# Acoustic Analysis of Voice Disorders from Clinical Perspective

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Doctor of Philosophy*
*in*
***Electronics and Communication Engineering***

by

Purva Barche
2018900030
purva.sharma@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology
Hyderabad, India

# CERTIFICATE

It is certified that the work contained in this thesis, titled “Acoustic Analysis of Voice Disorders from Clinical Perspective ” by Purva Barche, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Anil Kumar Vuppala

I would like to dedicate this thesis

To my husband Mr. Prafulla Sharma and my father Mr. Pramod Barche

My guide

Dr. Anil Kumar Vuppala

and

My Family, Teachers, Friends and Well wishers

# Acknowledgments

I am immensely grateful to numerous individuals who have provided invaluable support in making this thesis possible. Foremost, I extend my deepest respect and sincere gratitude to my guide, Dr. Anil Kumar Vuppala, for his guidance and encouragement at every stage of my research work. I am thankful to him for accepting me as his student. It was his lectures in speech signal processing that motivated me to join under him and pursue research in speech processing. His discipline and hard work are the constant source of inspiration for me to continue my journey as research student. No words can adequately express my appreciation for the valuable time and guidance he has bestowed upon me.

I would like to thank my teachers Dr. Suryakanth V. Gangashetty, Dr. Sachin Chaudhari, and Dr. Santosh Nannuru. Their distinct way of teaching and principled approach towards problems have been a source of motivation and energy for me to persist in research. I extend my sincere gratitude to Dr. Chiranjeevi for his invaluable guidance throughout my research work. The feedback provided during the lab meetings has been truly helped me in understanding the technical topics in detail. I am truly grateful for his support and expertise.

I am indebted to the numerous enthusiastic individuals in the Speech Processing Laboratory, both past and present, whose immense contributions have been integral to my journey. I would also like to thank each and every individual without their immense support it was impossible to complete research work. First and foremost, I express my sincere gratitude to Gurugubelli Krishna sir for being my mentor, supporter, and evaluator throughout my Ph.D. journey. Without his guidance, I would not have been able to reach this point of submitting my thesis. I am deeply grateful to him for all the valuable insights, feedback, and constructive criticism that he provided me with. The technical discussions I had with him were truly enjoyable and have contributed significantly to my thorough comprehension of the research problem. His contributions have played a pivotal role in shaping my research career. I am sincerely grateful to him for expertise and the in-depth knowledge he shared with me.

I would like to express my sincere gratitude to Javid for providing invaluable assistance during my course work, particularly in helping me grasp the fundamental concepts of signal processing. Additionally, I want to extend my heartfelt thanks to Ganesh, who has been a constant source of help and support since the second semester of my studies. His support helped me in overcoming numerous obstacles throughout my research journey. Ganesh has always been available to clarify technical concepts and create a friendly atmosphere in the lab, which I deeply appreciate.

I can't express my gratitude towards Tirusha, Her support and guidance have been indispensable in my deeper understanding of SMAI, Python, and ASR. I would like to express my heartfelt gratitude to Vishala ma'am for spending her valuable time with me in guiding during my proposal defense. Her in-depth knowledge shared during our meetings, have been incredibly valuable to me. I would also like to thank my seniors Ravi kumar sir, Vishnu sir, and Ramkrishna sir for providing me guidance.

I am fortunate to have Anuprabha and Priyanka as my friends. The discussions I had with both of them, whether technical or non-technical, have been incredibly fruitful. I truly enjoyed the coffee break with them. I am really thankful to both of them for always supporting me during my tough time. I would like to thank Nayan Vats, Nayan Jha, Hala, Jhansi, Meenakshi, Anshul, Sparsh, Utkarsh, Rashmi, Keshav, and Harsh for creating a very friendly atmosphere within and outside the lab. I cherish the moments of learning and fun we all shared and thank everyone for all the cooperation, understanding and the help I have received from all of them.

I am truly grateful to Zakir (SPCRC) for his continuous support and guidance from the second semester of my joining. Technical discussion with him were so fruitful. I would like to express my sincere gratitude to Sudershan sir, despite never having the opportunity to meet him in person, for his continuous guidance and support he provided me online.

I cherish the special bonding with Ruchi, Sudeepini and Hiranmayi. I feel lucky to have friends like them and really want to thank them enough for their love and moral support. Discussions during our course work and assignments were really helped me in gaining the knowledge. I cannot express my gratitude enough for their love and moral support they gave me during all these years. I would like to express my heartfelt gratitude to T. Sushmitha for her support. It was through her guidance that I became aware of the Ph.D. admission opportunity at IIIT-Hyderabad. I am really grateful to her for the support she has provided me throughout these years.

I would like to extend my sincere appreciation to Krishna sir, Javid, Ganesh, Tirusha, Vishala ma'am, Anuprapha, and Priyanka for generously proofreading my research work and offering invaluable feedback. Their suggestions played a crucial role in enhancing the quality of my work.

Finally I would like to thank my husband Prafulla Sharma for his support, guidance and for bearing with me very patiently all these year. During initial days of research work, he helped in understating many concept about programming language. I cannot express my gratitude for his moral support during my course work. I am grateful for all this wonderful year.

I would also like to express my heartfelt thanks to my kids Rushil and Aanya for their boundless love and patience throughout these years. Their smiles gave me positive energy through these years. I would like to also like to thank my parents (Mr. Pramod and Mrs. Sadhana), my sister (Megha) and my brother (Vaibhav) for their support and love for all these year. Their belief in me helped me to continue the studies thus far. This accomplishment would not have been possible without their support.

# Abstract

Voice disorders are caused due to abnormality in the laryngeal system. The signs and symptoms of voice disorder may include: abnormal pitch (too high pitch, too low pitch, pitch breaks), reduction in loudness, degradation of individual's voice quality (breathy, rough, and strained voice quality), loss of voice and so on. Instrumental assessment, auditory-perceptual assessment and objective assessment are most widely used methods for diagnosing the voice disorders. Instrumental assessment methods often involve the use of laryngoscopes and stroboscopes, but these procedures can be expensive and painful. Auditory-perceptual methods used by Speech-Language Pathologists (SLPs) is considered as a gold standard for detecting voice disorder. The decisions taken in the subjective intelligibility test vary with experience of SLPs, type of scale used, and also depend on the examiner's experience. To address these limitations, objective or automatic assessment methods have been extensively explored in the literature. These approaches extract acoustic features from speech signals, offering reliable, cost-effective, and repeatable assessments. Objective assessment methods have potential to be used as a pre-diagnostic measure for voice disorder assessment by SLPs. This thesis primarily focuses on the objective or automatic assessment methods of voice disorders.

Various objective assessment methods for the automatic detection of voice disorders have been explored in the literature. These methods aim to detect the presence or absence of voice disorders, as well as assess their severity ratings. However, clinical assessment of voice disorders relies on considering the underlying etiological diagnosis. Therefore, this study proposes a clinical approach to assess voice disorders. Along with the detection which was explored in the literature, this thesis explored an objective assessment method which can automatically identify the cause of voice disorders based on the acoustic features extracted from the speech signal. The resulting speech samples are categorized into four distinct categories: structural, neurogenic, functional, and psychogenic. To conduct a comprehensive clinical analysis, a multi-level classification approach is employed. This approach involves training four binary classifiers on acoustic features to achieve a thorough assessment from a clinical perspective.

Voice disorders are characterised by irregularities in the vocal fold vibration, incomplete glottal closure and opening, variation in the amplitude of consecutive opening and closing of the vocal folds. Hence the parameters, which can capture these disturbances in a better way will be able to discriminate the voice disorders from healthy samples. From the source-filter model of speech production these features can be captured in a better way from excitation source signals. Glottal flow waveform, zero frequency filtered (ZFF) signal and linear prediction (LP) residual signals are some evidence of exci-

tation source signal. Features derived from these evidences were used to capture the characteristics of voice disorders. First study explores perturbation (jitter, shimmer, noise to harmonic ratios etc.) and cepstral features derived from the excitation source evidence for detection and identification of voice disorders. In this regard state-of-art speech signal processing techniques, such as quasi-closed-phase (QCP) analysis, LP analysis and ZFF techniques, have been explored in this thesis in order to capture the excitation source information. From this study, it was concluded that perturbation parameters can capture voice disorder information in a better way. In addition it was also found that excitation source based features can discriminate between the organic voice disorder from non-organic voice disorder, as well as structural voice disorders from the neurogenic voice disorder category. However, distinguishing functional voice disorders from psychogenic voice disorders proved to be challenging in the study.

From the first study, it was found that excitation source based features are able to differentiate the various categories of voice disorders. Computation of these features involves the detection of epoch locations from speech. Therefore, accurate estimation of epoch locations is important for computing these features for the automatic detection and identification of voice disorders. Second study aimed to compare the various algorithms for detecting epoch locations from the speech associated with voice disorders. In this regard, nine state-of-the-art epoch extraction algorithms were considered, and their performance for different categories of voice disorders was evaluated. From the results it can be concluded that most of the epoch extraction methods showed better performance for healthy speech; however, their performance was degraded for speech associated with voice disorders. Furthermore, the performance of epoch extraction methods was degraded for the speech of structural and neurogenic disorders compared to the speech of psychogenic and functional disorders. This degradation in performance might be due to rapid change in fundamental frequency (F0) associated with subjects suffering with voice disorders as compared to healthy subjects. Some of the state-of-the-art epoch extraction methods depend on the average value of F0 for computation of epochs, hence if for these methods F0 is derived for each region for calculation of epoch locations then identified epoch locations might be more accurate. With this motivation to improve the performance, application of region-based processing as a pre-processing step on the state-of-the-art epoch extraction method was proposed for voice disorder scenarios. Results of this study showed that performance was improved for voice disorder scenarios with the application of region-based processing to state-of-the-art epoch extraction techniques which might be due to local F0 being used to estimate the epoch locations as compared to average F0 used in the state-of-the art epoch extraction algorithms. Moreover, to improve the performance of the voice disorder detection and identification system, the system was built using the features extracted by applying the region-wise processing to the state-of-the-art epoch extraction algorithm. From this study it was found that performance is improved as compared to the baseline features leading to the conclusion that the accurate identification of epoch locations plays an important role in case of voice disorder detection and identification.

