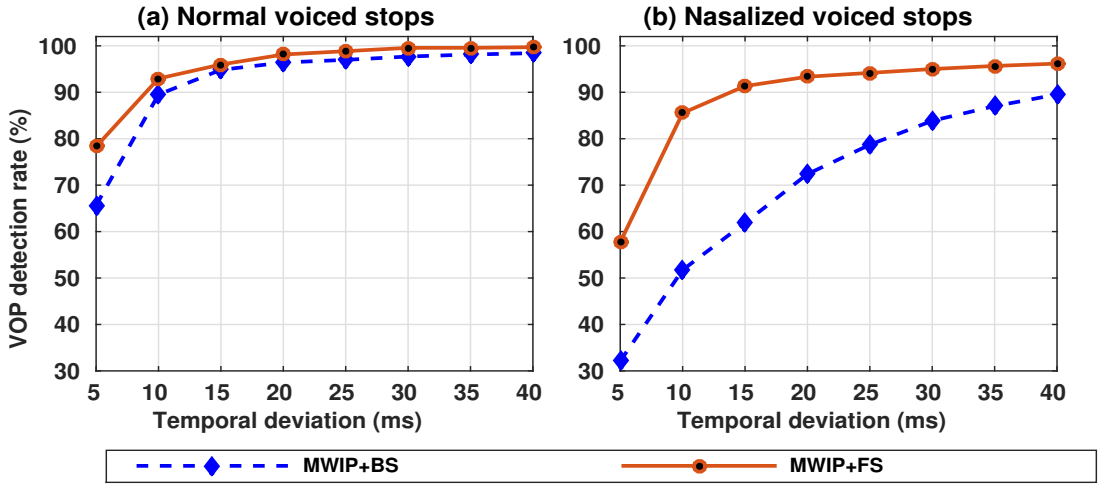

Figure 5.11: Variation of VOP detection rate as a function of temporal deviation for the syllables containing (a) normal and (b) nasalized voiced stops. The figure shows the performance of both MWIP+BS (backward searching) and MWIP+FS (forward searching) algorithms.

direction from the voice bar region. Since the modified VOP detection algorithm involves forward searching (FS), it is referred to as MWIP+FS. The VOP detection algorithm developed in Chapter 4 is based on the backward searching (BS) from syllable nucleus, which is referred as MWIP+BS. The performance of these algorithms for normal and nasalized voiced stops are shown in Figures 5.11(a) and (b), respectively. Figure 5.11(a) shows that both MWIP+BS and MWIP+FS algorithms almost give a similar performance for the voiced stop case; however, the forward searching improves the VOP detection accuracy slightly. The performance of MWIP+BS and MWIP+FS algorithms for nasalized voiced stops is compared in Figure 5.11(b). For the nasalized voiced stop case, the detected VOPs by MWIP+FS approach are more accurate than MWIP+BS. This improvement indicates that the forward searching approach significantly overcomes the VOP detection challenges caused due to smooth CV transitions. However, the forward searching algorithm critically depends on presence of voicing in the consonant region and it is suitable only for voiced consonants.

5.5.2 Evaluation of SVM classifier

The performances of SVM-consonant and SVM-syllable are evaluated at the phone level using the parameters: sensitivity (detection rate of nasalized voiced stops), specificity (detection rate of normal voiced stops), and overall classification accuracy. The performance of SVM is evaluated using leave-one-subject-out (LOSO) cross-validation technique. In this work, both SVM-consonant and SVM-syllable are developed using radial basis kernel function. Optimum values of kernel parameters (C, γ) are determined using grid search over the sets $C = [2^{-15}, 2^{-13}, ..., 2^{15}]$ and $\gamma = [2^{-15}, 2^{-13}, ..., 2^{0}]$ [10]. Experiments are carried out by training the SVM using the features extracted from manually labeled regions, whereas features computed from (1) manually labeled regions, (2) knowledge-based segmented regions, and (3) HMM-based segmented regions, are used for testing.

SVM classification experiments using manually labeled regions: The SVM-consonant and SVM-syllable classifiers are trained and tested using the features extracted from manually labeled regions. Consideration of manually labeled regions avoids the dependency of the classifier's performance on the segmentation errors, and help to analyze the contribution of features in a better way.

SVM-consonant: Correlation coefficients between each dimension of DCT and HELPR feature are ranging from -0.34 to 0.64. Table 5.4 shows that feature level combination of HELPR with DCT feature improves the performance of SVM-consonant when compared to that for DCT alone. This improvement shows the significance of the excitation source information in the detection of nasalized voiced stops.

Duration of transition region: Computation of sub-band 2D-DCT feature requires the segmentation of CV transition region anchored around VOP. To segment CV transition regions, the speech segments are considered on either side of VOP. For example, 40 ms of CV transition corresponds to the region 20 ms left and 20 ms right sides of VOP. Using manually marked VOPs, experiments are carried out by varying the transition duration from 20 to 80 ms, in steps of 10 ms. Cross-validated accuracy is found to be highest for the transition region of 60 ms duration, which is considered as the optimum duration of the transition region.

SVM-syllable: The performance of SVM-syllable for manually labeled regions is mentioned in Table 5.4. The experimental results in Table 5.4 indicate that sub-band 2D-DCT feature derived from CV transition performs better than DCT extracted from the consonant region. This result correlates with the perceptual experiments conducted on the consonant identification in [179], where the results

Table 5.4: Performance of nasalized and normal voiced stop classification system for the manually labeled regions.

Classifiers	Features	Nasalized (%)	Normal (%)	Accuracy (%)
SVM-consonant	DCT	85.01±16.05	89.87±11.24	87.57±13.83
	DCT+HELPR	85.99±12.19	90.28±10.11	88.86±10.80
	MFCC	79.54±28.89	83.12±12.80	82.12±17.67
SVM-syllable	Sub-band 2D-DCT	89.73±10.56	91.98±8.90	91.51±8.89
	Sub-band 2D-DCT+Δ_ρ	90.38±10.45	92.21±8.13	91.99±9.76
	MFCC	73.17±24.78	84.48±7.81	978.5±19.56
Score level fusion	All features	91.89±8.31	93.29±8.28	92.91±9.89
	MFCC	78.89±20.44	85.56±10.67	81.45±17.23

Table 5.5: Performance of nasalized and normal voiced stop classification system for the knowledge-based and HMM-based force alignment methods.

Classifier	Features	Knowledge-based segmentation method			HMM-based force alignment		
		Nasalized (%)	Normal (%)	Accuracy (%)	Nasalized (%)	Normal (%)	Accuracy (%)
SVM-consonant	DCT	83.49±16.03	87.28±10.89	86.13±14.90	79.12±17.79	86.89±10.78	83.50±16.89
	DCT+HELPR	85.12±13.59	89.89±10.91	87.45±12.90	80.89±16.89	86.12±8.78	83.52±17.12
	MFCC	78.68±25.67	81.89±12.78	80.17±19.32	72.56±25.11	79.56±14.91	75.22±20.40
SVM-syllable	Sub-band 2D-DCT	86.96±11.24	90.27±9.28	88.53±9.63	34.47±28.89	68.90±7.89	43.89±23.24
	Sub-band 2D-DCT+Δ_ρ	89.19±10.12	90.55±7.16	89.89±9.52	39.17±29.80	70.96±15.12	55.12±20.45
	MFCC	70.02±29.89	79.96±8.76	76.17±22.62	28.79±18.90	47.89±12.90	40.34±15.90
Score level fusion	All features	90.51±13.51	92.72±9.56	91.21±9.80	64.46±10.91	75.67±12.11	70.91±11.12
	MFCC	76.13±18.92	80.12±9.60	78.14±6.80	69.32±22.19	71.89±10.10	69.94±18.90

showed that CV transitions highly contribute to the perception of consonants than the sustained consonant regions. The correlation coefficients between each dimension of sub-band 2D-DCT and periodicity break feature are found be ranging from -0.19 to 0.33. The feature level combination of sub-band 2D-DCT and periodicity break features improves the performance of SVM-syllable. This improvement showed the significance of periodicity break measure in the discrimination of normal and nasalized voiced stops.

Score level fusion of SVM-consonant and SVM-syllable: Table 5.4 shows that the syllable level features perform better than the region level features. Further, score level fusion of SVM-consonant and SVM-syllable improves the classification results, which is observed in Table 5.4. Thus, compared to the

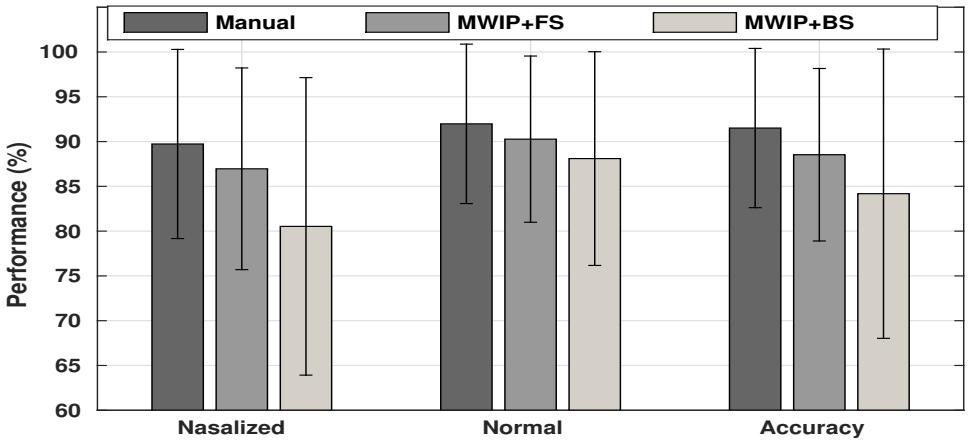

Figure 5.12: Comparison of performance of SVM-syllable for sub-band 2D-DCT feature computed using manually marked VOP, automatically detected using MWIP+FS (forward searching), and MWIP+BS (backward searching) approaches.

features derived individually from the consonant and transition regions, their combination is better able to characterize the effect of nasalization on the voiced stops.

Results for automatically segmented regions: Performance of the proposed system for the features computed using automatically segmented regions is evaluated. In Table 5.5, results for the various combination of epoch-synchronous features computed from knowledge-based segmented regions show a similar trend as that for manual labels. Due to the timing error in the segmented regions and falsely detected regions, the performance of SVM for automatically segmented regions is slightly lesser than that for the manual labels. For the knowledge-based segmentation case, the reduction in SVM's performance is found to be high for sub-band 2D-DCT feature. This result indicates that the spectro-temporal feature is highly sensitive to the temporal accuracy of VOP detection algorithm. The proposed algorithm estimates around 95% of VOPs within the temporal resolution of ± 25 ms. Temporal deviation in the detected VOPs reduces the performance of SVM-syllable for the sub-band 2D-DCT feature. Figure 5.12 shows the performance of sub-band 2D-DCT feature computed using manually marked VOPs, MWIP+FS, and MWIP+BS based VOP detection algorithms. As described in Figure 5.11, VOPs detected by MWIP+FS approach are more closure to the ground-truth than MWIP+BS. Therefore, the performance of SVM is found to be higher for the VOPs detected by MWIP+FS approach than MWIP+BS. The results indicate the importance of temporal accuracy of event detection algorithm in the event-synchronous processing of misarticulated stops.

5.5.3 Comparison with HMM-based segmentation

HMM-based force alignment method is used in [23] for the segmentation of consonant production errors in CLP speech. To compare the proposed knowledge-based segmentation algorithm with HMM-based force alignment method, HMM-based phoneme recognizer is developed using open source HTK toolkit [180]. There are 1200 [CVCV] words ([baba], [dada], [gaga], and [DaDa]) recorded from 30 CN children of age range 6-12 years (including the 12 CN children mentioned in Section II) used for training. HMM-based phoneme recognizer is developed for the 6 classes of sound units, i.e., /b/, /d/, /D/, /g/, /a/, and silence. Since the burst of the voiced stop has a very short duration, closure bar and burst are merged to form a single consonant unit. 39-dimensional MFCC feature vector (13-dimensional MFCCs, it's Δ and $\Delta\Delta$ variants) computed using 20 ms frame size with a frameshift of 10 ms is employed in the development of HMM. A context independent, 3-state left to right HMM trained with a 16 mixture continuous density diagonal covariance Gaussian mixture model per state is used to model each of the phoneme classes. Force alignment is conducted against the transcription of target [CVCV] words to get the phoneme boundaries. Table 5.3 shows the performance of HMM-based segmentation method. Results indicate that the correct detection rate of consonant region is slightly better for HMM-based segmentation method than the proposed knowledge-based approach. However, the false detection rate and timing error are high for HMM-based method than the knowledge-based approach. Increase in the timing error associated with HMM-based segmentation affects the performance of SVM-consonant and SVM-syllable as shown in Table 5.5. The onset of vowel is characterized by the abrupt transition in the signal energy. Hence, it is difficult to obtain the accurate VOP using HMM-based approach. Due to the erroneous segmentation of CV transitions by HMM, the performance of SVM-syllable is severely degraded for sub-band 2D-DCT feature. The training of HMM requires 1200 words, whereas the proposed approach uses only 50 words for the computation of thresholds. In the current work, 1200 words (2400 CV tokens) are recorded from 30 normal speakers for the target voiced stops. Since the performance of HMM will be good for a large number of training examples, all the samples present in the database are used to train HMM. Relatively small number of samples, i.e., 50 words (100 CV tokens) are used for the estimation of thresholds in case of knowledge-based segmentation algorithm. This indicates the advantages of the knowledge-based approach under limited data conditions.