Previous studies have revealed that features obtained from the excitation source signal can effectively distinguish between various categories of voice disorders. However, their effectiveness relies on the precise estimation of fundamental frequency and accurate epoch location. Detecting the pitch con-

tour is more straightforward in mild dysphonic voices compared to severely affected ones. Additionally, it has been observed that careful consideration should be given to the type of signal, gender, and fundamental frequency when calculating these features. Hence the following study in this thesis focused on the supra-segmental analysis (speech analysis with a frame size greater than 100 ms) of speech signal instead of short-term analysis (frame size of 20 ms) used in the previous study. Voice disorders affect the pitch, loudness, and voice quality, which are perceived at the supra-segmental level in the speech signal. To capture the voice quality feature, we explored the effectiveness of long term average spectrum (LTAS) features. For the detection and identification of voice disorders, this study explores the effectiveness of LTAS features using auditory filter banks like gammatone and Constant-Q. The performance of the system is also compared with LTAS features derived from critical band filter bank and single frequency filter (SFF) based filter bank. From the results it was observed that performance of the detection and identification system is improved using the gammatone and constant-Q based LTAS features as compared to the baseline features. The reason for improvement might be due to auditory filter banks which were designed to mimic the human auditory system. Compared to our previous study, significant improvement of performance for all the experiments was observed which might be due to the reason that long term features can capture the voice disorders information in a better way as compared to the features extracted using short-term analysis methods.

The previous study concluded the importance of spectral-temporal domain analysis for the voice disorder detection and identification system. Stockwell-Transform (S-Transform) is a time-frequency analysis method which provides better time-frequency localization as compared to other representations like short-time Fourier transform (STFT), wavelet-transform, etc. Therefore, S-Transform was explored for the classification of voice disorders from a clinical perspective. We proposed cepstral features derived from S-Transform for building the detection and identification system for assessing voice disorders. Additionally, we demonstrated the effectiveness of using the S-Transform method for capturing the acoustic characteristics of various voice qualities. As compared to baseline features, proposed features performed best in terms of classification accuracy for voice disorder detection task. Also, the proposed features performed better in case of identification tasks. Further, the experimental results reveal that the combination of cepstral coefficients derived from S-Transform with baseline features improved the performance of proposed systems by 8% and 4% for detection and identification task, respectively.

**Keywords:** *Clinical perspective, Voice disorders, Detection and identification of voice disorders, Excitation source features, Region-wise processing, Supra-segmental analysis, Long term average spectrum, Time-frequency analysis, Stockwell-Transform.*

# Contents

# List of Figures

# List of Tables

# Abbreviations

**List of Abbreviations:**

| | | |
|---|---|---|
| AME | – | Attenuated main excitation |
| ANOVA | – | N-way analysis of variances |
| ASHA | – | American speech language hearing association |
| CAPE-V | – | Consensus Auditory-Perceptual Evaluation-Voice |
| ComParE | – | The 2013 Interspeech Computational Paralinguistics Challenge |
| CQ | – | Close quotient |
| dEGG | – | Differenced Electroglottograph |
| DYPSA | – | Dynamic programming phase slope algorithm |
| eGeMAPS | – | extended Geneva Minimalistic Acoustic Parameter Set |
| EGG | – | Electroglottograph |
| FPR | – | False positive rate |
| FT | – | Fourier Transform |
| FVD | – | Functional voice disorder |
| GCIs | – | Glottal closure instants |
| GIF | – | Glottal inverse filtering |
| GVV | – | Glottal volume velocity |
| HNR | – | Harmonic to noise ratios |
| HUPA | – | Hospital Universitario Prıncipe de Asturias |
| IDA | – | Identification accuracy |
| IR | – | Identification rate |
| LP | – | Linear prediction |
| LR | – | Laryngeal airway resistance |
| LTAS | – | Long term average spectrum |
| MFCC | – | Mel frequency cepstral coefficients |
| MFCC-residual | – | MFCC of LP residual |
| MFCC-ZFF | – | MFCC of Zero frequency filtered signal |
| MR | – | Miss rate |
| NHR | – | Noise to harmonic ratio |
| NNE | – | Normalized noise energy |
| OpenSMILE | – | Open-source Speech and Music Interpretation by Large-space Extraction |
| PLP | – | Perceptual linear prediction |
| PSD | – | Power spectral density |
| PSP | – | Parabolic spectral parameters |
| PTP | – | Phonation threshold pressure |
| PVD | – | Psychogenic voice disorder |
| QCP | – | Quasi closed phase |
| RLNP | – | Recurrent laryngeal nerve palsy |
| SD | – | Structural voice disorder |
| SEDREAMS | – | Speech event detection using the residual excitation and a mean-based signal |

| | | |
|---|---|---|
| SFFB | – | Single frequency filter bank |
| SLPs | – | Speech language pathologists |
| SNR | – | Signal-to-noise ratio |
| SoE | – | Strength of excitation |
| STCC | – | Stockwell Transform cepstral coefficients |
| STFT | – | Short-time Fourier transform |
| SVD | – | Saarbruecken voice disorder |
| SVM | – | Support vector machine |
| VHI | – | Voice Handicap Index |
| VRQOL | – | Voice-Related Quality of Life |
| WLP | – | Weighted linear prediction |
| YAGA | – | Yet another GCI algorithm |
| ZFF | – | Zero frequency filter |
| ZP-ZFF | – | Zero-phase zero frequency filter |
| ZTW | – | Zero-time windowing |

*Chapter 1*

# Introduction

Speech is a natural way of communication used by human beings. It contains linguistic information like message and paralinguistic information like feelings, speaker's health, and speaker traits like gender, age, and personality. Speech production is a complex process. It requires coordination and control of five sub-systems: respiratory, laryngeal, articulatory, resonatory, and nervous systems [1, 2, 3, 4, 5]. Proper functioning of these sub-systems results in healthy speech. Abnormality in any of the sub-systems, results in disordered speech. Different categories of speech disorders include articulation, phonological, resonance, fluency, and voice disorders. Voice disorders are relevant to the interests of this thesis. Some professions, such as teachers, class instructors, factory workers, singers, have an excessive demand to use their voice, which can lead to degradation in voice quality. These professions are at high risk of developing voice disorders [6]. According to the National Institute on Deafness and Other Communication Disorders (NIDCD), approximately 7.5 million people in the United State (US) are suffering from voice disorder problems [7].

Voice disorders are caused due to abnormality in the laryngeal sub-system impacting the individual's ability to speak normally [8, 9]. The most common voice disorders include laryngitis, cyst, polyp, vocal cord paralysis, and recurrent laryngeal nerve palsy. The signs and symptoms of voice disorder may include: abnormal pitch (too high pitch, too low pitch, pitch breaks), reduction in loudness, degradation of individual's voice quality (breathy, rough, and strained voice quality), loss of voice [1, 10, 11] and so on. These problems arise when the vocal folds do not vibrate normally due to structural or functional abnormalities. Speech language pathologists (SLPs) diagnose voice disorders by conducting a comprehensive evaluation of an individual's voice, which includes assessing various aspects such as the pronunciation of constant and varying pitch vowels, sentences, breathing, vocal cord movement, and overall vocal quality. Voice disorder assessment methods can be broadly categorized as invasive or non-invasive. Invasive methods involve using a laryngoscope to examine the movements of the vocal cords for detecting the underlying cause of voice disorder but they are painful and costly. On the other hand, non-invasive methods that utilize acoustic information have gained significant attention. These approaches employ perceptual and objective assessment approaches to identify voice disorders. Although perceptual assessment ( which relies on listening the subject) is considered a reliable measure

for the assessment of voice disorder, it is subjective. Objective assessment methods ( which rely on analysing acoustic features) are effective, and require less time. Moreover, the acoustic features used in these methods are highly correlated to perceptual measures, so these methods are most widely explored for voice disorder detection [12, 13, 14]. These methods are used as viable techniques, as they have potential to provide relevant and perceptually correlated information about pathological speech [15]. From a clinical perspective, they can be used as an early diagnosis tool to detect the presence of pathology.

In literature, objective assessment methods have explored different machine learning algorithms and various signal processing techniques for voice disorder detection from speech. Due to advancements in deep learning, researchers have explored different architectures such as convolution neural networks (CNNs) [16, 17, 18], multi-layer perceptron (MLP) [19], long short-term memory (LSTM) [20, 21] for automatic voice disorder detection. In [22], combination of CNN with LSTM and MLP was explored for the detection of voice disorders. From various studies it can be understood that deep learning architectures require a huge amount of data for training the network. Hence, deep learning methods may not be suitable for developing pathological speech processing applications where the amount of data is small [23]. Therefore, various objective assessment methods use classical machine learning algorithms and they have been exploring different signal processing techniques to get the best feature representation for detecting voice disorders.

From an auditory-perceptual perspective, jitter and shimmer contribute to a rough perceptual effect (namely harshness); hence these perturbation measures were used in the literature to detect voice disorders [24, 25]. Jitter and shimmer model variation in the period and amplitude between the consecutive glottal cycle, respectively. Uncontrolled or irregular movement of vocal folds leads to a higher value of jitter and shimmer. Perceptual correlated information associated with jitter is roughness [26] while with shimmer it is breathiness. Different variations of jitter [27, 28, 29] and shimmer [26, 30, 31, 32] were used in the literature for automatic detection of voice disorders. Other popular measures to detect the presence of voice disorders are harmonic to noise ratio (HNR) [33, 34, 35, 36], signal to noise ratio (SNR) [37, 38], and glottal to noise excitation (GNE) [12, 39]. The physiological process of vocal fold vibration is represented by the glottal volume velocity (GVV) signal. Irregular vocal fold vibrations cause variation in the shape of the GVV signal. This time domain change in the GVV signal is also reflected in the frequency domain. Features like open quotient (OQ), closing quotient (ClQ), speed quotient (SQ), and Quasi-open quotient (QoQ) were the most widely used time duration ratios explored in the literature to detect the presence of voice disorders [40, 41]. Frequency-domain features like the difference between first and second harmonics (H1-H2) and harmonic richness factor (HRF) and parabolic spectral parameter (PSP) derived from GVV signal, were also used in the literature to discriminate voice disorders [40, 41]. In [42], different glottal signal parameters were explored to detect the vocal fold pathologies, namely nodules and unilateral paralysis. Cepstral peak prominence (CPP) [43, 44, 45, 46, 47, 48] was also used as a reliable measure for differentiating the disordered voice from the healthy voice. Even though most voice disorders affect the functioning and structure of the larynx, vocal tract features were also explored in the literature for discrimination of voice pathologies. Fea-

tures that capture vocal tract characteristics like mel frequency cepstral coefficients (MFCC) [49, 50], linear prediction cepstral coefficients (LPCC) [49], perceptual linear prediction coefficient (PLP) [51], and constant-Q cepstral coefficient (CQCC) [52] were also used in voice disorder detection.

In all the above mentioned approaches used in the literature, voice disorder detection was seen as two class problem which discriminates pathological voice from healthy voice. On the other hand, clinicians examine voice disorder in a different way. First, they detect the presence of voice disorder; later, they perform differential diagnosis to identify the type of voice disorder such as structural, neurogenic, functional or psychogenic [53]. This thesis focuses on automatic detection and identification method of voice disorders from a clinical perspective. To achieve this, we employed a multi-level classification approach that involved four binary classifiers for assessing voice disorders. The first classification step involved differentiating between healthy voices and voices with disorders. Subsequently, the voice disorder category was further classified into two classes: organic and non-organic. Organic disorders were further categorized as either structural or neurogenic, while non-organic disorders were classified as functional or psychogenic. This multi-level classification approach allowed us to comprehensively classify voice disorders based on their underlying causes, providing a more in-depth understanding of the types of disorders.

Voice disorders are often characterized by noticeable fluctuations in both amplitude and frequency during consecutive opening and closing of the vocal folds. Hence, the features derived from excitation source evidence like linear prediction (LP) residual [40, 54, 55], zero frequency filter (ZFF) signal [40, 56, 57, 58, 59], and glottal volume velocity (GVV) [60, 61, 62, 63, 64] are used in the thesis to study the importance of excitation source signal for different categories of the voice disorders. The accuracy of features derived from excitation source depends on the glottal closure instants (GCIs), also known as epoch. Therefore, the performance of the state-of-art epoch extraction methods is compared for different categories of the voice disorders. Region-based processing was applied to state-of-the-art epoch extraction methods to improve their performance in voice disorder scenario. Voice disorders affect the pitch, loudness, and voice quality, which are perceived at the suprasegmental level in the speech signal [1]. In this regard, to capture the feature related to voice disorders, long-term average spectrum (LTAS)-based features were also explored in this thesis for the detection and identification of voice disorders.

The presence of voice disorders can lead to a degradation in the acoustic characteristics of affected individuals, which can be observed as variations in the spectro-temporal domain. In the literature, various time-frequency representation methods were investigated for automatic detection of voice disorders. Stockwell-Transform (S-Transform) is a time-frequency analysis method which can localize information in both the time and frequency domains effectively. With this motivation, we investigated S-Transform for the automatic detection of voice disorders.

## 1.1 Objective and scope of the thesis

The primary objective of this thesis is to analyse the importance of acoustic features for the automatic detection and identification of voice disorder from a clinical perspective. Clinical perspective analysis may help SLPs to use this acoustic analysis as a pre-diagnosis tool in identifying voice disorders. To accomplish this analysis, a multi-level classification approach is used in which four binary classifiers were trained on the acoustic features. Pitch, loudness [10, 15, 65], and voice quality [11, 15] are some of the main acoustic characteristics affecting subjects suffering from voice disorders. Hence, the feature which captures these dimensions are explored in this thesis. The scope of the thesis is summarized as follows:

- Voice disorders tend to change the phonation (vocal fold vibration) characteristic, which in turn can be effectively captured by excitation source signal. Hence, thesis explores features derived from the excitation source evidences like ZFF, GVV derived from quasi closed phase (QCP) analysis method, and LP residual signal derived from LP analysis method for detection and identification of voice disorders.

- Features derived from the excitation source depends on accurate detection of epoch location, hence, the performance of state-of-the-art epoch extraction methods was compared for voice disorders scenario. Moreover to improve the performance, region-wise processing was applied to the state-of-the-art epoch extraction for voice disorder scenario.

- The perceptual methods were considered as golden standard in identifying the voice pathology. The voice quality like breathiness, roughness, loudness, and intonation from the speech signal are perceived in the long term [66]. Hence these features can be captured by LTAS. To capture the voice quality feature, we explored the effectiveness of LTAS features using four state-of-the-art filter banks designed with critical-band [67, 68], constant-Q, gammatone, and single-frequency filtering (SFF) [69] approaches for detection and identification of voice disorder.

- Individuals with voice disorders experience degradation in their acoustic characteristics, such as pitch, voice quality, and loudness, in comparison to those with healthy voices. These alterations in acoustic features manifest as variations in the spectro-temporal domain. In order to capture these characteristic S-Transform based cepstral features were also explored for detection and identification of voice disorder.

In nutshell, the main contribution of this thesis is analysis and detection of voice disorders from clinical perspective which in turn helps in knowing the category of voice disorders. Hierarchical approach was used in order to build the voice disorder system for detection and identification. Various acoustic features like excitation source, and long term average spectrum features from the speech signals are explored to perform the experiments. We also proposed extraction of cepstral features derived from S-Transform for performing the voice disorder detection and identification task.

## 1.2 Organisation of the thesis

The thesis is organised as follows:

- Chapter 2 presents the overview of voice disorder, and its different assessment methods. It also gives an overview of the speech signal processing methods used in the literature used for automatic detection of voice disorder.
- Chapter 3 explored the different features derived from excitation source signals for developing the automatic system for detecting voice disorders. Additionally, the chapter also investigated the identification of voice disorder in clinical way using the various excitation source based features.
- Chapter 4 explored the application of region-based processing for state-of-the-art epoch extraction method for performing the detection and identification of voice disorders.
- Chapter 5 presents the LTAS features derived from the auditory filter banks like gammatone and constant-Q for automatic detection of voice disorders in clinical way.
- Chapter 6 proposed the features derived from the S-Transform for building the system for automatic detection and identification of voice disorders. Different variants of S-Transform are explored for analysing voice qualities (such as breathiness, harshness, creakiness etc.) associated with voice disorders.

*Chapter 2*

# Background and literature review

Speech production requires airflow from the lungs to be phonated through vocal folds of the larynx and resonated in the vocal cavities shaped by the tongue, jaw, soft palate, lips, and other articulators. Phonation is a process by which the vocal folds produce sounds through quasi-periodic vibration, also known as voicing. Any abnormality in the larynx that affects voicing in speech production is referred to as voice disorder. From an auditory-perceptual point of view, voice disorders affect voice quality, pitch, and loudness [11]. This chapter covers the review of speech production mechanisms and an overview of existing literature related to automatic detection of voice disorders.

The rest of the chapter is organized as follows. Section 2.1 briefly discusses the speech production mechanism. Section 2.2 describes the phonation process and various types of phonation. Section 2.3 gives the basic definition of speech disorder. Voice disorders and their classification based on etiology are described in section 2.4. Various methods used to assess voice disorder are explained in section 2.5. Section 2.6 discusses the significant gaps identified from the literature. Section 2.7 briefly discusses the existing database for performing voice disorder detection. The conclusion and summary of the chapter are presented in section 2.8.

## 2.1 Anatomy and physiology of speech production

Speech production is a complex process and involves the control and coordination of many sub-systems. From the physiological point of view, the speech production system is subdivided into three main systems: subglottal, glottal (larynx), and supralaryngeal system [70, 71] as illustrated in Figure 2.1. Speech is produced when air is exhaled out from the lungs via the trachea. The subglottal system provides airflow to the glottal system. The larynx modulates airflow from the lungs and provides either quasi-periodic or noisy pulses to the supra-laryngeal system. The supralaryngeal system consists of the pharynx, oral and nasal cavities, and further shapes (or filters) the spectrum of the airflow. The resulting signal is radiated by the lips.

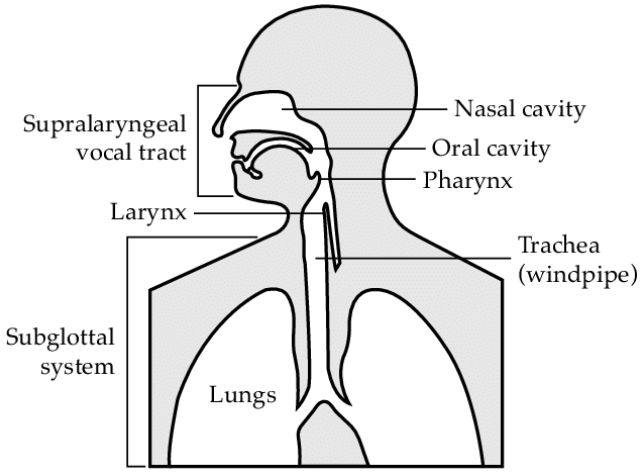



Figure 2.1: Speech production system. (After Lieberman 1992, [134].)

### 2.1.1 Subglottal system

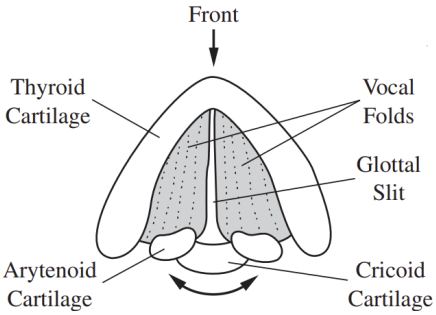
***Subglottal system*** consists of lungs, ribcage, chest muscles, diaphragm, and trachea. Lungs act as a power supply and its main function is to facilitate the respiration process [70]. Respiration cycle includes one inspiration and one expiration. Adults typically complete 12 to 18 respiration cycles per minute during normal breathing. During these cycles, inspiration takes up about 40%, while expiration takes up 60% of the respiration cycle. However, during speaking, the proportions are different, with inspiration taking up only 10% and expiration taking up 90% of the respiratory cycle. When speaking, the lungs are filled to approximately 48% of their vital capacity (VC) and a breath is taken when they reach a level just below the resting lung volume, at around 35% of VC. The loudness and pitch of sound can be varied by changing the glottal airflow and lung pressure (subglottal pressure).

### 2.1.2 Larynx

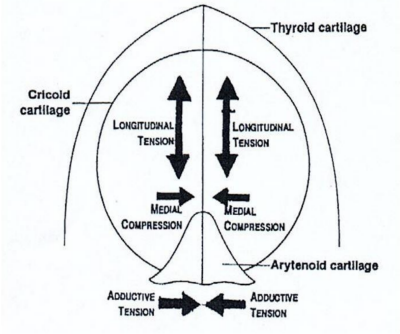
The larynx, commonly known as the voice box situated at the top of the trachea and below the pharynx. It is made up of cartilage, ligaments, and muscles. The vocal folds are located at the top of the larynx and have a V-shape appearance when viewed from the top. The front and side part of the vocal folds is attached to the stationary thyroid cartilage. Hence front part of the vocal folds can not move. Vocal folds are free to move at the the back and sides of the larynx, connected to arytenoid and cricoid cartilages. The area between the two vocal folds is called the glottis [70, 72]. The recurrent laryngeal nerve and the superior laryngeal nerve perform the muscle control of the larynx. The larynx has three main functions: protection, respiration, and speech production. Epiglottis is located at the root of the tongue and provides the protection to the trachea against unwanted substances.

Proper laryngeal adjustments, such as longitudinal tension, adduction/abduction tension, and medial compression [73] (as shown in Figure 2.2 (b)), can control the movement of the vocal folds which in turn

impact the pitch of our voice. These tensions also determine the state of the vocal folds (phonation type). In speech production, vocal folds can be either in two states: voiced and unvoiced. In the voiced state, vocal folds are tensed, which causes self-sustained oscillation of vocal folds [70]. In unvoiced state, vocal folds are relaxed, which allows the airflow to continue through the vocal tract until it is blocked by articulators of the vocal tract. Fig 2.2 (a) and (b) shows the top view of the larynx and laryngeal adjustments for producing the different phonation, respectively. Adductive tension is responsible for bringing the arytenoid cartilages together. For certain sounds, such as the glottal stop, a high degree of adductive tension is necessary to fully close the vocal folds and create a complete obstruction of airflow. On the other hand, for voiced sounds, a lower value of adductive tension is required to allow the vocal folds to vibrate freely and produce sound. Medial compression controls the closing and opening of the glottis. Longitudinal tension is important in regulating the tension along the length of the vocal folds. By adjusting this tension, the pitch of speech sounds can be varied. A larger value of this tension lengthened the vocal folds, which in turn resulted in a higher frequency of vibration.