5.5.4 Comparison with MFCCs

Experiments are carried out to compare the performance of proposed epoch-synchronous features with MFCCs, which is the most commonly used block-based feature [23, 102]. For the comparison, 39-dimensional MFCC feature is computed using Hamming windowed speech frames of 20 ms duration with a shift of 5 ms. SVM-consonant is trained and tested using MFCCs computed for each frame. For SVM-syllable, twelve frames of 39-dimensional MFCCs are computed from 60 ms transition region and concatenated to form a 468-dimensional feature vector. Further, principal component analysis (PCA) is applied for the dimensionality reduction. The 150-dimensional feature vector resulted after the application of PCA is used to train and test the SVM-syllable. Tables 5.4 and 5.5 indicate the performance of MFCC feature for manual and automatic segmented regions, respectively. The performance of SVM-consonant for MFCC feature is comparable to DCT. However, the performance of SVM-syllable for MFCC is severely degraded, which indicates that MFCCs may not be efficient to capture the spectro-temporal dynamics of transition regions. Automatic classification of articulation errors in CLP speech based on the combination of HMM-based segmentation and MFCC feature is proposed in [23]. The combination proposed knowledge-based segmentation and epoch-synchronous features outperformed the HMM-based segmentation and MFCC feature for the classification of normal and nasalized voiced stops. This improvement signifies the importance of epoch-synchronous features in the detection of nasalized voiced stops.

5.6 Classification of nasalized and non-nasalized misarticulations

In the previous section, experiments are carried out for the detection of nasalized voiced stops in presence of normal voiced stops. The applications of proposed segmentation algorithm and features are extended in this section for the classification of nasalized and non-nasalized misarticulations. Here, nasalized group is formed by nasalized voiced stops. Whereas, the non-nasalized group comprised of weak, palatal, and velar substitutions produced for the target voiced stops. Both nasalized and non-nasalized groups have similar voicing characteristics, but they are mainly differentiable in terms of the manner of articulation. Whereas, remaining errors produced for the voiced stops, i.e., glottal and devoiced stops differ in terms of voicing characteristics. Table 5.6 shows the details of database used for nasalized and non-nasalized misarticulation classification experiments. The production of nasalized stops depends on the place of articulation and voicing characteristics of targets. For example, speaker

'A' produced nasalized stop for bilabial target /b/, weak stop for velar target /g/. Hence, in the current database, the speakers who produced nasalized stops are also present in the non-nasalized group. To avoid the repetition of the same speaker's samples between nasalized and non-nasalized groups during training phase, the following strategies have been adopted.

- If a particular speaker has both nasalized and non-nasalized realizations, then only nasalized misarticulations are used during training.
- Both nasalized and non-nasalized misarticulations are used in the testing phase.

Table 5.6: Database for nasalized vs non-nasalized stop classification experiment. Non-nasalized voiced stop class comprised of weak, velar, and palatal stops produced for the voiced targets (details of symbols used in the table: B-boys, G-girls, μ-mean and σ-stranded deviation).

No. of children	Age ($\mu \pm \sigma$)	Nasalized	Non-nasalized
12 (7B+5G)	9.08±2.46	600	178
12 (8B+4G)	9.54±1.96	0	730

Similar to normal and nasalized voiced stops classification system, the SPF-DCT and HELPR features computed from the consonant region are used train SVM-consonant. The sub-band 2D-DCT and periodicity break features are used to train SVM-syllable. Further, the score level fusion of SVM-consonant and SVM-syllable is carried out. The LOSO cross-validated results of SVM-consonant and SVM-syllable are presented in Table 5.7. The results indicate that the feature level combination of DCT and HELPR feature showed an improvement over DCT alone. Similarly, concatenation of sub-band 2D-DCT and periodicity break features showed an improvement over the sub-band 2D-DCT alone. Further, score level fusion of SVM-consonant and SVM-syllable enhances the detection rate of nasalized stops in presence of non-nasalized misarticulations. Also, Table 5.7 shows that the classification results obtained for the knowledge-based segmentation algorithm are comparable with that of manually labeled regions. These experimental results showed a similar trend as that of nasalized vs normal voiced stop classification. The results indicate that the epoch-synchronous framework proposed for classification of nasalized vs. normal voiced stops is also suitable for the discrimination of nasalized vs. non-nasalized misarticulated stops.

Table 5.7: Performance of nasalized vs. non-nasalized misarticulated stop classifier.

Classifiers	Features	Manual labels			Knowledge-based segmentation		
		Nasalized (%)	Non-nasalized (%)	Accuracy (%)	Nasalized (%)	Non-nasalized (%)	Accuracy (%)
SVM-consonant	DCT	78.66±17.27	81.88±18.89	79.85±16.02	76.07±17.55	83.07±15.55	80.01±14.84
	DCT+HELPR	78.56±16.77	81.95±18.96	79.94±15.79	76.07±17.55	83.37±15.35	80.27±14.67
SVM-syllable	Ssub-band 2D-DCT	79.55±15.54	84.81±23.71	85.86±11.77	78.17±12.22	85.21±17.56	84.36±13.36
	Sub-band 2D-DCT+Δ_ρ	80.09±16.42	88.21±17.89	86.33±11.95	79.02±16.16	84.74±18.56	84.7±13.46
Score level fusion	All features	81.81±13.70	89.64±16.15	87.79±11.88	80.64±12.91	88.62±15.96	85.85±14.22

5.7 Summary

In this chapter, automatic segmentation and detection of nasalized voiced stops in CLP speech are carried out. Knowledge of the glottal activity, syllable nuclei, low-frequency dominance, and VOP are utilized for the segmentation of consonant and CV transition regions. From the segmented regions, epoch-synchronous spectral, spectro-temporal, excitation source, and periodicity break features, which capture the effect of nasalization on voice bar, CV transition, abruptness of glottal closure, and burst, respectively are computed. Combination of these features showed an improvement in the performance of SVM-based normal and nasalized voiced stop classifier. This improvement signifies that the misarticulation information is heterogeneously distributed among the different sub-phonetic units of stops; hence the combination of information from different subphonetic units helps to characterize the misarticulations in a better way. The experimental results showed that the proposed knowledge-based segmentation algorithm and epoch-synchronous features outperform HMM-based segmentation and MFCC feature. Further, the efficacy of proposed framework is tested for the detection of nasalized stops in presence of non-nasalized misarticulations.

The works presented till now in the thesis are focused towards the development of event-synchronous methods for the classification of normal vs. misarticulated, and normal vs. nasalized voiced stops. The significance of event-synchronous features is evaluated for the discrimination of different categories of misarticulations, such as glottal vs. non-glottal, and nasalized vs. non-nasalized. In the next chapter, different binary classifiers are combined to demonstrate a framework for the assessment of misarticulated stops in CLP speech.

6

Combined Framework for Assessment of Misarticulated Stops

Contents

Objective

The objectives of assessment of stop consonant production errors include the identification of presence/absence of misarticulation and classification of the category of misarticulation. In this chapter, previously proposed event detection algorithms and event-synchronous features are combined to demonstrate their applications in the assessment of misarticulated stops in CLP speech. The VOP-based normal vs. misarticulated stop classifier is used to construct a spider plot-based representation for the screening of misarticulated stops from normal. Spider plot provides a visual interpretation for assessing the stop consonant articulation capability of an individual. Further, using event-synchronous features a multi-class classification system is demonstrated for the detection of the category of misarticulation. The binary classifiers developed for the classification of normal vs. misarticulated stop, glottal vs. non-glottal, nasalized vs. non-nasalized misarticulation, are combined hierarchically for the multi-class classification of misarticulated stops.

6.1 Introduction

In the previous chapters, the focus is made on the (a) identification of significant events for the processing of misarticulated stops, (b) development of algorithms for the detection of events and (c) extraction of features around the events that capture the information about the deviant articulation. Epochs and VOP are considered as the events for the analysis of misarticulated stops. By addressing the issues related to epoch extraction from pathological children speech, an epoch extraction algorithm is proposed in Chapter 3. Epochs are extracted by processing SPF-based TFR. In Chapter 4, the knowledge of epochs is used for the detection of VOP. The detected VOPs are used for the segmentation of CV transition regions. The SPF-based 2D-DCT features computed from the CV transition regions are used to develop an SVM-based system for the classification of normal and misarticulated stops. Also, in Chapter 4 the application of VOPs in the discrimination of glottal stops from non-glottal misarticulations is demonstrated. In Chapter 5, the effect of nasalization on the voiced stops is analyzed using consonant and CV transition regions. The consonants present in [CVCV] utterance are identified using the voice bar component of voiced stops. Based on the knowledge of the glottal activity, low-frequency dominance, and VOP events, an algorithm for the segmentation of consonant and CV transition regions is developed. The epoch-synchronously computed spectral, spectro-temporal, excitation source and periodicity break features are used for the classification of normal and nasalized

voiced stops. Further, the effectiveness of the proposed epoch-synchronous features is analyzed for the discrimination of nasalized voiced stops from non-nasalized misarticulations.

In clinical scenarios, objectives of the assessment of misarticulated stops include the screening of misarticulated stops from normal and detection of category of misarticulation [48]. The screening test gives a primary information related to the presence or absence of misarticulation. This information will help to evaluate the outcome of surgery or speech therapy. However, only binary-level classification of normal and misarticulated stops is not sufficient to decide the kind of speech therapy or surgery required to correct the misarticulation. Identification of the type of misarticulation provides an information about the structural/functional deviations present in the speech production system, and also the compensatory mechanism adapted by the speaker. Therefore, by identifying the type of misarticulation, an appropriate plan for speech therapy or surgery can be suggested [48, 181].

To meet clinical requirements, the speech processing based assessment system should be able to differentiate the misarticulated stops from normal and classify the type of misarticulation. The binary classifiers developed in the previous chapters have an ability to discriminate (a) normal vs. misarticulated stops, (b) glottal vs. non-glottal, and (c) nasalized vs. non-nasalized misarticulations. In the present chapter, these binary classifiers are combined to demonstrate the application of event-synchronous features in the assessment of misarticulated stops. The major contributions of this chapter are mentioned as follows:

- SVM scores derived from normal and misarticulated stop classifier for the eight target stops (/k/, /t/, /T/, /p/, /g/, /d/, /D/, and /b/) are used to construct spider plots for the visual interpretation of stop consonant production errors.
- A three class classificication system is proposed for the classification of normal, glottal, and non-glottal misarticulations produced for the target unvoiced stops (/k/, /t/, /T/, and /p/).
- A hierarchical system is developed for the classification of normal, nasalized, non-nasalized, glottal, and devoicing errors produced for the target voiced stops (/g/, /d/, /D/, and /b/).

The chapter is organized as follows. Section 6.2 describes the construction of spider plots for the screening of normal and misarticulated stops. The multi-class classification systems are explained in Section 6.3. Finally, the chapter is summarized in Section 6.4.

6.2 Spider plot based representation

In Chapter 4, eight different SVM classifiers are developed for the detection of misarticulations produced for the eight target stops i.e., /k/, /t/, /T/, /p/, /g/, /d/, /D/, and /b/ (refer block diagram shown in Figure 4.8). SVM classifier for each target stop is trained using 27-dimensional SPF-based 2D-DCT features computed from CV transition regions. The SVM scores computed for eight target stops are combined to construct a spider plot. Procedure for the construction of spider plot is explained as follows.