(a) Top view of the Larynx [70].




(b) Laryngeal adjustments for phonation [70].

Figure 2.2: Larynx.

### 2.1.3 Supralarangeal system

The supralarangeal system is comprised of the pharynx, oral, and nasal cavities.The airflow that comes from the larynx is further modified by the vocal tract system to produce different speech sounds. Different sounds can be produced by altering the vocal tract's length and shape through articulators' movement. Articulators such as the tongue, lips, jaw, and soft palate are movable, while the alveolar ridge, hard palate, and teeth are fixed. The average length of the vocal tract for adult males and females is 17 cm and 15 cm, respectively.

## 2.2 Phonation

Phonation refers to the state of the vocal folds [74]. In general, two states of vocal folds are possible: relaxed and tensed vocal folds. In the relaxed state, vocal folds are far apart to vibrate but close enough to

cause turbulence of airflow, resulting in an unvoiced phonation. In the tensed state, arytenoid cartilages move towards one another, partially closing the vocal folds. This partial closing of the glottis and increased vocal fold tensions result in the oscillation of vocal folds. The oscillatory vocal folds convert the expiratory airflow into intermittent airflow pulses which result in a buzzing sound or voice phonation. The complete process of vocal fold vibration can be described as follows. At the starting phase of the phonation cycle, vocal folds are closed (Figure 2.3A). During exhalation, air comes out of the lungs, increasing pressure in the sub-glottal system (system below the glottis). When this pressure is stronger than the muscle tension of vocal folds, it causes the lower part to open, followed by the upper part (Figure 2.3 B-D). Once the vocal folds are opened completely (Figure 2.3 E), the air will pass from the sub-glottal system to the glottal system [71]. From Bernoulli's principle, it can be inferred that as the speed of the air increases, the pressure in the glottis decreases. This decrease in pressure causes the vocal folds to come together again (Figure 2.3 F-G), which occurs in a single glottal cycle. The rate of vibration of the vocal folds is referred to as the fundamental frequency (F0). When the vocal folds come together, it is called adduction, whereas if they are apart from each other, this is called abduction. Phonations such as modal, creaky, breathy, harsh, and falsetto are voiced phonations, while whisper is an unvoiced or voiceless phonation [75].

### 2.2.1 Modal phonation

Modal phonation is a neutral mode of phonation in which vocal folds vibrate normally with the vocal folds fully adducted such that there is no air leakage through the glottis during the closed phase of the glottal cycle. Vocal folds have moderate longitudinal tension, medial compression, and adductive tension, producing quasi-periodic vibrations [76]. It is used as reference phonation to compare all other phonations.

### 2.2.2 Creaky phonation

In creaky phonation, vocal folds are adducted with weak longitudinal tension which causes the thickening of the vocal fold. Additionally, the inferior surfaces of the false folds may sometimes come in contact with the superior surfaces of the true vocal folds creating an unusually thick and slack structure before the initiation of phonation. These laryngeal settings result in heavy vibrating mass which in turn causes vocal folds to vibrate at very low frequency with low airflow rate [74, 77]. Creaky phonation is characterized by low and irregular F0, weak or damped pulses, and alternating longer and shorter pulses (period-doubled vibration).

### 2.2.3 Breathy phonation

Breathy phonation is produced due to incomplete closure of vocal folds, which causes constant leakage of air through the glottis. Vocal folds will vibrate, but will not be able to make good contact

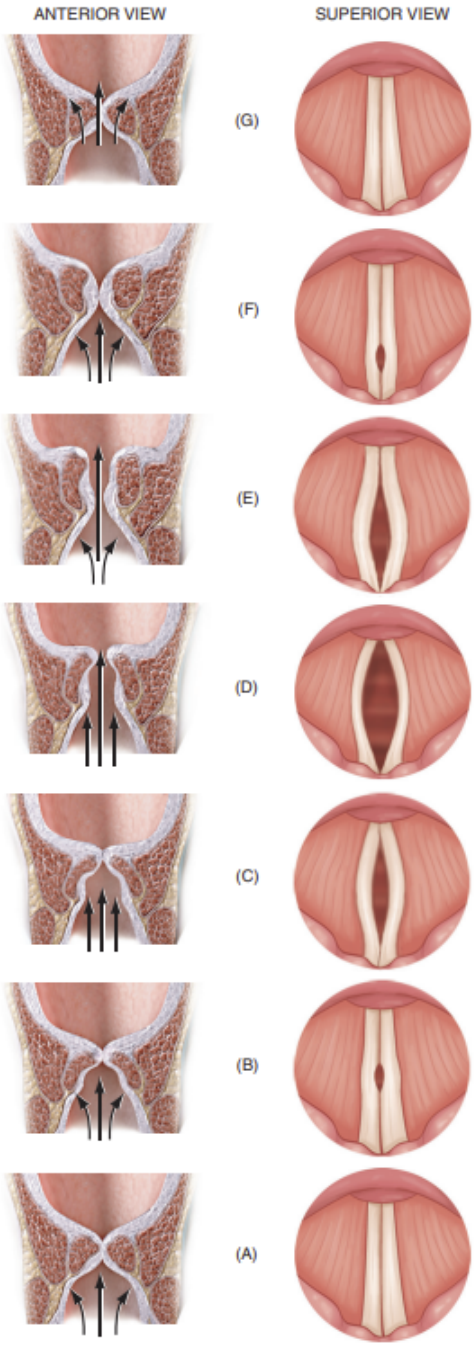



Figure 2.3: One complete cycle of vocal fold vibrations. A: Airflow moves toward the adducted vocal folds. B and C: Once subglottal pressure exceeds the tension between the vocal folds, maintaining them in adduction, the lower part of the vibrating vocal folds starts opening. D and E: The air pressure moves towards the upper part of the vocal folds. F and G: The increased velocity of airflow results in decreased air pressure due to Bernoulli's effect, causing vocal folds to come back to its original place. Source: From Seikel/Drumright/King. Anatomy & Physiology for Speech, Language, and Hearing, 5th Ed [71].

which results in turbulence noise or friction noise. Air will be leaked throughout the glottal cycle [78]. The breathy phonation is described by low muscle tension, medium longitudinal tension, and weak medial compression, which results in minimum adduction of vocal folds [65, 74]. It is characterized by increased spectral noise, especially at high frequency, which is due to constant leakage of air through the glottis.

### 2.2.4 Harsh phonation

Harsh phonation is also known as pressed or tense voice. It is described as a rasp or unpleasant sound associated with excessive approximation of the vocal folds. High medial compression and strong adductive tension along with increased tension in laryngeal and pharyngeal parts of vocal tract results in excessive approximation (over adduction) of vocal folds [78]. Acoustic characteristics associated with harsh phonation are low pitch and an increase in the overall intensity of sound.

### 2.2.5 Falsetto phonation

Falsetto phonation is described as having high longitudinal tension, along with strong adductive tension and medial compression. The vertical cross-section of the edges of the vocal folds is relatively thin due to the longitudinal stretch of vocal folds, resulting in a small vibrating mass. Hence, the frequency of vibration of vocal folds is very high, and the intensity of sound is low [74]. Compared to modal phonation, sub-glottal air pressure is small, due to which glottis remains slightly apart. Frictional noises sometimes accompany falsetto phonation.

### 2.2.6 Whisper phonation

Whisper phonation is produced due to low adductive tension, moderately high longitudinal tension with moderate compression of vocal folds [79]. This is unvoiced phonation in which vocal folds do not vibrate due to insufficient vocal fold adduction. Rigid vocal folds prevent the vibration.

## 2.3 Speech disorders

Speech disorders affect the the individual's ability to talk [80]. A subject suffering from speech disorders will not be able to articulate the words properly, which in turn affects their ability to communicate effectively [81]. Some speech disorders are due to physical abnormality, while others might be due to neurological problems. The most common categories of speech disorders are articulation, fluency, and voice disorders.

- An articulation disorder is a type of speech disorder where an individual has difficulty in articulating some speech sounds. Sounds may be distorted, omitted or substituted by another sound.

These difficulties can be caused by a variety of factors, such as learning difficulties, neurological issues (like dysarthria or apraxia), or structural abnormalities (like cleft lip and palate).

- Fluency disorders refers to a type of speech disorder characterized by disruptions in the smooth flow of speech. An individual suffering from fluency disorder may hesitate, repeat, or prolong sounds, words or phrases. Stuttering and cluttering are two common fluency disorders.

- Voice disorders occur due to anatomic or functional abnormality of the larynx which in turn affects the vocal fold vibrations. As a consequence, the sound produced by the larynx can vary in pitch, quality, or intensity.

Voice disorders are relevant to the interests of this thesis and will be discussed in more detail in the next section.

## 2.4 Voice disorders

Any abnormality in the larynx that affects voicing in speech production is referred to as a voice disorder. It may occur due to a poor respiratory system, incomplete glottal closure, growth of an extra lesion on the vocal fold, irregularity in the vibration of the vocal fold, or muscle weakness. These factors change regular quasi-periodic vibration into irregular and aperiodic vibration. The typical symptoms of a voice disorder include degradation of an individual's voice quality, reduction in loudness, loss of voice, and more effort in speaking or singing. From an auditory-perceptual point of view, the following acoustic dimensions are altered for subjects suffering from voice disorder:

**Fundamental frequency**: Fundamental frequency is a function of mass, elasticity, length of vocal folds, and sub-glottal pressure [1]. Voice disorders are due to extra growth on the vocal folds, insufficient tension, and improper coordination of laryngeal muscles, which results in irregular vocal fold vibrations and hence changes the fundamental frequency (F0) of vocal fold vibrations. If F0 increase is inappropriate to age and gender, it might cause the voice to sound shrilly, whereas a decrease in the F0 value might cause the voice to sound harsh or rough [15].

**Voice quality**: Voice disorders result in degradation of voice quality. The speech associated with voice disorders is identified as hoarse. For one of the categories of voice disorder (spasmodic dysphonia), the voice may sound strained as spasms cause the movement of the vocal folds to be a little difficult. For another category, vocal cord paralysis, the voice sounds breathy as the paralyzed vocal fold will not be able to move, which results in constant leakage of air during speech production.

**Loudness**: Loudness was found to have a high correlation with sound pressure level and sub-glottal pressure level; higher values of these parameters are associated with a loud voice. A subject suffering from voice disorder may not be able to produce a loud voice due to insufficient value of this pressure. Some of the voice disorders like vocal cord polyps, cysts, and nodules are characterized by a higher degree of loudness when compared to others like vocal fold bowing, presbyopia, and vocal fold atrophy [82].

### 2.4.1 Classification of voice disorders based on the etiology

Based on the etiology [53], voice disorders can be broadly classified into organic and non-organic voice disorders. Figure 2.4 shows the classification of voice disorders.

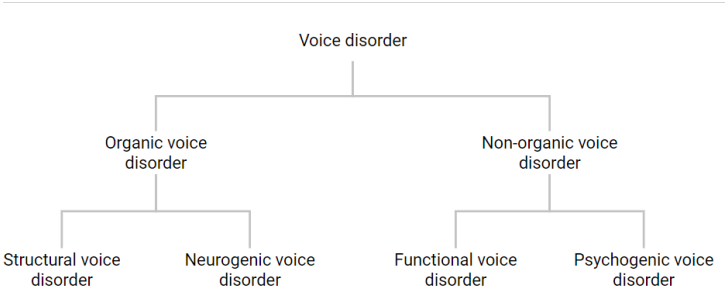



Figure 2.4: Classification of Voice disorders [53].

1. Organic Voice Disorders (OVD) are physiological voice disorders due to anatomic abnormalities in the larynx or muscle strain, which result in incomplete glottal closure. Patients suffering from OVD will not be able to produce normal phonation, which might be due to the presence of extra mass on the vocal folds or insufficient tension in the muscles controlling the larynx [83]. The onset of this pathology may be sudden or gradual. OVD, in a broad sense, can be categorized into two sub-types: structural and neurogenic.

   (a) Structural voice disorder (SD) is due to abnormal or extra growth on vocal folds, which cause irregular glottal open and close phases. Vocal cord polyps, nodules, leukoplakia, and laryngitis, are some of the structural voice disorders. Excessive use of voice, singing, yelling, and shouting may lead to swelling on the vocal cords. With time, swelling becomes hard, like callous at the middle part of the vocal folds, which results in a vocal cord nodule. Vocal cord polyps look like blisters or long growths or bumps on either vocal folds. Polyps are usually bigger than nodules as they have more blood vessels. Like nodules vocal card polyps are also due to loud singing, shouting, or smoking. Overuse of the larynx, infection or allergies in the larynx, or too much alcohol drinking results in inflamed vocal folds. The inflammation of vocal folds leads to laryngitis. Leukoplakia of the larynx is a voice disorder due to the presence of plaque-like cells on mucous membranes of the larynx which is mainly due to excess smoking. Leukoplakia is a Greek word which means white plaque. Figure 2.5 shows healthy vocal folds along with vocal folds suffering from structural voice disorders.

   (b) Neurogenic voice disorder (NVD) is caused by a problem in the central or peripheral nervous system that can weaken the muscle of the larynx . It affects the functioning of the phonation. Spasmodic dysphonia and recurrent laryngeal nerve palsy (RLNP) are the main common disorders that fall into the category of neurogenic voice disorders. Spasmodic dysphonia is

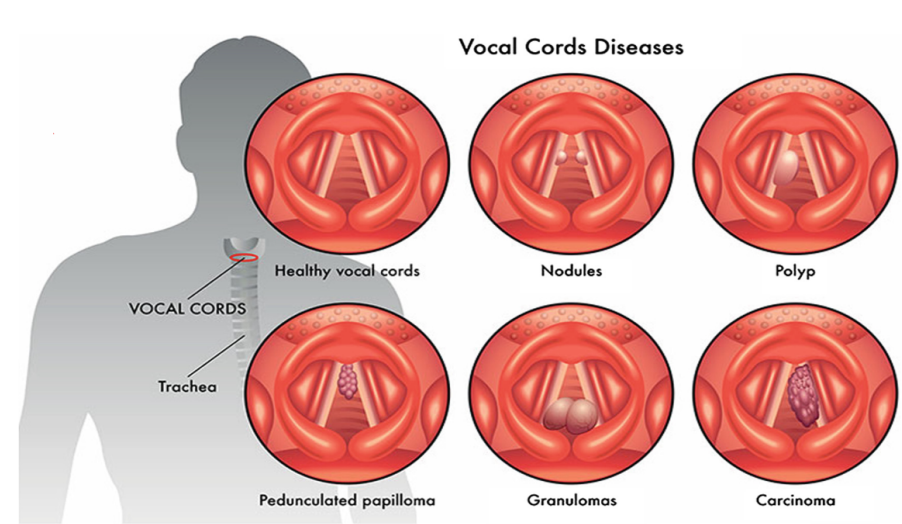



Figure 2.5: Top view of healthy vocal folds and vocal folds suffering with structural voice disorders [53].

also known as laryngeal dystonia [8]. Dystonia is a neurological disorder in which sudden, involuntary movements (spasms) occur in body parts. Dystonia can affect many parts of the body. If dystonia affects the voice box, it is called spasmodic dysphonia. One of the important characteristics of spasmodic dysphonia is voice break during speech [84]. The recurrent laryngeal nerves (RLNs) are responsible for the adduction and abduction of the vocal folds, as well as adjustment of the tension of the vocal fold. RLNP is a paralysis of the RLN on either one side or on both sides of the vocal folds. If there is no movement, it is known as paralysis, and if movement slows down, it is called paresis. The resulting effect of RLNP is that the vocal folds do not move close to each other, and the voice may sound breathy and rough. Figure 2.6 shows the top view of the vocal cord with and without paralysis during respiration and phonation.

2. Non-organic voice disorders are caused by ineffective use of the vocal mechanism or poor muscle control in subjects with normal physical structure. The phonation, in this case, is characterised by excessive laryngeal activity, excessive tension, and reduced vocal capacity [8]. It is broadly categorised into functional voice disorder and psychogenic voice disorder.

   (a) Functional voice disorders (FVD) are also known by another name as muscle tension voice disorders (MTVD). FVD is due to improper coordination of the laryngeal muscle and breathing pattern [85]. It is characterized by excessive force, tension or laryngeal muscle activity which is due to high vocal demand [86]. It is more common at the age of 40 to 50 years, and women have more chances of getting FVD than men [86].

   (b) Psychogenic voice disorder (PVD) occurs due to emotional stress or psychogenic trauma in the absence of organic pathology [86]. Subjects suffering from PVD will lose control

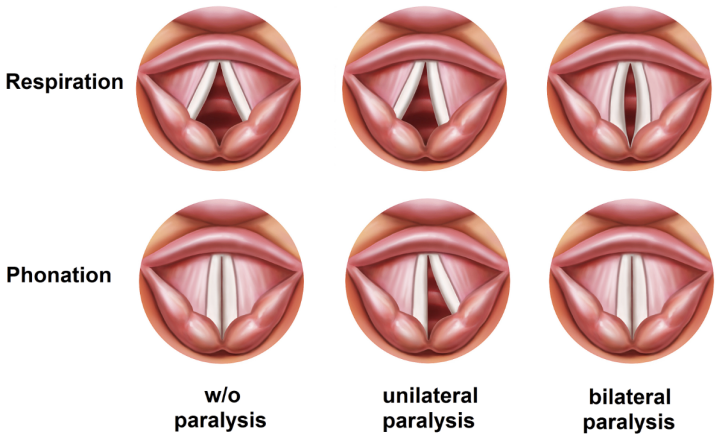



Figure 2.6: Top view of vocal folds during the respiration and phonation without paralysis, with unilateral and bilateral paralysis [70].

over the initiation and maintenance of phonation during speech production due to disturbed psychological processes like anxiety, depression, conversion reaction, or personality disorder. These are more common in women than men, with approximately in the ratio of 8:1. Psychogenic aphonia, puberphonia, and psychogenic spasmodic dysphonia are some of the disorders that fall into the PVD category [87].

## 2.5 Assessment of voice disorders

The assessment of voice disorders involves an examination of the patient to detect the presence of voice disorders, identify their underlying cause, and determine their severity. Assessment of voice disorders is crucial to avoid any further repercussions and to provide the subject with an opportunity to live a life of better quality. Voice disorders can be diagnosed by an SLP or an otolaryngologist (ear, nose, and throat doctor) through various methods. The methods used to assess voice disorders can be grouped into four categories: aerodynamic measurement, perceptual, visual imaging, and acoustic methods.

### 2.5.1 Aerodynamic measurement

The most widely used theory that describes the phonation or vocal fold vibration process is myoelastic-aerodynamic theory [11, 88]. In terms myoelastic-aerodynamic the word myo means muscles, which are used to denote that vocal folds are made up of muscles; elastic means vocal folds are associated with elasticity property, and aerodynamic refers to air flow and air pressure. According to this theory, the aerodynamic and muscular influences set the vocal folds into the vibration. Aerodynamic measurement helps the SLPs to evaluate the respiration function, laryngeal function, and coordination between them. To differentiate voice disorder from healthy voice, aerodynamics measurements like sub-glottal

air pressure, air flow rate, and laryngeal airway resistance were used, which involves measurement of air flows and air pressure [89]. Table 2.1 shows the aerodynamic measure and corresponding perceptual correlate used for the assessment of voice disorder.

Table 2.1: Perceptual correlates of aerodynamic measures [89].

| Measures | Perceptual correlate |
|---|---|
| Sub-glottal air pressure | Phonetory effort |
| Phonation threshold pressure | Effort to initiate phonation |
| Airflow | Breathiness |
| Laryngeal airway resistance | Phonatory effort, vocal strength, strain |
| Velopharyngeal measures | Nasal emission, strength of pressure consonants |

- Subglottal air pressure is defined as the pressure created below the glottis or pressure generated by lungs [8]. It is a very important parameter in phonation. To produce speech, when vocal folds are brought together, enough pressure must be built up below the glottis to initiate the phonation. Subglottal pressure will be different for different pathologies. For example, a person having polyp, laryngeal cancer or nodule will have large subglottal pressure as compared to a person having vocal fold ulcer. Phonation threshold pressure (PTP) is defined as the minimum value of subglottal air pressure required for vocal fold vibration. Organic voice disorders and adductor spasmodic dysphonia will have larger value of PTP when compared to the healthy speaker [90].

- Mean flow rate (MFR) is defined as the average volume of air passing through the glottis over a specified time. It is measured in mL per second. An increase in the value of MFR is observed in organic voice disorder due to incomplete glottal closure [91].

- Laryngeal airway resistance (LR) is a ratio of subglottal pressure to the glottal airflow. It indicates laryngeal constriction [9]. It is also used to differentiate voice quality. Depending on the type of phonation, the laryngeal resistance (LR) may be high or low. For breathy phonation, LR is small, but for pressed phonation, it is large compared to normal phonation. [92]. As voice disorders affect voice quality, LR was used as a reliable measure [93] for assessing voice disorder.