- **Data:** The responses of an individual for the eight [CVCV] words corresponding to eight target stops are recorded.
- **Computation of SVM scores:** For each [CVCV] word, VOPs are detected using the MWIP feature and the knowledge of syllable nucleus. The detected VOPs are used to segment the CV transition region, and 2D-DCT features are computed. Based on the *a priori* knowledge of target stop, an appropriate SVM is selected. The selected SVM is tested using 2D-DCT feature. The SVM gives the scores corresponding to normal and misarticulated classes. SVM scores are normalized between 0 and 1 using logistic regression technique. Further, SVM scores computed for the misarticulated stop class are considered. For the misarticulated stop class, if the normalized SVM score is equal to 1, then it indicates the presence of misarticulation. Whereas, if the normalized SVM score is equal to 0, then it indicates the absence of misarticulation. Since each word contains two instants of target stop, two SVM scores are obtained for each word. These SVM scores are averaged to get a single score per word.
- **Generation of Spider plots:** The SVM scores obtained for all eight target stops are used to construct the spider plot. Figure 6.1(a) demonstrates the spider plot constructed for the normal children. The spider plot consists of three octagons. 'Outermost octagon' represents the highest value of SVM score for misarticulated class, which is equal to 1. The 'Outermost octagon' is represented using white color. A regular octagon constructed with the circumcircle radius equals to 0.5 is considered as the reference for normal articulation and this octogon is referred as 'reference octagon'. The reference octagon is indicated by blue color. The octagon constructed using the SVM scores corresponding to the subject's response is referred as 'test octagon', which is indicated by red color.

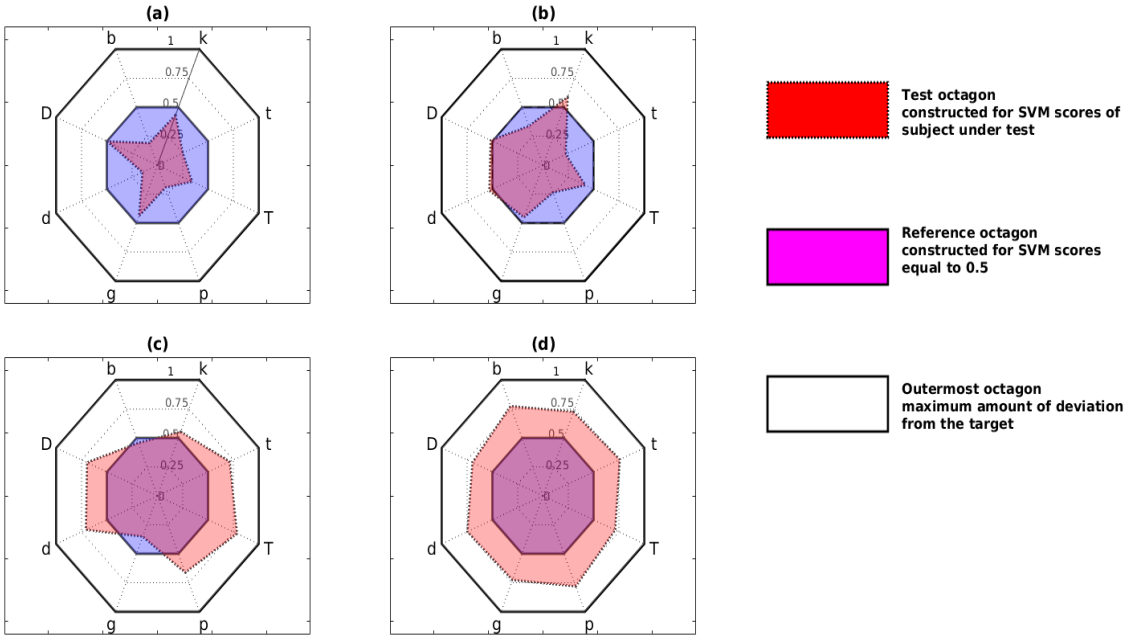

Figure 6.1: Illustration of spider plots constructed using SVM-scores for 8 target stops. (a) represents for normal children, (b), (c), and (d) represent for different CLP cases. Subject in (b) showed normal articulation. Subject in (c) diagnosed with the presence of velar substitutions for dental (/t/ and /d/) and retroflex (/T/, and /D/), and weak bilabial /p/. Subject in (d) reported with the presence of nasalization error for the eight target stops. The SCAI values for the cases (a)-(d), equal to 8.25%, 16.45%, 37.79%, and 56.65%, respectively. In each octagon, white colored octagon indicates the outermost octagon, which represents the maximum amount of deviation from the normal. Octagon filled with blue color represents the reference octagon. Red colored octagon is constructed for the subject under the test using the SVM scores of misarticulated stops.

6.2.1 Analysis of spider plots

In this work, eight SVM classifiers for the normal and misarticulated stops produced for the targets /k/, /g/, /t/, /d/, /T/, /D/, /p/, and /b/. During testing, the target information is known to be *a prior* and the appropriate SVM is chosen out of 8-SVMs trained for the different targets. For the normal articulation, the normalized SVM scores obtained for the misarticulated stop class are expected to lie within 0.5. Wheareas, for the misarticulated one, the SVM score is expected to be grater than 0.5. To analyze the significance of spider plots in the assessment of misarticulated stops, four speakers are considered. The spider plots constructed for the considered four speakers are shown in Figures 6.1(a)-(d), respectively. Figure 6.1(a) represents the spider plot corresponding to a normal subject. Hence, in Figure 6.1(a), all the vertices of octagon obtained for this subject lie within the reference octagon and indicate the presence of normal articulation. The spider plot shown in Figure 6.1(b) corresponds to a CLP subject, who successfully attained normal articulation after

speech therapy. Therefore, corresponding SVM scores are nearly equal to the normative value, i.e., 0.5. The spider plot shown in Figure 6.1(c) corresponds to the CLP case, diagnosed with the presence of velar substitutions for the dental (/t/, /d/) and retroflex (/T/, /D/) stops, and weakly articulated /k/ and /p/. Hence, the spider plot shows the deviation only for /t/, /T/, /d/, /D/, and /p/. Spider plot demonstrated in Figure 6.1(d) corresponds to a CLP case, who produced nasalization error for all eight targets. Therefore, the octagon corresponding to the subject's response exceeds the normal limits for all eight targets. Based on the observations from spider plots shown in Figure 6.1, it can be noticed that the spider plot provides a visual interpretation for the screening of misarticulation produced for each target stop.

6.2.2 Stop consonant articulation index

Different cases illustrated in Figure 6.1 indicated that area of test octagon increases as the production of the number of misarticulated stops increases. Based on the observation from Figure 6.1, a measure called stop consonant articulation index (SCAI) is derived. The SCAI (ψ) is computed as

$$\psi = \frac{A_t}{A_m} \times 100 \tag{6.1}$$

where, A_t represents the area of test octagon and A_m corresponds to the area of outermost octagon. The SCAI values for the speakers shown in Figure 6.1(a)-(d) are equal to 8.25%, 16.45%, 37.79%, and 56.65%, respectively.

The SCAI values are computed for the 30 normal and 31 CLP children present in the current database (refer Chapter 4, Section 4.2). For each child, one session of recording, which comprised of all eight target [CVCV] words are considered. Figure 6.2 shows the relationship of the SCAI with the number of misarticulated stops. In this work, eight stop consonants are used to compute SCAI; hence, the number of misarticulated stops can reach maximum value up to '8'. For the controlled normal children, the number of misarticulated stops is equal to '0', which indicates all the eight targets are produced correctly. As shown in Figure 6.2, SCAI value for normal children has a range of 0-20 %. Figure 6.2 shows that the SCAI value increases in proportionate to the number of misarticulated stops. Figure 6.2 indicates that the children who exhibit a lesser number of misarticulations, their SCAI value is very near to that of normal. Therefore, SCAI can be used as an objective measure to assess the stop consonant articulation capability of an individual.

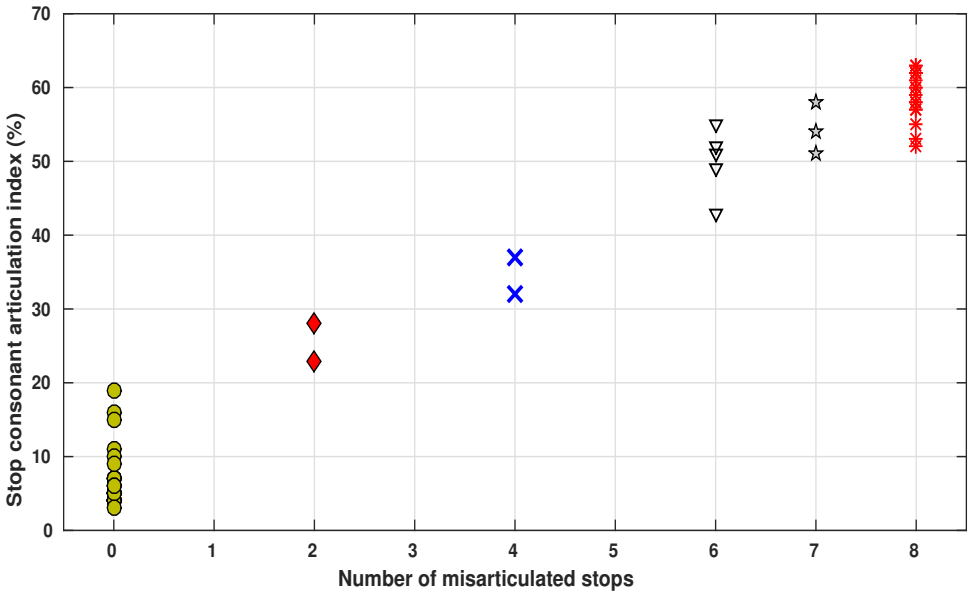

Figure 6.2: Stop consonant articulation index (SCAI) values computed for 30 normal and 31 CLP children. X-axis of graph corresponds to the number of misarticulated stops and y-axis represents the SCAI values in %.

6.3 Multi-class classification systems

The spider plot can be used to assess the presence/absence of misarticulation, but it will not provide information about the category of misarticulation. The misarticulated stops in CLP speech are categorized into weak, nasalized, palatal, velar, pharyngeal, devoiced, and glottal stops. As per the observations from the database (refer Table 4.2), production of a particular category of misarticulation depends on the voicing and place of articulation of target stops. The current database contains very less number of samples for pharyngeal and nasalized stops produced for target unvoiced stops. Based on the availability of speech samples in the current database, there are two different multi-class classification systems developed. The first system is developed for the classification of the normal, glottal, and non-glottal misarticulations produced for the target unvoiced stops. Second system is implemented for the classification of the normal, nasalized, non-nasalized voiced, glottal, and devoiced misarticulations produced for the target voiced stops.

6.3.1 Three-class classification system

Detection of glottal stop gives information about the presence of compensatory articulation. In chapter 4, detection of glottal stops in the presence of non-glottal misarticulations (weak, palatal, and velar) is carried out using CV transition regions. However, during the articulation test, it is necessary

to distinguish the glottal stops from normal as well as other misarticulations. By addressing this issue, a two-stage hierarchical system is developed for the classification of the normal unvoiced stop, glottal, and non-glottal misarticulations. Figure 6.3 shows the proposed hierarchical three-class classification system. Different stages involved in this system are explained as follows.

1. **Stage-1**: This stage aims to classify the subject's response to the target stop as normal or misarticulated. There are four SVM classifies trained for the taget unvoiced stops (/k/, /t/, /T/, and /p/) and corresponding misarticulated stops. While testing, based on *a priori* knowledge of target stop, an appropriate SVM is chosen and the speaker's response is classified as normal or misarticulated.

2. **Stage-2**: This stage consists of an SVM classifier trained for glottal and non-glottal misarticulations. The detected misarticulated stops from stage-1 are passed to stage-2, which categorizes the misarticulations into glottal or non-glottal.

The classifiers in stage-1 and stage-2 are trained for the SPF-based 2D-DCT features computed using manually marked VOPs. Whereas, testing is carried out using the automatically detected VOPs by MWIP-based backward searching algorithm.

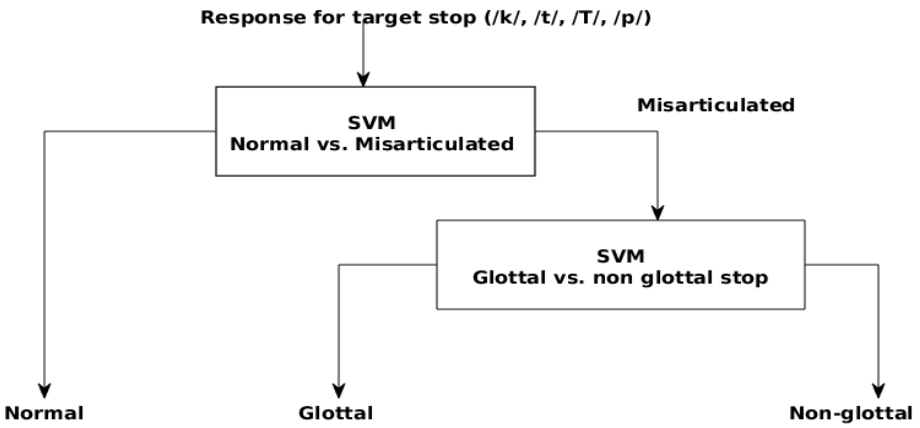

Figure 6.3: Hierarchical approach for the categorization of normal unvoiced stop, glottal and non-glottal misarticulations.