### 2.5.2 Perceptual methods

Perceptual methods are considered as "gold standard" for the assessment of voice disorder. SLPs use some perceptual scales to evaluate voice quality, while the patient uses others for rating their own voice quality [9]. This method depends on the auditory perceptual attribute of the speech and is used as the main part of routine clinical assessment for assessing the voice quality [94, 95]. These methods were widely used as an early diagnosis tool to judge the severity. It is generally influenced by personal and professional experience, cultural differences, relationships with patients, and the type of scale used for

assessing voice disorders. Due to subjectivity in nature, these scales have some limitations, but they are still the most widely used methods as they are designed based on perceptual phenomenon [96].

1. Clinical-based scales: These scales are used by SLPs to assess the voice by listening to the patients. These are used because voice has greater intuitive meaning than the instrumental methods [96]. The main voice quality associated with voice disorder is "roughness", "breathiness" and "strained voice". Hence, the scales used by SLPs are designed to assess these voice quality. The most frequently used and accepted scales for perceptual evaluation are GRBAS and CAPE-V.

    - GRBAS scale is the most common scale used by SLPs to rate the severity of voice disorder developed by Japan Society of Speech Therapy for perceptual measurement of voice [9]. It is a 4-point scale, in which G indicates a grade of hoarseness or overall severity, other 4 represents overall voice quality. R indicates roughness, B for breathiness, A for asthenia, and S indicates strain (as shown in Table 2.2). 0 indicates the absence of disorder, 1 indicates mild deficit, 2 indicates moderate deficit, and 3 indicates a severe deficit [97]. Roughness indicates irregularity in vocal fold vibrations and is present mainly in disorders such as vocal cord polyps, polypoid vocal cords, and laryngeal cancer. Breathiness is perceived as air leakage through the glottis and is present mainly in voice disorders such as recurrent laryngeal nerve paralysis, nodules, and laryngeal cancer [98]. Asthenic indicates the degree of weakness and can be heard mainly in psychosomatic aphonia. A strained voice indicates an effortful voice and is present mainly in spasmodic dysphonia and laryngeal cancer.

Table 2.2: Auditory-Perceptual Evaluation [9].

| Parameter | Description |
|---|---|
| G-Grade | Degree of hoarseness of the voice |
| R-Roughness | Impression of regularity of the vibration of the vocal folds |
| B-Breathiness | Degree to which air escaping from the vocal folds |
| A-Asthenia | Degree of weakness heard in the voice |
| S-Strain | Degree to which strain or hyperfunction use of phonation is heard |

    - Consensus Auditory-Perceptual Evaluation-Voice (CAPE-V) is an analog scale used by SLPs to rate the patient's voice quality [99]. It was developed by the American Speech-Language-Hearing Association's (ASHA's) Special Interest Division 3 for voice and voice disorders after 2002. SLPs use a CAPE-V form to rate a voice disorder patient using six parameters, including overall severity, roughness, breathiness, strain, pitch, and loudness. Additional parameters are used using a 100-mm visual analogue. scale [9].

2. **Patient's scale**: Depending on the profession and daily requirements of voice, individuals have different satisfaction levels with their voice quality. Hence, to evaluate the voice from an individual's perspective, a patient's scale was designed. Patient scales are very important in measuring

patients' general health and quality of life, knowing the onset of problems and their profession so that SLPs can plan their treatment accordingly. These scales provide novel information and are used as an initial step for diagnosing voice disorders. Based on these scales, SLPs can discuss the problem in more detail. Different scales were designed to assess the voice in different aspects. The most widely used scales are Voice Handicap Index (VHI) and Voice-Related Quality of Life (VRQOL).

- Voice Handicap Index (VHI), developed by Jacobson [100], is most widely used for providing details about subject's voice quality. It is used by SLPs to understand the social, function or environmental disturbances caused due to voice impairment. This self-questionnaire form to be filled by the patient or sometimes by care taker. It has three parts: functional, physical and emotional, each with 10 questions. The functional part has questions based on the disorder's effect on daily activity, the questions in the physical part are related to the subject's perception of voice quality, emotional part explores the patient's response to the the disorder [101]. Each question should be given a numeric value between 0 to 4 based on the frequency of occurrence. In this scale 0 indicates frequency as never, 1 rarely, 2 sometimes, 3 almost always and 4 indicates always. This index is applicable for all types of voice disorder [102].
- Voice-Related Quality of Life (VRQOL) is a 10-item self-administered instrument or scale designed to help SLPs. It measures the social-emotional and physical-functional aspects of voice. It comprises 10 questions to be filled by patient [103]. These questions are divided into two parts: physical and social-emotional domains. It is a 5-point rating questionnaire; a score of "1" indicates normal health, and a score of "5" indicates voice disorder is very severe.

### 2.5.3 Visual Imaging methods

The visual imaging method of diagnosing voice disorders utilizes special instrument to understand the functioning of vocal cord [8, 99]. These methods are used to analyse the structure of vocal folds the complete functioning of the larynx, and measure vocal fold vibrations. There are many methods available to examine the larynx visually; the most commonly used methods by SLPs are laryngoscopy, stroboscopy and their variations [9].

- Mirror laryngoscope: In this method voice box is examined by inserting a mirror into the mouth [104]. The image of the vocal fold can be seen by the tilted mirror. It is the oldest method for examining the larynx. The examiner will ask to protrude the tongue and then will place a mirror at the posterior oropharynx with gentle pressure at the soft palate. This process of examining the larynx may require anaesthesia. It is the most accurate method but painful for the patients.

- Direct laryngoscopy: This method can be used to examine the voice box or larynx by rigid or flexible endoscope. 70-degree or 90-degree scopes are used to examine voice box [9].

- Flexible endoscope: The most popular method of examining the larynx. The flexible endoscope examines the larynx during natural functions like singing and speech. An endoscope (thin, flexible tube) is inserted from one of the nostrils to the throat. The endoscope has an eyepiece and a fibre optic light inside the tube for examining the voice box and throat. It is important to analyse organic and neurological voice disorder [8, 99].

- Rigid endoscope: In this method, a rigid endoscope (usually 70 degrees or 90 degrees) is inserted into the patient's mouth while an Otolaryngologist or SLP holds the tongue. It gives a magnified and clear image of the voice box [9]. It is only suitable for the vowel 'ee'. Images are taken using the camera for analysis purposes. It is not suitable for muscle tension dysphonia and spasmodic dysphonia.

- Video Stroboscope: The rate of vibration of the vocal folds is very high in general, 100 to 400 vibrations per second; it is difficult to capture this vibration by the human eye. Hence, a special light known as strobe light is used. It flashes the light synchronising with the fundamental frequency of vocal fold vibrations. It provides an optical illusion of image [105]. Video stroboscope consists of a stroboscope with a flexible or rigid laryngoscope to analyse vocal fold vibrations. A stroboscope is an instrument that uses a pulse of light with frequency such that moving objects appear to be slow. Hence, by using the light at regular intervals, the shape, vibration, and movement of the vocal cords can be observed. As the shape of vocal folds can be easily observed, this method is useful for examining the stiffness of vocal folds and voice disorders related to structure abnormalities [99].

### 2.5.4 Objective assessment methods

Objective assessment methods rely on the features extracted from the speech signal for the automatic detection of voice disorders. Voice disorders are due to the asymmetrical distribution of mass, tension, uncoordinated movement of vocal folds, and insufficient sub-glottal pressure, which in turn changes the aerodynamic and acoustic characteristics of the voice. These methods can analyse the speech signal and also measure the different characteristics of the speech signal that were found to be perceptually correlated to voice disorders. Subjective methods (as discussed in the previous section) are influenced by personal and professional experience, cultural differences, and relationships with patients, and they are also laborious and time-consuming. On the other hand, objective assessment methods are repeatable, more effective in time, are economical [106]. Hence, these objective assessment methods are gaining popularity in the automatic detection of voice disorders from speech. This thesis focuses on the automatic detection of voice disorders from the speech signal using different acoustic features. In the literature, various acoustic features extracted from speech signals were explored to detect voice disorders

automatically. The following section discusses the different acoustic methods and features explored in the literature for assessing voice disorders.

#### 2.5.4.1 Studies based on the acoustic features

**Studies related to perturbation parameters** : In the source-filter theory of speech production, the source provides the energy for vocal fold vibration, which is then modified by the vocal tract system to produce speech [3]. The quasi-periodic vibration of the vocal folds results in phonation, which is due to the adduction and abduction of vocal folds [2]. Voice disorders are characterized by irregularities in vocal fold vibration, incomplete glottal closure and opening, and variation in the amplitude of consecutive opening and closing of the vocal folds. Hence, parameters that capture disturbances in vocal fold vibrations were used in the literature to distinguish voice disorders from healthy speech [107]. These parameters were divided into three categories: frequency perturbations, amplitude perturbations, and spectral noise parameters. The cycle-to-cycle perturbation of the glottal cycle is defined as jitter, which indicates the dysperiodicity of the glottal cycle [26, 30, 108, 109, 110] and disturbances in the amplitude of the successive laryngeal cycle is called the shimmer [111]. Jitter and shimmer were derived in the literature on steady vowel [111, 112], and on the running speech as well [108]. For a healthy speaker, these values are very small due to the acoustic stability of the excitation source signal. In contrast, a large value of these parameters indicates the presence of vocal fold pathology. Studies have revealed that jitter models aperiodicity in voice, and the voice quality associated with jitter is roughness [26]. However, the estimation of jitter requires the exact calculation of the fundamental frequency contour or pitch period contour. Spectral based method for the calculation of jitter was also explored [108]. Moreover, variations like absolute jitter(in ms), percentile jitter, pitch perturbation quotient, and jitter based on the autoregressive model were used in the literature [107]. The shimmer indicates the presence of noise and breathiness due to lesions on the vocal folds [26]. Similar variations like shimmer in dB, percent shimmer, and amplitude perturbation quotients were also used for the calculation of shimmer [30, 31, 32]. In [113] effect of parameters like gender, vowel, SPL, and F0 was studied using ANOVA analysis on jitter and shimmer. The results of this study concluded that voice intensity has a larger effect on calculation jitter and shimmer. It was also concluded that the importance of vowel /a/in pathological study and setting the threshold based on gender would help clinicians in the detection of pathology. In [25] perturbation parameters (jitter, shimmer, HNR etc) were derived from the zero frequency filtered (ZFF) signal for discrimination of pathological voice from healthy voice. Epoch locations were derived from positive to negative crossings of ZFF signal; from these locations, pitch contour was obtained, which in turn was used to derive jitter and its variations. The strength of excitation (SoE), indicates the strength of the glottal signal. It is used to compute the amplitude perturbation parameters. The results of this study showed that perturbation parameters measured from ZFF signals are better than PRAAT-based perturbation features for both clean and noisy conditions. The increased value of perturbation parameters is observed for pathological voice compared to healthy voice.