6.3.2 Five-class classification system

A hierarchical system is developed for the classification of normal stop, nasalized, non-nasalized, glottal, and devoicing errors produced for the voiced targets (/g/, /d/, /D/, and /b/). Figures 6.4(a),

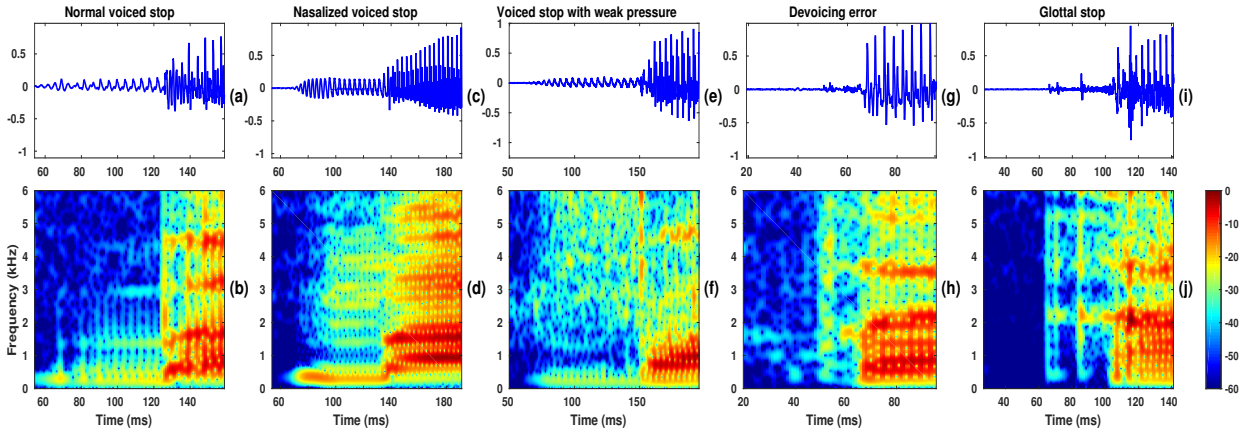

Figure 6.4: Illustration of different misarticulations produced for the target voiced stop in CLP. (a), (c), (e), (g), and (i), represent the waveforms of normal voiced stop, nasalized voiced stop, voiced stop distorted by nasal air emission, devoicing error, and glottal stop, and corresponding spectrograms are shown in (b), (d), (f), (h), and (j), respectively.

(c), (e), (g), and (i), represent the waveforms of normal voiced stop, nasalized voiced stop, weak stop, devoicing error, and glottal stop, corresponding spectrograms are depicted in Figures 6.4(b), (d), (f), (h), and (j), respectively. Based on the voicing, nasalization, and glottalization aspects, the misarticulated stops depicted in Figure 6.4 are discriminated as follows.

- **Voicing:** Misarticulations such as weak, palatal, velar, and nasalized stops are produced with the presence of glottal vibrations. In contrast to these errors, during the production of glottal stops the glottis is completely closed, whereas it is opened for devoiced stops. Due to completely opened or closed glottis, the glottal vibrations are absent for glottal and devoicing errors.

- **Nasalization:** Among the misarticulations produced with the glottal vibrations, nasalization is an important attribute to discriminate the nasalized stops from weak, velar, and palatalization errors. Based on the nasalization, the voiced misarticulations are grouped into nasalized and non-nasalized (weak, palatal, and velar).

- **Glottalization:** Misarticulations produced with the absence of glottal vibrations, i.e., glottal and devoicing errors are mainly distinguishable in terms of place of articulation. In case of glottal stops, the place of articulation lies at the glottis, where it will be somewhere within the oral cavity for devoiced stops.

Based on the voicing, nasalization, and glottalization characteristics, a hierarchical approach is proposed for the classification of normal, nasalized, non-nasalized, glottal, and devoiced stops, which is

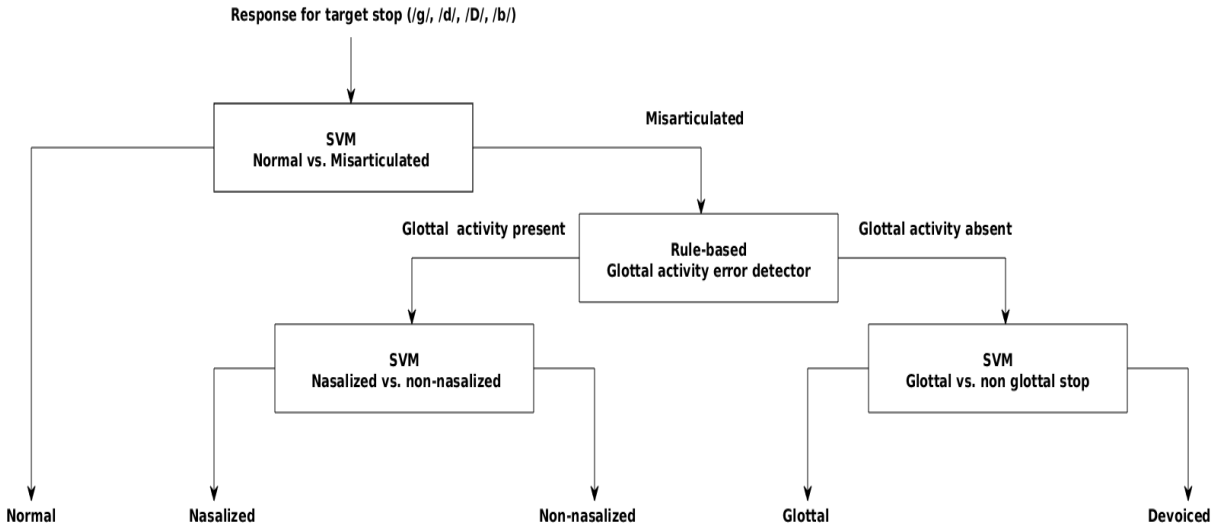

Figure 6.5: Hierarchical approach for the categorization of normal, nasalized, non-nasalized, glottal, and devoiced stops for the target voiced stops (/g/, /d/, /D/, and /b/).

demonstrated in Figure 6.5. Different stages involved in the proposed 5-class classification system for the target voiced stops (/g/, /D/, /d/, and /b/) are explained as follows.

1. **Stage-1:** In this stage, the speaker's response for the target voiced stop is evaluated for the presence or absence of misarticulation. The SVM-based normal and misarticulated stop classifier developed in chapter 4 is used for the categorization of normal and misarticulated stops.

2. **Stage-2:** The misarticulated stops by stage-1 are categorized into consonants with the presence or absence of glottal vibrations. The presence/absence of glottal vibrations is detected using the voiced consonant detection algorithm proposed in the Chapter 5. Figures 6.6 (a)-(d), (e)-(h), and (i)-(l) represent the speech waveform, strength of excitation (SoE) superimposed with glottal activity decision, zero-frequency filtered signal (ZFFS) to band-pass filtered signal (BPFS) energy ratio, and voice consonant decision for the [CVCV] words containing weak, devoiced, and glottal misarticulations, respectively. Due to the absence of quasi-periodic glottal vibrations and low-frequency spectral dominance, the ZFFS to BPFS ratio is not high at the regions corresponding to devoicing error. Hence, the voiced consonant evidence shows the absence of glottal vibrations during the articulatory closure phase. Similarly, Figures 6.6 (g)-(i) indicate the absence of voiced consonant regions for the glottal stop case. A rule-based algorithm is developed for the detection

of presence or absence of glottal vibrations. The algorithm involves the following stages.

(a) Syllable nuclei are detected using the procedure explained in Section 4.3 of Chapter 4. The set of detected syllable nuclei can be represented using $V = \{v_1, v_2, ..., v_N\}$, where N is the total number of detected nuclei.

(b) Based on the ratio of ZFFS to BPFS, the decision of voiced consonant region (d_{vcr})is computed using the equation 5.1 (refer Section 5.3 of Chapter 5).

(c) A search interval t_j defined around the j^{th} syllable nucleus v_j. For word initial syllable v_1, the search interval t_1 is chosen from the beginning of the utterance to the location of syllable nucleus v_1, i.e., $t_j \in [0, v_j], j = 1$. Whereas, for word medial syllable, t_j is chosen as the interval between previous and current syllable nuclei i.e., $t_j \in [v_{j-1}, v_j], j \neq 1$.

(d) Within the search interval t_j, if the evidence d_{vcr} shows the presence of voiced consonant region at least for 30 ms duration, then the consonant associated with the j^{th} syllable (v_j) is decided to be produced with the glottal vibrations.

(e) If the syllable v_j is identified to contain a voiced consonant, then it is considered to be associated with the glottal vibrations, other wise as the absence of glottal vibrations.

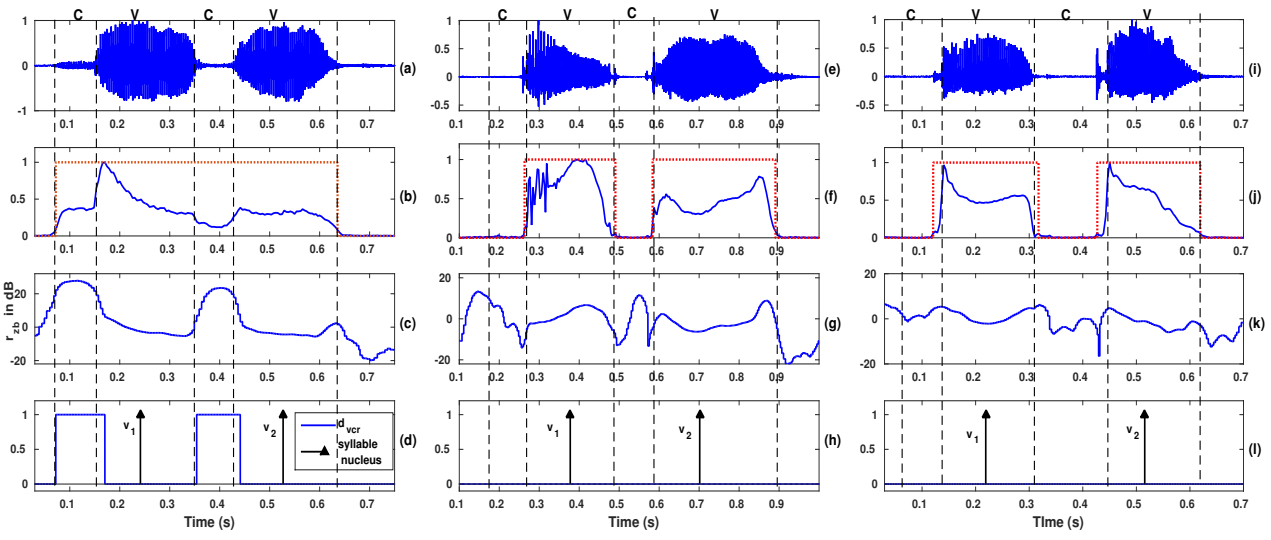

Figure 6.6: Illustration of nature of ratio of ZFFS to BPFS energy ratio. (a)-(d), (e)-(h), and (i)-(l) represent the speech waveform, SoE (blue colored) superimposed with glottal activity decision (red colored dotted line), ZFFS to BPFS energy ratio, and voice consonant decision for the [CVCV] words containing weak, devoiced, and glottal stops, respectively. In (d), (h), and (l), the vertical arrows indicate the positions of detected syllable nuclei.

3. **Stage-3:** Misarticulations with the presence of glottal vibration are categorized into nasalized and non-nasalized misarticulations (weak, palatal, and velar voiced misarticulations). SVM clas-

sifier developed for the classification of nasalized and non-nasalized misarticulations in Chapter 5 is used in stage-3.

3 **Stage-4:** The misarticulated stops categorized with the absence of glottal vibrations are passed to stage-4. The stage-4 aims to classify the misarticulations produced with the lack of glottal vibrations into glottal and devoicing errors. The acoustic characteristics of devoiced stops are similar to that of the unvoiced stops. Hence, the SVM-based glottal and non-glottal misarticulated stops developed in Chapter 4 is used for the classification of glottal and devoiced stops.

The SVM classifiers associated with different stages of five-class classification system are trained using the features computed from manual labels. During testing, the automatically segmented regions are used. In stage-1 and stage-4, the MWIP based backward searching VOP detection algorithm is used. The inputs for the stage-2 is assumed to be voiced consonant. Hence, in stage-2, the voiced consonant and CV transitions segmented using the knowledge-based segmentation algorithm proposed in Chapter 5.

Table 6.1: Database for three-class classification experiments (details of symbols used in the table: B-boys, G-girls, μ-mean and σ-stranded deviation).

Sl. No.	Category	No. of speakers	No. of tokens
1	Normal	30 (14B+16G)	2400
2	Glottal	14 (9B+5G)	690
3	Non-glottal	29 (9B+6G)	1418

Table 6.2: LOSO cross-validated performance of three-class classification system.

	Normal	Glottal	Non-glottal
Normal	**89.7±13.19**	4.54±3.17	8.78±5.67
Glottal	7.26±3.35	**79.45±12.18**	20.34±8.91
Oral	19.51±12.14	21.69±11.15	**68.53±11.45**

6.3.3 Results and discussion

Performance of the classifiers is evaluated using LOSO cross-validation method, which ensures that the training and testing of the classifier is in a speaker independent manner. The results of there-class and five-class classification systems are explained in this section.

Performance of three-class classification system: Details of the database used for the validation of the 3-class classification system are mentioned in Table 6.1. The non-glottal stop class comprised of weak, velar, and palatal stops produced for target unvoiced stops. Confusion matrix mentioned in Table 6.2 represents LOSO-cross-validated performance of three-class classification system. The classifier gives an average accuracy of 79.22%. The confusion matrix shows that the normal stops have high degree of confusion with the non-glottal misarticulations. This confusion is because of the existence of similarity between the normal unvoiced stops and non-glottal misarticulations, where the place of articulation lies within the oral cavity for both groups of sounds. Confusion between the glottal and non-glottal groups may be due to the presence of co-articulated glottal stops in CLP speakers [17]. The co-articulated glottal stops are produced by making the constriction simultaneously at the glottis and oral cavity. Most of the time, an SLP categorizes the co-articulated glottal stops into either glottal or non-glottal group. Therefore, in the current database co-articulated stops may present either in glottal or non-glottal group. Due to such problems in the ground truth, confusion between the glottal and non-glottal misarticulations is noticed. Therefore, furthermore evaluation needs to be conducted to refine the ground truth.

Performance of five-class classification system: Details of the database used for the validation of the 5-class classification system are mentioned in Table 6.3. In Table 6.3, non-nasalized misarticulations corresponds to weak, velar, and palatal stops produced for the target voiced stops.

Table 6.3: Database for five-class classification experiments (details of symbols used in the table: B-boys, G-girls, μ-mean and σ-stranded deviation).

Sl. No.	Category	No. of speakers	No. of tokens
1	Normal	30 (14B+16G)	2400
2	Nasalized	09 (4B+5G)	600
3	Non-nasalized	12 (8B+4G)	908
4	Glottal	10 (8B+1G)	202
5	Devoiced	13 (9B+4G)	296

Performance of 5-class classification system is presented in Table 6.4. The average accuracy of proposed 5-class classifier is equal to 79.38%. Compared to glottal, nasalized, and devoiced stops, the non-nasalized misarticulations have a high-degree of similarity with respect to normal stops. Due to the presence of similarity, 11.67% of the normal voiced stops are falsely detected as non-nasalized misarticulations and 17.22% of non-nasalized misarticulations as normal. There are 14.71% of non-nasalized misarticulations detected as nasalized and 20.20% of nasalized as non-nasalized. This high

Table 6.4: LOSO cross-validated performance of five-class classification system.

	Normal	Nasalized	Non-nasalized	Glottal	Devoiced
Normal	**91.37±8.11**	3.84±2.62	11.67±19.51	2.31±1.3	2.26±1.3
Nasalized	15.1±16.1	**77.9±19.41**	20.2±21.45	3.86±4.54	4.93±10.01
Non-nasalized	17.22±11.49	14.71±14.64	**67.22±21.17**	5.09±3.81	14.74±11.49
Glottal	15±0.1	12.56±16.81	5.21±1.61	**76.67±20.55**	16.94±12.13
Devoiced	8.34±4.64	6.31±4.59	13.41±9.23	14.95±8.3	**77.77±16.78**

degree of confusion between nasalized and non-nasalized misarticulations is because of the presence of VPD. Similar to nasalized stops, the VPD is present in the speakers who produced non-nasalized misarticulations; however, the velopharyngeal gap size is smaller for the non-nasalized case. Due to the small velopharyngeal gap, weak nasal resonances may persist in the non-nasalized misarticulations. But, it may be difficult to identify weak nasal resonances by perceptual evaluation method. Similarly, glottal and devoicing errors show increased confusion due to the presence of co-articulated glottal stops. Therefore, the presence of co-articulated glottal stops and weak nasalized resonances in non-nasalized misarticulations are the major challenges for the multi-class classification of misarticulated stops.

The multi-class classification results showed a high degree of confusion among the various categories of misarticulations. This confusion is due to the presence of similar acoustic characteristics and limitations of perceptual evaluation. Therefore, a reliable ground truth is required for the classification of misarticulated stops. Hence, the augmentation of perceptual evaluation along with the waveform and spectrogram analysis, and instrumental methods may be help to get a reliable ground truth.

6.4 Summary

In this chapter, different event-based modules proposed in the previous chapters are combined to demonstrate their application in the assessment of misarticulated stops in CLP speech. The applications are mainly focused on the screening of misarticulated stops from normal and multi-class classification of misarticulations. Normal vs. misarticulated stop classification results obtained for 8-target stops are combined to construct a spider plot. The spider plot provides a visual interpretation of the presence or absence of misarticulation for each target stop. Therefore, spider plot can be used as a screening tool in the assessment of misarticulated stops. Normal vs. misarticulated stop, glottal activity error detector, nasalized vs. non-nasalized and glottal vs. non-glottal misarticulation

classifiers are combined to develop 3-class and 5-class classification systems. These systems showed the potential of event-synchronous features for the development of multi-class classification systems. The proposed multi-class classification systems detect the glottal, devoiced, and nasalized stops. Hence, the results of multi-class classifation can help to plan the kind of surgery and therapy required to correct the misarticulated stops. Next chapter provides the summary of various contributions of the present thesis.

7

Summary and Conclusions

Contents

7.1 Summary of the work

The objective of the work presented in this thesis is to develop event-synchronous methods for the analysis and detection of misarticulated stops. The epochs and VOP are considered as the significant events for the segmentation of misarticulated regions. To analyze the rapidly varying signal characteristics around these events, SPF-based TFR is utilized. Using SPF-based TFR, an epoch extraction algorithm for the pathological children speech is proposed. Then, based on the knowledge of epochs, VOP detection algorithm is developed for the segmentation of CV transition regions. Using spectro-temporal features extracted from CV transitions, SVM-based system is proposed for the classification of normal and misarticulated stops in CLP speech. Also, the significance of CV transitions is demonstrated for the classification of glottal vs. non-glottal stops, and nasalized vs. non-nasalized misarticulations. The combination of evidence derived from consonant and CV transition regions are used to analyze the effect of nasalization on voiced stops. Further, by combining proposed event-synchronous features, a hierarchical approach for the multi-class classification of misarticulated stops in CLP speech is demonstrated. Some of the important results of the thesis are summarized as follows:

(i) **Epoch extraction using single pole filtering approach:** Accurate detection of epochs is necessary for the epoch-synchronous analysis of speech. In the first work of the thesis, an epoch extraction algorithm is proposed by addressing the issues related to pathological children speech. The algorithm is based on the processing of impulse-like excitation information from the time-frequency domain of speech. SPF based filter-bank approach is used to estimate TFR that offers a better spectro-temporal resolution than STFT. To locate the epochs, SPF-based TFR is processed in two different ways, namely, time marginal and multi-scale product methods. The proposed algorithms are evaluated over the database containing simultaneously recorded speech and EGG signals of pathological children. The identification accuracy of proposed multi-scale product based approach is found to be better than the state-of-the-art methods. In the subsequent chapters, applications of SPF-based TFR and epoch extraction algorithm are demonstrated for detection of misarticulated stops produced by children with CLP.

(ii) **Vowel onset point based processing of misarticulated stops:** The misarticulated stops in CLP speech have deviated in terms of voicing, place, and manner of articulation. Information related to deviant articulation gets reflected in the CV transition regions. The VOPs are

used as the anchor points to segment the CV transition regions. VOPs are detected based on the knowledge of syllable nucleus and epoch-synchronously computed maximum weighted inner product (MWIP) feature. The proposed VOP detection algorithm is compared with state-of-the-art algorithms. Due to the epoch-synchronous processing of speech, the proposed algorithm locates VOPs more accurately than the state-of-the-art VOP detection methods.

The CV transition region anchored around VOP is analyzed using 2D-DCT based joint spectro-temporal features. SPF-based TFR is used to compute the 2D-DCT feature. Further, an SVM classifier is trained for the 2D-DCT features extracted from the CV units containing normal and misarticulated stops. The classification results showed the potential of VOP in the discrimination of various categories of misarticulated stops in CLP speech from normal. The normal vs. misarticulated stop classification results obtained for the proposed VOP detection method are found to be almost comparable with the manually marked VOPs. Whereas, the performance of the classifier for state-of-art-art VOP detection algorithms showed severe degradation due to inaccurately detected VOPs. This result demonstrated the importance of accuracy of the event detection algorithm in event-synchronous processing of misarticulated stops. Also, the proposed SPF-based 2D-DCT feature showed improvement in the classification accuracy over the STFT-based 2D-DCT and MFCC features. This improvement revealed the significance of spectro-temporal resolution of TFR in the analysis of CV transition regions. Subsequently, the glottal vs. non-glottal misarticulations classification system are developed using VOP and SPF-based 2D-DCT features. The classification results indicated the usefulness of VOP event in the discrimination of two different categories of misarticulated stops.

Motivated by the significance of CV transition regions anchored around VOPs is analysed for the detection of misarticulated stops. Initially, to segment the CV transition regions, VOP detection algorithm is developed using the knowledge of syllable nucleus, and epoch-synchronously computed MWIP feature. Due to the epoch-synchronous processing, the proposed VOP detection algorithm detects the VOPs more accurately than state-of-the-art methods. The spectro-temporal dynamics of CV transition regions anchored around VOP are parameterized by 2D-DCT coefficients, which are computed using SPF-based TFR. Using SPF-based 2D-DCT feature, SVM classifier is trained for the classification of normal and misarticulated stops. Since SPF gives better spectro-temporal resolution than STFT, the classifier showed better performance for the

proposed SPF-based 2D-DCT feature than STFT-based 2D-DCT and MFCC features. The performance of the proposed misarticulated stop detection system is tested using both manually marked and automatically detected VOPs. The performance of the normal vs. misarticulated stop classification system for the proposed VOP detection algorithm is near to the manually marked VOPs and better than the state-of-the-art VOP detection methods. Application of the proposed VOP-based framework is extended for the classification of glottal vs. non-glottal misarticulations.

(iii) **Detection of Nasalized Voiced Stops:** In CLP speech assessment, the categorization of articulation error is necessary to analyze the structural and functional deviations associated with speech production system. In this thesis, a detailed analysis of nasalized voiced stops is carried out. Nasalized voiced stops provide a significant information about the presence of moderate-severe VPD. In nasalized voiced stops, the effect of VPD gets reflected on both consonant and CV transition regions. To segment these regions, knowledge-based algorithm is developed using the information of glottal activity, low-frequency spectral dominance, and VOPs. From the segmented regions, spectral, spectro-temporal, excitation source, and periodicity aspects of nasalization are analyzed. Epoch-synchronously estimated spectrum from SPF-based TFR is used to compute spectral and spectro-temporal features. Effect of VPD on the excitation source is analyzed using the peaks of Hilbert envelope of linear prediction residual (HELPR) present around the epochs. The cosine kernel based comparison of successive inter-epoch intervals is carried out to analyze the periodicity characteristics. These features are used to train an SVM-based classifier for the normal and nasalized voiced stops. The classifier showed better performance for spectro-temporal feature than spectral feature. This result indicates the importance of CV transitions over the consonant region in the identification of nasalization. Further, the combination of spectral, spectro-temporal, excitation source, and periodicity break features showed significant improvement than their individual contribution for in detection of nasalized voiced stops.