**Studies based on excitation source information**: An important feature in the identification of pathological voice is the degree of vocal fold adduction towards the glottis. Hence, glottal signal parameters like open quotient (OQ), speed quotient (SQ), and close quotient (CQ), along with the difference between first and second harmonics (H1-H2) and harmonic richness factor (HRF) were used for detection of the voice quality. In [42], different glottal signal parameters were explored to detect the vocal fold pathologies, namely nodules and unilateral paralysis; this study found that glottal signal parameters discriminate pathologies better than MFCC feature. Glottal source time and frequency domain features derived from the quasi-closed phase (QCP) method in [40] were used to detect voice disorder. Along with this, glottal source feature derived from speech signals and features were also derived from ZFF method (like SoE, energy of excitation (EoE), loudness, and ZFF signal energy) were used to extract the excitation source information of speech signals. In [114] explored power spectral density (PSD) derived from the glottal source waveform was used as a bio-metrical signature for pathological voice. In the study [115] pitch strength were investigated as a good measure for classifying the dysphonic voices before and after surgical/behavioral treatment. The author in the study [116] explored the feature derived from the interlaced derivative pattern of glottal source waveform as a promising indicator for pathology detection. In [117] residual signal obtained from inverse filtering analysis is used as an appropriate measure for identification of the laryngeal pathology. This study is based on the knowledge that residual signal is obtained by removing the supra-glottal signal from the speech, which might capture better information about laryngeal pathology than the original speech signal. The study found that a healthy or homophonic signal has sharp peaks at the start of each pitch period with a relatively low noise level between the periods in the residual signal, whereas this is not the same for the pathological voice. Pathological voices have aperiodicity and noisy-like characteristics in the residual signal [117].

**Studies on noise measures**: The incomplete vocal fold closure causes turbulent airflow from the vibratory vocal folds, this constant leakage of air results in noisy components in the speech signal. As regularity in vibration is not present in most of the pathologies, the features that can capture the information about the source of excitation (whether it is noisy or voiced excitation) will be better in understanding pathologies. Hence the parameters like signal to noise ratio (SNR) [38, 118, 119], normalized noise energy (NNE) [39, 120], harmonic to noise ratio (HNR) [14, 26, 39], noise to harmonic ratio (NHR) [121], glottal to noise excitation (GNE) [39, 122, 123] were used in the literature to indicate the noise parameter for discrimination of voice pathology. In [118] SNR was measured in both time and frequency domains and was calculated for different laryngeal pathologies like functional voice disorders, RLNP, laryngitis, and papillomatosis. In this study, SNR was found to be correlated with hoarseness. SNR [38] in this study was derived on the running speech signal as the ratio of the energy of correlated signal to uncorrelated signal. For the calculation of uncorrelated signal, first, long-term and short-term correlated components were calculated from inverse filtering and residue signal was considered as noise components. The GNE indicates that speech signal originates from the quasi-periodic vibration of vocal fold or by turbulent noise and is used in many studies as an indicator of breathiness [122]. The motivation for using GNE as an indicator of pathology is that vocal folds, when excited by a voiced

signal, are found to have a highly correlated Hilbert envelope (HE) across all frequency bands, while when excited by noise, HEs are uncorrelated [122].NNE is the ratio of noise energy to the total energy of the signal, which is inversely related to the cepstral-based Harmonic-to-Noise Ratio (HNR), and it indicates the amount of turbulent noise due to incomplete glottal closure during phonation."[39, 122]. HNR was calculated in the time domain [34], frequency domain [34] and cepstral domain as well [33]. In [120], NNE was used to detect voice pathologies like glottal cancer, recurrent nerve palsy and vocal cord nodules. According to the study NNE performed better than two noise parameters like relative harmonic intensity and HNR. For pathological voice, a large value of these noise parameters was found to be due to noisy excitation source characteristics associated with the voice.

**Studies based on spectral and cepstral features**: Some studies in the literature investigated the features obtained from the different frequency bands for dysphonic voice detection. Spectral energy and band power correlation time [124], normalized autocorrelation function [125], of the filter banks were used as good indicator to identify the voice pathology. Importance of frequency bands with features like peak and lag of autocorrelation function along with entropy feature were also explored for voice pathology detection [7]. According to study [126], it was also found that spectrum coefficients obtained from lower frequency ranges between 0 Hz to 3000 Hz are more significant than other frequency ranges for diagnosing the voice disorder. Harshness, roughness, breathiness, and strained voice are the main symptoms associated with voice disorder [10]. These voice quality, loudness, and intonation from the speech signal are perceived in the long term [66]. Hence, the features that are present in the long term will be better captured by the Long Term Average Spectrum (LTAS) instead of the short time variation present in the speech. Many researchers used LTAS in clinical applications to detect the presence of different voice pathology before and after surgery and in the quantification of voice quality. Some studies claim that LTAS can be used for voice classification [127]. In [67] LTAS was used as a good acoustic measure to differentiate the male and female. In [128], features derived from long-term spectra were used to study voice quality changes before and after surgery. Other works in this direction were finding differences related to age [68], professional singers, different styles of singing [129], speaking and singing [130] and quantifying the quality of voice [131].

Even though most voice disorders affect the functioning and structure of the larynx, vocal tract features were also explored in the literature. Features like mel frequency cepstral coefficients (MFCC) [49], linear prediction cepstral coefficients (LPCC) [49], perceptually linear prediction coefficient (PLP) [51], which were used to capture the vocal tract characteristics were also used in voice pathology detection. In the [132], the vocal tract area was explored for detection of voice pathology based on the assumption that for healthy subjects vocal tract area does not change significantly across the frames while this area shows irregularity for pathological voice due to irregular vocal fold vibration. In [133] mean and standard deviation of the first three formant frequencies and its dynamic features were used to discriminate vocal fold pathologies. Cepstral peak prominence (CPP) [43, 44, 45, 47, 48, 134] was also used as reliable measure for differentiating the dysphonic voice from healthy voice. CPP measure is based on the concept that periodic signals have a higher amplitude at the fundamental frequency and its har-

monic frequency. Hence, periodic signals have a prominent peak in their cepstral, which is present at the fundamental period. As it was found that the cepstral peak depends on window size, overall energy and periodicity were used instead of the amplitude prominence of the peak. CPP was measured as a difference between the linear regression line and the cepstral peak. Another variant of CPP was also explored, known as smoothed CPP (CPPS) [65], which was measured by first averaging the cepstra over all the frames and then calculating peak prominence. Both of these parameters were found to be better measures for pathological voice detection (small value of parameters), with CPPS being better.

#### 2.5.4.2 Studies on voice quality analysis

Voice quality is a perceptual attribute defined by phonation type. Based on the different tension present on laryngeal muscles and respiratory effort, human is capable of producing various types of phonation. The literature used different features derived from the epoch and GVV waveform to analyse the phonation types. Modal phonation is considered as a reference phonation for analysing the different type of phonation [3]. In the study [135], phonations like breathy, modal, and pressed phonation were analysed using the features derived from ZFF method, zero time windowing (ZTW) method, and single frequency filtering (SFF) method for normal and singing speech. To discriminate different phonation types, time domain and frequency domain parameters derived from GVV waveform were used [76, 136]. Breathy phonation has more influence of the sub-glottal system. In contrast, the pressed phonation has less influence of the sub-glottal system than modal phonation; hence study in this [78] used low-frequency spectral density (LFSD) for classifying the different phonation. In the study [137], different acoustic parameters like jitter, shimmer, SNR, and peaks derived from LP residual signal [138] were used to identify creaky phonation. In the study [139] performance of different state-of-the-art epoch extraction algorithms was compared for different modal and non-modal phonations. It was found that non-modal phonation in which there is variation in the glottal source characteristics, is more challenging than modal phonation. Voice disorders affect the structure and functioning of the larynx, subjects with voice disorders require more vocal effort. The loudness of speech signal is associated with the vocal effort [140]. Hence study of loudness will help to understand voice disorders better. In the study [141] strength of excitation (SOE) derived from the Hilbert envelope of LP residual signal is used as a parameter to relate of loudness. The result of the paper concludes that impulses like excitation are more sharper for loud sounds than soft and normal sounds as greater SoE is present when an amount of energy is present for a short duration than the same energy is present for a longer duration of time. Another work related to this study explored the feature from excitation source signal like discrete cosine transform of integrated linear prediction residual (DCT-ILPR), mel-power difference of spectrum in sub-bands (MPDSS), and residual mel-frequency cepstral coefficient (RMFCC), for classification of shouted and normal speech [142]. The author used DCT-ILPR, MPDSS and RMFCC to capture information of glottal shape, periodicity and spectral information respectively, from the excitation source signal, which provide more relevant information to discriminate shout speech from normal speech. The maximum airflow declination rate (MFDR) was found to be highly correlated to sound pressure level (SPL), which

is found to be lower for soft voice than normal and loud voice [143]. The study in this [144] observed different parameters like fundamental frequency, the difference between the first and second harmonics (H1-H2), normalized amplitude quotient (NAQ) and SPL. This study found that smaller spectral tilt, high F0 and vocal energy, and increased duration are some of the characteristics associated with shouted speech.

One very important perceptual attribute for analysing voice disorders is breathiness. The incomplete closure of vocal folds during the closed phase of the phonation cycle and sub-glottal coupling cause constant air leakage through the glottis, giving rise to turbulence, which results in breathy voice [145]. The sub-glottal coupling increases the width and decreases the amplitude of the first formant frequency. The incomplete glottal closure results in a symmetrical open and closed phase, which is responsible for the relatively increased amplitude of the first harmonic in the spectrum. Moreover, it is also responsible for the decrease in the amplitude at high frequencies. Noise parameters were also used to correlate with breathiness, as constant air leakage is associated throughout the breathy sound, which in turn results in noise. The spectrum associated with breathy voice was found to have high spectral noise at high and medium frequency [145]. All perceptual characteristics of breathiness are found to be associated with noise, aperiodicity, spectral tilt and perturbation. Hence acoustic features like the difference between the amplitude of first and second harmonic (H1-H2) [65, 146], GNE [12], HNR [33], NNR [45], CHNR [45] and amplitude of first harmonic [65, 145], NNE [12, 120] were used as an acoustic correlate for breathiness. CPP was found as a correlate of breathiness and roughness (perceptual attribute of voice pathology) in some studies [45, 65]. For normal speech, which has a comparatively good harmonic structure, which indicates the large value of CPP, whereas the breathy voice has a relatively flat spectrum, CPP is found to have a small value. The spectral differences like H1-H2, H1-A1, H1-A2, and H1-A3 were used to indicate the presence of spectral noise and breathy voice associated with hyperfunction voice disorder and vocal nodules [147, 148]. A1, A2, and A3 were used to define as amplitude of the most robust harmonic in the region of first, second and third formant frequency, respectively. These spectral tilt were used as they indicate the degree of vocal fold closure, as incomplete glottal closure is a strong characteristic associated with voice disorders. Insufficient vocal fold adduction might lower the amplitudes of higher frequency harmonics, resulting in higher spectral noise.