The consonant and CV transition boundaries estimated by proposed knowledge-based approach are found to be more accurate than HMM-based force-alignment method. Hence, nasalized voiced stop detection system showed better performance for the proposed segmentation algorithm, than HMM-based force-alignment. Also, the proposed epoch-synchronous features

outperformed the conventional block-based MFCCs for the detection of nasalized voiced stops. Subsequently, the efficacy of proposed features is tested for the discrimination of nasalized voiced stops from the non-nasalized misarticulations.

(iv) **Combined framework for the assessment of misarticulated stops:** The event-synchronous systems proposed for the classification of normal vs. misarticulated stop, glottal vs. non-glottal, and nasalized vs. non-nasalized misarticulations are combined to demonstrate their applications in the assessment of misarticulated stops. Normal vs. misarticulated stop classifiers developed for the 8-target stops are combined to construct a spider plot representation. The spider plot provides a visual interpretation about the presence or absence of misarticulation for each target stop. Therefore, spider plot can be used as a screening tool in the assessment of misarticulated stops. Further, using spider plots, a measure called stop consonant articulated index is proposed to evaluate the overall stop consonant production capability of an individual. The normal vs. misarticulated stop, rule-based voiced consonant detector, nasalized vs. non-nasalized, and glottal vs. non-glottal misarticulation classifiers are combined to develop 3-class and 5-class classification systems. The multi-class classifiers are caplable of detecting nasalized, devoicing, and glottal stops, in presence of normal as well as other categories of misarticulations. Hence, these multi-class classification systems can be used for the detection of glottal, devoicing and nasalized misarticulations during the diagnosis of CLP speakers.

7.2 Contributions of this thesis

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

- Single pole filter (SPF) based time-frequency representation is proposed for the spectro-temporal analysis of speech. In particular, SPF is used for the epoch extraction and processing of CV transition regions.
- The time-localized vertical striations present in the SPF-based time-frequency representation are considered as the representative candidates of epochs. This vertical striation information is utilized to develop an epoch extraction algorithm for the pathological children speech.
- An epoch-synchronous feature is proposed for the detection of VOP. The detected VOPs are used to develop a method for the screening of misarticulated stops produced by CLP children

from normal.

- A method for the detection of nasalized voiced stops is proposed using epoch synchronously computed spectral, spectro-temporal, excitation source, and periodicity features.
- A hierarchical framework is developed for the automatic detection of glottal, nasalized, and devoiced stops by combining the proposed event-synchronous features.

7.3 Directions for future work

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

- In the proposed SPF-based epoch extraction algorithm, the epoch locations are primarily detected by zero frequency filtered signal. Further, multi-scale product of SPF-based envelopes is used to modify the accuracy of detected epochs. Zero frequency filtering approach requires *a priori* pitch information for the trend removal operation. However, estimation of pitch is a challenge due to the presence of aperiodic voicing characteristics in pathological speech. The application of peak picking algorithms such as dynamic plosion index on the SPF-based multi-scale product signal may help to overcome the pitch dependency issue of proposed epoch extraction algorithm.
- In the present work, the potential of glottal closure and vowel onset events are explored for the processing of misarticulated stops. In addition to these events, region around the burst onset and vowel offset events also carry the information about misarticulation. Hence, the significance of vowel offset and burst onset events need to be explored. Augmentation of burst and vowel offest events, along with VOP may help to characterize the misarticulated stops in a better way than VOP alone. Also, combination of different events can be used for the detection of errors in CLP such as weak stops, pharyngeal stops, palatal, velar substitutions, etc. which are not addressed in this thesis.
- Proposed VOP detection algorithm assumes the presence of periodicity in the vowel region. Due to presence of creaky phonation, detected VOPs for glottal stops are not found to be accurate. Similar to vowels followed by the glottal stops in CLP speech, vowels in dysarthria and hearing impaired speech are characterized by the presence of pitch perturbation. Hence, using features

related to voice quality measures may help to improve the detection rate of VOPs in presence of aperiodic voicing.

- Nasalization of voiced stops is considered as an important clue in the severity grading of hypernasality in CLP speech. Most of the existing approaches involve the analysis of hypernasality from the vowel spectrum, where it is difficult to detect the weak nasalized cues superimposed on high energy formants of vowel. Production of nasalized stops and normal voiced stops differs only in terms of the velopharyngeal activity and the voice bars have relatively simpler spectral structure than vowels. Therefore, the proposed features for the analysis of nasalized voiced stops can be used for the detection of hypernasality and its severity levels.

- The proposed event-synchronous approach developed for sound treated room data is tested for only one language. Hence, its efficacy and robustness need to be evaluated for various languages and noisy conditions.

- In addition to perceptual evaluation, instrumental methods, such as aerodynamic and articulograph can be used to obtain a reliable ground truth for the development of speech processing algorithms.

- The event-synchronous approach developed in this thesis can be extended for detection of misarticulated stops in other speech sound disorders. The methods can be used directly or with some modifications. Devoicing error is one of most reported misarticulations in the individuals with hearing disorders. The proposed framework for the detection of devoicing errors has an immediate application in hearing disorders. Similarly, SPF-based spectro-temporal features can be used for the detection of backing (/t/ and /k/) and fronting errors (/k/ to /t/) in hearing disorders. The nasalized voiced stop detection method can be used for the assessment of VPD in non-cleft cases such as dysarthria, apraxia, Huntington disorder, etc.

Bibliography

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

[2] B. J. Philips and R. D. Kent, "Acoustic–phonetic descriptions of speech production in speakers with cleft palate and other velopharyngeal disorders," *Speech and Language*, vol. 11, pp. 113 – 168, 1984. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780126086119500085

[3] Y. Kim, H. Shim, and S. Kim, "Articulation and phonological disorders: speech sound disorders in children," *Seoul: Pakhaksa*, 2012.

[4] R. Waring and R. Knight, "How should children with speech sound disorders be classified? a review and critical evaluation of current classification systems," *International Journal of Language & Communication Disorders*, vol. 48, no. 1, pp. 25–40, 2013.

[5] F. E. Gibbon, "Research and practice in developmental phonological disorders," in *Phonology in context*. Springer, 2007, pp. 245–273.

[6] C. M. Combes and J. Martin, "The pronunciation of stops in consonant-vowel-consonant (CVC) words: A comparison between normal and speech-disordered preschool children," *Journal of communication disorders*, vol. 20, no. 4, pp. 281–294, 1987.

[7] A. W. Kummer, *Cleft Palate & Craniofacial Anomalies: Effects on Speech and Resonance*. Cengage Learning, 2013.

[8] F. E. Gibbon, "Undifferentiated lingual gestures in children with articulation/phonological disorders," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 2, pp. 382–397, 1999.

[9] Y. Kim, R. D. Kent, and G. Weismer, "An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 2, pp. 417–429, 2011.

[10] M. Novotny, J. Rusz, R. Cmejla, and E. Ruzicka, "Automatic evaluation of articulatory disorders in parkinson's disease," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 9, pp. 1366–1378, 2014.

[11] M. A. Mines, B. F. Hanson, and J. E. Shoup, "Frequency of occurrence of phonemes in conversational english," *Language and speech*, vol. 21, no. 3, pp. 221–241, 1978.

[12] J. S. Damico, N. Müller, and M. J. Ball, *The handbook of language and speech disorders*. Wiley Online Library, 2010.

[13] American, Speech-Language-Hearing, Association *et al.*, "SLP health care survey: Caseload characteristics," 2011.

[14] D. Sell, "Issues in perceptual speech analysis in cleft palate and related disorders: a review," *International journal of language & communication disorders*, vol. 40, no. 2, pp. 103–121, 2005.

[15] J. E. Trost, "Articulatory additions to the classical description of the speech of persons with cleft palate," *Cleft Palate J*, vol. 18, no. 3, pp. 193–203, 1981.

[16] T. Bressmann, B. Radovanovic, G. V. Kulkarni, P. Klaiman, and D. Fisher, "An ultrasonographic investigation of cleft-type compensatory articulations of voiceless velar stops," *Clinical linguistics & phonetics*, vol. 25, no. 11-12, pp. 1028–1033, 2011.

[17] F. E. Gibbon, "Abnormal patterns of tongue-palate contact in the speech of individuals with cleft palate," *Clinical linguistics & phonetics*, vol. 18, no. 4-5, pp. 285–311, 2004.

[18] P. A. Dagenais, "Electropalatography in the treatment of articulation/phonological disorders," *Journal of Communication Disorders*, vol. 28, no. 4, pp. 303–329, 1995.

[19] M. J. Mcauliffe, E. C. Ward, and B. E. Murdoch, "Speech production in parkinson's disease: I. an electropalatographic investigation of tongue-palate contact patterns," *Clinical linguistics & phonetics*, vol. 20, no. 1, pp. 1–18, 2006.

[20] J. L. Preston, N. Brick, and N. Landi, "Ultrasound biofeedback treatment for persisting childhood apraxia of speech," *American Journal of Speech-Language Pathology*, vol. 22, no. 4, pp. 627–643, 2013.

[21] H. W. Catts and P. J. Jensen, "Speech timing of phonologically disordered children: voicing contrast of initial and final stop consonants," *Journal of Speech, Language, and Hearing Research*, vol. 26, no. 4, pp. 501–510, 1983.

[22] S. Dudy, M. Asgari, and A. Kain, "Pronunciation analysis for children with speech sound disorders," in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 5573–5576.

[23] A. Maier, F. Hönig, T. Bocklet, E. Nöth, F. Stelzle, E. Nkenke, and M. Schuster, "Automatic detection of articulation disorders in children with cleft lip and palate," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2589–2602, 2009.

[24] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, "The goodness of pronunciation algorithm applied to disordered speech," in *Intterspeech*, 2014.

[25] P. Vijayalakshmi, M. R. Reddy, and D. O'Shaughnessy, "Acoustic analysis and detection of hypernasality using a group delay function," *IEEE Transactions on biomedical engineering*, vol. 54, no. 4, pp. 621–629, 2007.

[26] W. Rodríguez, O. Saz, E. Lleida, C. Vaquero, and A. Escartín, "Comunica-tools for speech and language therapy," in *First Workshop on Child, Computer and Interaction*, 2008.

[27] R. Prasad and B. Yegnanarayana, "Acoustic segmentation of speech using zero time liftering (ztl)." 2013.

[28] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996.

[29] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 460–471, 2014.

[30] C.-Y. Lin and H.-C. Wang, "Burst onset landmark detection and its application to speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1253–1264, 2011.

[31] A. Suchato, "Classification of stop consonant place of articulation," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[32] S. R. M. Prasanna, "Event based analysis of speech," *Dept. of Computer Science and Engineering, Ph. D. Thesis: Indian Institute of Technology Madras, India*, 2004.

[33] A. P. Prathosh, "Temporal processing for event-based speech analysis with focus on stop consonants," Ph.D. dissertation, Indian Institute of Science Bangalore-560 012, India.

[34] J. C. Vasquez-Correa, T. Arias-Vergara, J. Orozco-Arroyave, B. M. Eskofier, J. Klucken, and E. Noth, "Multimodal assessment of parkinson's disease: a deep learning approach," *IEEE journal of biomedical and health informatics*, 2018.

[35] 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.

[36] N. Dhananjaya, "Signal processing for excitation-based analysis of acoustic events in speech," Ph.D. dissertation, Ph. D. thesis, Dept. of Computer Science and Engineering IIT Madras, Chennai, URL http://speech. iiit. ac. in/svlpubs/phdthesis/dhanu-phd-2011. pdf (Last viewed Jan. 20, 2012), 2011.

[37] S. H. Dumpala, B. T. Nellore, R. R. Nevali, S. V. Gangashetty, and B. Yegnanarayana, "Robust vowel landmark detection using epoch-based features." 2016.

[38] S. R. M. Prasanna, S. V. Gangashetty, and B. Yegnanarayana, "Significance of vowel onset point for speech analysis," in *Proc. of int. conf. signal processing and communications*, 2001, pp. 81–88.

[39] K. S. Nathan and H. F. Silverman, "Time-varying feature selection and classification of unvoiced stop consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 395–405, 1994.

[40] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Spotting consonant-vowel units in continuous speech using alitoassociative neural networks and support vector machines," in *Proceedings of 14th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, 2004, pp. 401–410.

[41] C. C. Sekhar, "Neural network models for recognition of stop consonant-vowel (scv) segments in continuous speech," *Ph. D. thesis, Dept. of Comput. Sci. and Eng., Indian Inst. of Technol. Madras*, 1996.

[42] K. N. Stevens and D. H. Klatt, "Role of formant transitions in the voiced-voiceless distinction for stops," *The Journal of the Acoustical Society of America*, vol. 55, no. 3, pp. 653–659, 1974.

[43] D. J. Sharf and T. Hemeyer, "Identification of place of consonant articulation from vowel formant transitions," *The Journal of the Acoustical Society of America*, vol. 51, no. 2B, pp. 652–658, 1972.

[44] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, "The role of consonant-vowel transitions in the perception of the stop and nasal consonants." *Psychological Monographs: General and Applied*, vol. 68, no. 8, p. 1, 1954.

[45] P. A. Mossey, E. E. Catilla *et al.*, "Global registry and database on craniofacial anomalies: report of a who registry meeting on craniofacial anomalies," 2003.

[46] A. Harding, R. Razzell, and K. Harland, "Cleft audit protocol for speech," *Eur J Disord Commun*, vol. 31, pp. 331–357, 1996.

[47] G. Henningsson, D. P. Kuehn, D. Sell, T. Sweeney, J. E. Trost-Cardamone, and T. L. Whitehill, "Universal parameters for reporting speech outcomes in individuals with cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 45, no. 1, pp. 1–17, 2008.

[48] S. J. Peterson-Falzone, M. A. Hardin-Jones, and M. P. Karnell, *Cleft palate speech.* Mosby St. Louis, 2001.

[49] S. Witt and S. Young, "Computer-assisted pronunciation teaching based on automatic speech recognition," *Language Teaching and Language Technology Groningen, The Netherlands*, 1997.

[50] I. Laaridh, C. Fredouille, and C. Meunier, "Evaluation of a phone-based anomaly detection approach for dysarthric speech." in *Proc. of Interspeech*, 2016, pp. 223–227.

[51] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," *Computer speech & language*, vol. 50, pp. 62–84, 2018.

[52] S. Strömbergsson, G. Salvi, and D. House, "Acoustic and perceptual evaluation of category goodness of/t/and/k/in typical and misarticulated children's speech," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3422–3435, 2015.

[53] C. Bhat, B. Vachhani, and S. Kopparapu, "Automatic assessment of articulation errors in hindi speech at phone level," in *TENCON 2015*, 2015, pp. 1–4.

[54] G. Van Nuffelen, C. Middag, M. De Bodt, and J.-P. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *International journal of language & communication disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[55] M. Shahin, B. Ahmed, A. Parnandi, V. Karappa, J. McKechnie, K. J. Ballard, and R. Gutierrez-Osuna, "Tabby talks: An automated tool for the assessment of childhood apraxia of speech," *Speech Communication*, vol. 70, pp. 49–64, 2015.

[56] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1998, pp. 433–436.

[57] J.-P. Hosom, "Automatic time alignment of phonemes using acoustic-phonetic information."

[58] A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5552–5556.

[59] V. Karjigi and P. Rao, "Classification of place of articulation in unvoiced stops with spectro-temporal surface modeling," *Speech Communication*, vol. 54, no. 10, 2012.

[60] P. Niyogi and P. Ramesh, "The voicing feature for stop consonants: Recognition experiments with continuously spoken alphabets," *Speech Communication*, vol. 41, no. 2-3, pp. 349–367, 2003.

[61] B. D. Sarma and S. R. M. Prasanna, "Acoustic–phonetic analysis for speech recognition: A review," *IETE Technical Review*, vol. 35, no. 3, pp. 305–327, 2018.

[62] M. Sonderegger and J. Keshet, "Automatic measurement of voice onset time using discriminative structured predictiona)," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3965–3979, 2012.

[63] P. Auzou, C. Ozsancak, R. J. Morris, M. Jan, F. Eustache, and D. Hannequin, "Voice onset time in aphasia, apraxia of speech and dysarthria: a review," *Clinical Linguistics & Phonetics*, vol. 14, no. 2, pp. 131–150, 2000.

[64] N. N. Bitar, *Acoustic analysis and modeling of speech based on phonetic features.* Boston University, 1998.

[65] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, 2000.

[66] J. Hou, L. Rabiner, and S. Dusan, "Automatic speech attribute transcription (asat)-the front end processor," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2006, pp. I333–I336.

[67] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1296–1305, 2004.

[68] P. Niyogi and M. Sondhi, "Detecting stop consonants in continuous speech," *the Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 1063–1076, 2002.

[69] M. Halle, G. W. Hughes, and J.-P. Radley, "Acoustic properties of stop consonants," *The Journal of the Acoustical Society of America*, vol. 29, no. 1, pp. 107–116, 1957.

[70] S. E. Blumstein and K. N. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," *The Journal of the Acoustical Society of America*, vol. 66, no. 4, pp. 1001–1017, 1979.

[71] K. Forrest, G. Weismer, P. Milenkovic, and R. N. Dougall, "Statistical analysis of word-initial voiceless obstruents: preliminary data," *The Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 115–123, 1988.

[72] E. Chodroff and C. Wilson, "Burst spectrum as a cue for the stop voicing contrast in american english," *The Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2762–2772, 2014.

[73] K. Forrest, G. Weismer, M. Elbert, and D. A. Dinnsen, "Spectral analysis of target-appropriate/t/and/k/produced by phonologically disordered and normally articulating children," *Clinical linguistics & phonetics*, vol. 8, no. 4, pp. 267–281, 1994.

[74] D. J. Zajac, L. Cevidanes, S. Shah, and K. L. Haley, "Maxillary arch dimensions and spectral characteristics of children with cleft lip and palate who produce middorsum palatal stops," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 6, pp. 1876–1886, 2012.

[75] M. Eshghi, D. J. Zajac, M. Bijankhan, and M. Shirazi, "Spectral analysis of word-initial alveolar and velar plosives produced by iranian children with cleft lip and palate," *Clinical linguistics & phonetics*, vol. 27, no. 3, pp. 213–219, 2013.

[76] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[77] D. H. Klatt, "Voice onset time, frication, and aspiration in word-initial consonant clusters," *Journal of Speech and Hearing Research*, vol. 18, no. 4, pp. 686–706, 1975.

[78] V. Stouten *et al.*, "Automatic voice onset time estimation from reassignment spectra," *Speech Communication*, vol. 51, no. 12, pp. 1194–1205, 2009.

[79] J. H. Hansen, S. S. Gray, and W. Kim, "Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification," *Speech Communication*, vol. 52, no. 10, pp. 777–789, 2010.

[80] C.-Y. Lin and H.-C. Wang, "Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 514–525, 2011.

[81] A. P. Prathosh, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Estimation of voice-onset time in continuous speech using temporal measures," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. EL122–EL128, 2014.

[82] Y. Adi, J. Keshet, O. Dmitrieva, and M. Goldrick, "Automatic measurement of voice onset time and prevoicing using recurrent neural networks," *Proc. of Interspeech*, pp. 3152–3155, 2016.

[83] A. A. Tyler, G. R. Figurski, and T. Langsdale, "Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 4, pp. 746–759, 1993.

[84] A. P. Prathosh, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Classification of place-of-articulation of stop consonants using temporal analysis," in *Proc. of Interspeech*, 2015.

[85] L. Forner, "Speech segment durations produced by five and six year old speakers with and without cleft palates." *The Cleft palate journal*, vol. 20, no. 3, pp. 185–198, 1983.

[86] D. Montaña, Y. Campos-Roca, and C. J. Pérez, "A diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of parkinsons disease," *Computer methods and programs in biomedicine*, vol. 154, pp. 89–97, 2018.

[87] B. H. Repp, "Perceptual integration and differentiation of spectral cues for intervocalic stop consonants," *Attention, Perception, & Psychophysics*, vol. 24, no. 5, pp. 471–485, 1978.

[88] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 769–773, 1955.

[89] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-hall Englewood Cliffs, NJ, 1978, vol. 100.

[90] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 2, pp. 135–141, 1974.

[91] M. Temkin Martinez and H. K. Boone, "On the presence of voiceless nasalization in apparently effaced somali chizigula prenasalized stops," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 2218–2218, 2016.

[92] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 3, pp. 596–607, 2014.

[93] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5120–5124.

[94] T. T. Wang and T. F. Quatieri, "High-pitch formant estimation by exploiting temporal change of pitch," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 1, pp. 171–186, 2010.

[95] 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.

[96] B. Yegnanarayana and R. N. 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.

[97] H. M. Sussman and J. Shore, "Locus equations as phonetic descriptors of consonantal place of articulation," *Perception & psychophysics*, vol. 58, no. 6, pp. 936–946, 1996.

[98] M. Tabain, "Coarticulation in cv syllables: a comparison of locus equation and epg data," *Journal of Phonetics*, vol. 28, no. 2, pp. 137–159, 2000.

[99] S. Watanabe, K. Arasaki, H. Nagata, and S. Shouji, "Analysis of dysarthria in amyotrophic lateral sclerosis–MRI of the tongue and formant analysis of vowels," *Rinsho shinkeigaku= Clinical neurology*, vol. 34, no. 3, pp. 217–223, 1994.

[100] B. Lindblom, D. Krull, L. Hartelius, and E. Schalling, "Formant transitions in normal and disordered speech: An acoustic measure of articulatory dynamics," in *Proceedings of FONETIK*. Citeseer, 2009, pp. 18–23.

[101] M. Béchet, F. Hirsch, C. Fauth, and R. Sock, "Consonantal space area in children with a cleft palate: An acoustic study," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[102] L. He, J. Zhang, Q. Liu, J. Zhang, H. Yin, and M. Lech, "Automatic detection of glottal stop in cleft palate speech," *Biomedical Signal Processing and Control*, vol. 39, pp. 230–236, 2018.

[103] N. P. Condor, C. L. Ludlow, and G. M. Schulz, "Stop consonant production in isolated and repeated syllables in parkinson's disease," *Neuropsychologia*, vol. 27, no. 6, pp. 829–838, 1989.

[104] D. Kewley-Port, "Time-varying features as correlates of place of articulation in stop consonants," *The Journal of the Acoustical Society of America*, vol. 73, no. 1, pp. 322–335, 1983.

[105] A. Lahiri and S. E. Blumstein, "A reconsideration of acoustic invariance for place of articulation in stop consonants: Evidence from cross-language studies," *The Journal of the Acoustical Society of America*, vol. 70, no. S1, pp. S39–S39, 1981.

[106] Z. B. Nossair and S. A. Zahorian, "Dynamic spectral shape features as acoustic correlates for initial stop consonants," *The Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2978–2991, 1991.

[107] C. C. Sekhar and B. Yegnanarayana, "Recognition of stop-consonant-vowel (SCV) segments in continuous speech using neural network models," *IETE Journal of Research*, vol. 42, no. 4-5, pp. 269–280, 1996.

[108] D. J. Hermes, "Vowel-onset detection," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 866–873, 1990.

[109] S. V. Gangashetty, "Neural network models for recognition of consonant-vowel units of speech in multiple languages," Ph.D. dissertation, PhD thesis, IIT Madras, 2004.

[110] S. R. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation information," in *Ninth European conference on speech communication and technology*, 2005.

[111] S. R. M. Prasanna, B. V. 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.

[112] A. K. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, "Vowel onset point detection for low bit rate coded speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1894–1903, 2012.

[113] A. K. Vuppala and K. S. Rao, "Vowel onset point detection for noisy speech using spectral energy at formant frequencies," *International Journal of Speech Technology*, vol. 16, no. 2, pp. 229–235, 2013.

[114] G. Pradhan and S. R. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 854–867, 2013.

[115] B. D. Sarma, S. S. Prajwal, and S. R. M. Prasanna, "Improved vowel onset and offset points detection using bessel features," in *International Conference on Signal Processing and Communications (SPCOM)*, 2014, pp. 1–6.

[116] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with parkinson's disease." in *Proc. of Interspeech*, 2017, pp. 314–318.

[117] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[118] N. Dhananjaya, S. Rajendran, and B. Yegnanarayana, "Features for automatic detection of voice bars in continuous speech." in *Proc. of Interspeech*, 2008, pp. 1321–1324.

[119] D. W. Warren and S. B. Mackler, "Duration of oral port constriction in normal and cleft palate speech," *Journal of speech and Hearing Research*, vol. 11, no. 2, pp. 391–401, 1968.

[120] B. Hutters and K. Brøndsted, "Strategies in cleft palate speech–with special reference to danish," *Cleft Palate Journal*, vol. 24, no. 2, pp. 126–136, 1987.

[121] G. Henningsson, D. P. Kuehn, D. Sell, T. Sweeney, J. E. Trost-Cardamone, and T. L. Whitehill, "Universal parameters for reporting speech outcomes in individuals with cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 45, no. 1, pp. 1–17, 2008.

[122] L. Nord and G. Ericsson, "Acoustic investigation of cleft palate speech before and after speech therapy," *Journal of STL-QPSR*, vol. 26, no. 4, pp. 15–27, 1985.

[123] F. E. Gibbon and L. Crampin, B, "An electropalatographic investigation of middorsum palatal stops in an adult with repaired cleft palate," *The cleft palate-craniofacial Journal*, vol. 38, no. 2, pp. 96–105, 2001.

[124] A. Harding, R. Razzell, and K. Harland, "Cleft audit protocol for speech," *Eur J Disord Commun*, vol. 31, pp. 331–357, 1996.

[125] A. Butcher and K. Ahmad, "Some acoustic and aerodynamic characteristics of pharyngeal consonants in iraqi arabic," *Phonetica*, vol. 44, no. 3, pp. 156–172, 1987.

[126] D. H. Klatt and K. N. Stevens, "Pharyngeal consonants," *Quarterly Progress Report, Research Laboratory of Electronics, MIT*, vol. 93, pp. 207–216, 1969.

[127] Y. Xiao, Y. Feng, Q. Zhao, L. Ma, J. Qian, and Y. Yan, "Acoustic analysis and detection of glottal stops substituted for alveolar stops in cleft palate speech," *Shengxue Xuebao/Acta Acustica*, vol. 40, no. 2, pp. 285–293, 2015, cited By 1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84927131057&partnerID=40&md5=2a099ff13a50f0af1ebf46739ad2da6a

[128] M. Golabbakhsh, F. Abnavi, M. Kadkhodaei Elyaderani, F. Derakhshandeh, F. Khanlar, P. Rong, and D. P. Kuehn, "Automatic identification of hypernasality in normal and cleft lip and palate patients with acoustic analysis of speech," *The Journal of the Acoustical Society of America*, vol. 141, no. 2, pp. 929–935, 2017.

[129] L. He, J. Zhang, Q. Liu, H. Yin, and M. Lech, "Automatic evaluation of hypernasality and consonant misarticulation in cleft palate speech," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1298–1301, 2014.

[130] K. Forrest, G. Weismer, M. Hodge, D. A. Dinnsen, and M. Elbert, "Statistical analysis of word-initial/k/and/t/produced by normal and phonologically disordered children," *Clinical Linguistics & Phonetics*, vol. 4, no. 4, pp. 327–340, 1990.

[131] M. Eshghi, J. S. Preisser, M. Bijankhan, and D. J. Zajac, "Acoustic-temporal aspects of stop-plosives in the speech of persian-speaking children with cleft lip and palate," *International journal of speech-language pathology*, vol. 19, no. 6, pp. 578–586, 2017.

[132] J. L. Flanagan, *Speech Analysis Synthesis and Perception.* NewYork, NY, USA: Springer.

[133] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 23, no. 6, pp. 562–570, 1975.

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

[135] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[136] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972–980, 2006.

[137] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using lp residual signal," *IEEE Trans. Speech and Audio Process.*, vol. 8, no. 3, pp. 267–281, 2000.

[138] K. S. R. 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.

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

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

[141] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762–765, 2007.

[142] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 1–4.

[143] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 1, pp. 82–91, 2012.

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

[145] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals." in *Proc. of Interspeech*, 2009, pp. 2891–2894.

[146] V. N. Tuan and C. Alessandro, "Robust glottal closure detection using the wavelet transform." in *Proc. of Eurospeech*, 1999.

[147] J. L. Navarro-Mesa, E. Lleida-Solano, and A. Moreno-Bilbao, "A new method for epoch detection based on the cohen's class of time frequency representations," *IEEE Signal Process. Lett.*, vol. 8, no. 8, pp. 225–227, 2001.

[148] R. L. Vikram, K. V. V. Girish, S. Harshavardhan, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Subband analysis of linear prediction residual for the estimation of glottal closure instants," in *IEEE Int. Conf. Acoustics, Speech and Signal Process.* IEEE, 2014, pp. 945–949.

[149] V. R. Lakkavalli, K. V. V. Girish, and A. G. Ramakrishnan, "Sub-band envelope approach to obtain instants of significant excitation in speech," in *Proc. of National Conf. Communications*. IEEE, 2012, pp. 1–5.

[150] 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.

[151] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52–63, 2017.

[152] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A quantitative comparison of glottal closure instant estimation algorithmson a large variety of singing sounds," in *Proc. of Interspeech*, 2013, pp. 1–5.

[153] M. Amin, "On the application of the single-pole filter in recursive power spectrum estimation," *Proceedings of the IEEE*, vol. 75, no. 5, pp. 729–730, May 1987.

[154] M. G. Amin and K. Di Feng, "Short-time fourier transforms using cascade filter structures," *IEEE Trans. Circuits and Systems II: Analog and Digital Signal Process.*, vol. 42, no. 10, pp. 631–641, 1995.

[155] S. Tomazic, "On short-time fourier transform with single-sided exponential window," *Signal processing*, vol. 55, no. 2, pp. 141–148, 1996.

[156] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 4, pp. 705–717, 2015.

[157] D. Gowda, R. Saeidi, and P. Alku, "AM–FM based filter bank analysis for estimation of spectro-temporal envelopes and its application for speaker recognition in noisy reverberant environments," in *Proc. of Interspeech*, 2015.

[158] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 2, pp. 435–444, 2006.

[159] B. Boashash, *Time-frequency signal analysis and processing: a comprehensive reference: Part I.* Academic Press, 2003.

[160] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3795–3806, 1996.

[161] M. Slaney, "Auditory toolbox," *Interval Research Corporation, Tech. Rep*, vol. 10, p. 1998, 1998. [Online]. Available: http://rvl4.ecn.purdue.edu/malcolm/interval/1998010

[162] "G.191 : Software tools for speech and audio coding standardization." [Online]. Available: https://www.itu.int/rec/T-REC-G.191/en

[163] T. F. Website. [Online]. Available: http://festvox.org

[164] "ITU, 1996. recommendation G.712: transmission performance charac- teristics of pulse code modulation channels."

[165] A. W. Black, S. King, and K. Tokuda, "The blizzard challenge 2009," 2009.

[166] B. Woldert-Jokisz, "Saarbruecken voice database," 2007.

[167] M. R. Thomas and P. A. Naylor, "The sigma algorithm: A glottal activity detector for electroglottographic signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1557–1566, 2009.

[168] K. Sjölander and J. Beskow, "Wavesurfer-an open source speech tool," in *Sixth International Conference on Spoken Language Processing*, 2000.

[169] Noisex-92. [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Sectionl/Data/noisex.html

[170] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *The Journal of the Acoustical Society of America*, vol. 64, no. 5, pp. 1358–1368, 1978.

[171] H. M. Sussman, H. A. McCaffrey, and S. A. Matthews, "An investigation of locus equations as a source of relational invariance for stop place categorization," *The Journal of the Acoustical Society of America*, vol. 90, no. 3, pp. 1309–1325, 1991.

[172] S. Ha, H. Sim, M. Zhi, and D. P. Kuehn, "An acoustic study of the temporal characteristics of nasalization in children with and without cleft palate," *The Cleft palate-craniofacial journal*, vol. 41, no. 5, pp. 535–543, 2004.

[173] A. K. Vuppala, K. S. Rao, and S. Chakrabarti, "Spotting and recognition of consonant-vowel units from continuous speech using accurate detection of vowel onset points," *Circuits, Systems, and Signal Processing*, vol. 31, no. 4, pp. 1459–1474, 2012.

[174] R. Smits, "Accuracy of quasistationary analysis of highly dynamic speech signals," *The Journal of the Acoustical Society of America*, vol. 96, no. 6, pp. 3401–3415, 1994.

[175] A. Harding and P. Grunwell, "Characteristics of cleft palate speech," *International Journal of Language & Communication Disorders*, vol. 31, no. 4, pp. 331–357, 1996.

[176] K. Ueda, Y. Nakajima, and Y. Satsukawa, "Effects of frequency-band elimination on syllable identification of japanese noise-vocoded speech: Analyses of confusion matrices," *Proceedings of Fechner Day*, vol. 26, no. 1, pp. 33–38, 2010.

[177] K. B. Guyton, M. J. Sandage, D. Bailey, N. Haak, L. Molt, and A. Plumb, "Acquired velopharyngeal dysfunction: Survey, literature review, and clinical recommendations," *American journal of speech-language pathology*, pp. 1–26, 2018.

[178] J. Van den Berg, "Subglottic pressures and vibrations of the vocal folds," *Folia Phoniatrica et Logopaedica*, vol. 9, no. 2, pp. 65–71, 1957.

[179] S. Furui, "On the role of spectral transition for speech perception," *The Journal of the Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1986.

[180] "HTK toolkit." [Online]. Available: http://htk.eng.cam.ac.uk/

[181] A. W. Kummer, "Disorders of resonance and airflow secondary to cleft palate and/or velopharyngeal dysfunction," in *Seminars in speech and language*, vol. 32, no. 02. © Thieme Medical Publishers, 2011, pp. 141–149.

List of Publications

Journal Publications

1. **C. M. Vikram** and S. R. Mahadeva Prasanna, "Epoch extraction from telephone speech using single pole filter," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 25, no. 3, pp. 624-636, March 2017.

2. **C. M. Vikram**, Nagaraj Adiga, and S. R. Mahadeva Prasanna, "Detection of nasalized voiced stops in cleft palate speech using epoch synchronous features," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 27, no. 7, pp. 1189-1200, July 2019.

3. **C. M. Vikram** and S. R. Mahadeva Prasanna, "Vowel onset point based screening of misarticulated stops in cleft lip and palate speech," IEEE/ACM Transactions on Audio, Speech and Language Processing (under first review).

4. **C. M. Vikram**, Sishir Kalita, and S. R. Mahadeva Prasanna, "Automatic Classification of Misarticulated stops in Cleft Lip and Palate Speech," IEEE Journal of Selected Topics in Signal Processing (under first review).

Conference and Workshop Publications

1. **C. M. Vikram**, S. R. M. Prasanna, "Epoch extraction from pathological children speech," Interspeech 2018, Hyderabad, India.

2. **C. M. Vikram**, S. R. M. Prasanna, Ajish K. Abraham, Pushpavathi M., Girish K. S. "Detection of glottal activity Errors in production of stop consonants in children with cleft lip and palate," Interspeech 2018, Hyderabad, India.

3. **C. M. Vikram**, Pushpavathi M., Ajish K. Abraham, Girish K. S., S. R. M. Prasanna, "Spectro-temporal analysis of stop consonant production errors in children with cleft lip and palate," Workshop on voice, speech, and hearing disorders, 2018, Mysore, India.

4. **C. M. Vikram**, and S. R. M. Prasanna, "Study of Voicebars and Nasals for the Detection of Hypernasality," Workshop on voice, speech, and hearing disorders, 2018, Mysore, India.

Other Publications

1. **C. M. Vikram**, Sashank Kumar Macha, Sishir Kalita, S. R. M. Prasanna, "Acoustic analysis of misarticulated trills in cleft lip and palate children," Journal of Acoustical Society of America-Express Letters, June, 2018.

2. **C. M. Vikram**, Ayush Tripathi, Sishir Kalita, S. R. M. Prasanna, "Estimation of hypernasality scores from cleft lip and palate speech," Interspeech 2018, Hyderabad, India .

3. Ravi Shankar, **C. M. Vikram**. and S. R. Mahadeva Prasanna, "Spoken keyword detection using joint DTW-CNN," Interspeech 2018, Hyderabad, India.

4. Nagaraj Adiga, **C. M. Vikram**, Keerthi Pullela and S. R. M. Prasanna, "Zero frequency filter based analysis of voice disorders," Interspeech, 2017, Stockholm, Sweden.

5. Nikitha K., Sishir Kalita, **C. M. Vikram**, M. Pushpavathi and S. R. M. Prasanna, "Hypernasality Severity Analysis in Cleft Lip and Palate Speech Using Vowel Space Area," Interspeech, 2017, Stockholm, Sweden.

6. Protima Nomo Sudro, **C. M. Vikram**, S. R. M. Prasanna, "Vowel onset point based characterization of velopharyngeal activity using imaging techniques," National Conference on Communication, 2017, IIT Madras, India.

7. **C. M. Vikram** , Nagaraj Adiga, and S. R. M. Prasanna, "Spectral enhancement of cleft lip and palate speech," Interspeech, 2016, San Francisco, California.

8. Ravi Shankar, Arpit Jain, Deepak K. T., **C. M. Vikram**, and S. R. M. Prasanna, "Spoken term detection from continuous speech using ANN posteriors and image processing techniques," National conference on communication, 2016, IIT Guwahati, India.