## 2.6 Significant gaps

- In most of the literature, voice disorders detection was considered as two-class problem, where voice disorders were discriminated from healthy samples using the acoustic features. In literature, assessment of voice disorders was not explored. There is a need for detailed identification of voice disorders. Hence, there is a need for a detailed analysis of voice disorders from the clinical point.
- The epoch locations estimated from the speech signals were used to obtain the perturbation parameters like jitter, shimmer, and fundamental frequency contour, which are very important for detecting voice pathology. Methods used for the calculation of epoch location work efficiently

for clean speech. Some voice disorders affect the structure of vocal folds, whereas disorders like functional voice disorders are due to excessive or inappropriate muscle force. Should it be studied whether state-of-the-art methods perform accurately for different voice disorders in a similar way or not?

- The perceptual methods were considered as the golden standard in identifying voice pathology. By incorporating the knowledge of the human auditory system performance of voice disorder detection and assessment system may work better. Many auditory filter banks were explored for speech analysis in the literature to improve speech systems' performance. In the literature, critical band filter bank-based features were explored, but different perceptually motivated filter banks were not explored. The open problem here we found is that considering the perceptually motivated filter for different feature extraction might help to understand the voice disorders in a better way. Moreover, importance of different frequency bands for different types of disorders can be explored.

- Loudness is one of the important perceptual characteristics used by SLPs for voice disorder detection as a subject with a voice disorder will not be able to produce loud sounds compared to a healthy subject. The open issue here is whether the loudness and voice quality affect the different organic and non-organic voice disorders in the same way.

- Voice disorders affect one of the dimensions of speech which is voice quality. From the literature, it was found that the phase spectrum of speech signal captures the information about voice quality. Incorporating the information about the phase spectrum along with the magnitude spectrum which provides complete information about the speech, might improve automatic detection and assessment of voice pathology. In the literature, phase spectrum features derived from group delay function were used, while the analytic phase was not explored for voice pathology.

## 2.7 Voice disorder databases

Automatic detection of voice disorders relies on the availability of databases that contain recordings of both healthy individuals and those with voice disorders. There are many public and private databases that have been collected for the purposes of automatic detection, identification, or assessment of voice pathologies. Massachusetts Eye and Ear Infirmary Database (MEEI) [149], Saarbruecken Voice Database (SVD) [150], Hospital Universitario Prıncipe de Asturias (HUPA) [151], Arabic Voice Pathology Database (AVPD) [152], Hospital Gregorio Maranon (GMar) [153] are most commonly used publicly available databases. SVD corpus contains more than 2000 voice recordings out of which 687 are collected from healthy subjects (428 females and 259 males) and 1356 are collected from subjects (629 males and 727 females) with voice disorders. HUPA database contains recordings of the vowel /a/ for a total of 440 subjects. Out of total of 366 recordings, 201 recordings are from pathological subjects, and 239 recordings are from normal subjects. AVPD database contains a total of 366 samples of normal

and pathological subjects. 188 samples are from healthy subjects, and 178 are from pathological subjects. Recordings are available for vowels (a,i,u), isolated word (like Arabic numbers and words) and running speech. All samples are recorded at the sampling frequency of 48 KHz with 16 bits of resolution. The GMAR database contains recordings of Spanish speakers of the vowels/a/, /i/, and /u/. All the samples are recorded at a sampling frequency of 22050 Hz. For the vowel /a/ 202 (107 disorder samples and 95 healthy samples), for vowel /i/ 190 (96 disorder samples and 94 healthy samples) , and for vowel /u/ 176 (90 voice disorder samples and 86 healthy samples), samples are available. MEEI database is commercially available and the most widely used database in the field of voice disorder detection. It contains a recording of vowel /a/ and rainbow passages for 684 subjects. Out of 684 subjects, 53 samples belong to healthy subjects, whereas 631 samples belong to subjects suffering from voice disorders. Table 2.3 shows the list of voice disorder databases along with a number of samples and speech stimuli available in the literature.

Table 2.3: Details of voice disorder database available in literature, its corresponding number of samples and speech stimuli.

| Database | Number of samples | | Speech stimuli |
|---|---|---|---|
| | Healthy samples | Disorder samples | |
| Saarbruecken Voice disorder database (SVD) | 687 | 1356 | Vowel (a,i,u) and Sentence |
| Hospital Universitario Prıncipe de Asturias database (HUPA) | 239 | 201 | Vowel (a) |
| Arabic Voice Pathology database (AVPD) | 188 | 178 | Vowel, words, Sentences |
| Hospital Gregorio Maranon database (GMar) | 95 | 107 | Vowel (a) |
| | 94 | 96 | Vowel (i) |
| | 86 | 90 | Vowel (u) |
| Massachusetts Eye and Ear Infirmary database (MEEI) | 53 | 631 | Vowel (a) and sentences |

### 2.7.1 Database used in this thesis

Databases used in this thesis are Saarbruecken voice disorder (SVD) dataset [150], and Hospital Universitario Prıncipe de Asturias (HUPA) database [151].

1. SVD database is the most widely explored database due its availability on [1]. It contains more than 2000 (from 71 different voice disorder categories) voice recordings sampled at 50 kHz. The recording session consists of a German sentence and vowels of /a/, /i/, and /u/ in normal, high, low and rising-falling pitch. 625 samples were considered from the healthy class, and total of 950 voice samples were considered from different voice disorders categories for vowel /a/, /i/, and /u/ in normal, high, low and rising-falling pitch. In our study, all recordings were down-sampled to 8000 Hz.

[1] http://www.stimmdatenbank.coli.uni-saarland.de/

2. HUPA database is considered to perform detection tasks. It contains recordings of the vowel /a/ for a total of 440 subjects. Auditory-perceptual ratings according to GRBAS [99] scale is available for HUPA database. It contains the five different components: Grade of hoarseness (G), Roughness (R), Breathiness (B), Asthenia (A), and Strain (S). Each component is rated as 0, 1, 2, or 3, where 0 indicates normal, 1 mild, 2 moderate, and 3 indicates a more severe voice disorder. Table 2.4 shows the database details and the the number of samples used in this thesis for performing the voice disorder detection task. A total of 659 and 950 samples are considered from the SVD dataset for healthy and voice disorder classes, respectively. 239 samples and 201 samples are considered from healthy and voice disorder classes from the HUPA database to perform experiments.

Table 2.4: Details of the number of samples used for the detection task in our study from SVD and HUPA database.

| SVD database | | HUPA database | |
|---|---|---|---|
| Healthy | Voice Disorder | Healthy | Voice Disorder |
| 625 | 950 | 239 | 201 |

## 2.8 Summary and conclusions

This chapter overviews the speech production process, phonation, and voice disorders. It also explores various methods used for assessing voice disorders. Additionally, it conducts a literature survey of different acoustic methods utilized for automatically detecting voice disorders. This analysis identifies gaps in the existing literature, and some of these issues are addressed in the present thesis. This chapter also discussed the standard database commonly used in previous studies for the automatic detection of voice disorders. Furthermore, it details the specific databases used in this study for the automatic detection of voice disorders.

*Chapter 3*

# Exploring the excitation source based information for detection and identification of voice disorders

Voice disorders may alter the phonation (vocal fold vibration) characteristics of speech by affecting muscle tension and sub-glottal pressure. The fundamental frequency, voice quality, and loudness of speech are the main features that can be impacted by voice disorders, as reported in studies [10, 15, 65]. These dimensions of speech were found to be effectively captured from the excitation source information. From the literature, it can be concluded that most of the studies used these features to discriminate healthy speech from voice-disordered speech. However, the clinical way of assessing the voice disorder requires a more detailed analysis of the voice disorder. Hence, this chapter explored the excitation source-based features for the detection and identification of voice disorders in a clinical way. A more detailed analysis of voice disorders was performed to know whether the disorder is structural, neurogenic, functional or psychogenic. The excitation source features used in this chapter are intonation features, glottal features, and cepstral coefficients derived from the excitation source signal. These excitation source features were compared with state-of-the-art MFCC, LPCC, and openSMILE features.

The rest of the chapter is organised as follows. Section 3.1 describes the clinical perspective of identification of voice disorder, section 3.2 discusses the different evidence of excitation source. Section 3.3 presents the experimental setup with details of the database, extraction of excitation source evidence, feature extraction and classifier. Results and discussion of the voice disorder detection and identification system are presented in Section 3.4. Finally, the summary and conclusion of this work are discussed in Section 3.5.

## 3.1 Clinical way of identification of voice disorder

The aim of this thesis is to investigate an objective method for detecting and identifying voice disorders from a clinical perspective. Such a method can be used by SLPs as pre-diagnostic tool for the assessment of voice disorder. The detection task is the discrimination of healthy subjects from the voice disorder subjects (as shown in Figure 3.1), whereas identification requires a more in-depth analysis to

determine the underlying cause of the voice disorder. ASHA classifies voice disorders, based on their etiology, into organic voice disorders (OVD) and non-organic voice disorders (NOVD) (as discussed in the previous chapter). Organic voice disorders can be further categorized as either structural or neurogenic, while NOVD can be classified as functional or psychogenic. Therefore, identification refers to the process of determining the specific category or type of voice disorder. In order to perform the clinical way of identification, SVD database from voice disorder subjects was grouped into four classes. Out of the 71 different disorder categories present in the SVD database, our work focused only on the categories that had more than 30 subject recordings. Further details about the database are discussed in the experiment set-up section.

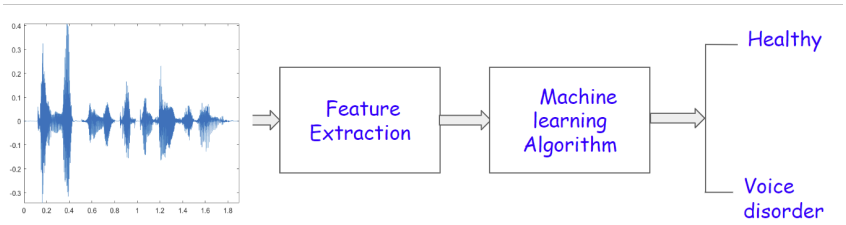



Figure 3.1: Voice disorder detection task.

This thesis explored a multi-level classification approach employing four binary classifiers to assess voice disorders (as shown in Figure 3.2). The first classifier distinguished healthy samples from voice disorder samples, referred to as voice disorder detection. The other classifiers are trained to identify the cause of the voice disorder. The second classifier is trained to distinguish organic voice disorder from the non-organic voice disorder category. The organic voice disorder class was further subdivided into structural or neurogenic classes, while the non-organic class was classified as functional or psychogenic. This detailed analysis of voice disorders can assist SLPs in planning appropriate surgical interventions or speech therapy. The binary classifiers were trained using a machine learning classifier algorithm. Throughout the thesis, the same approach is followed to perform the experiments.

## 3.2 Excitation source evidences

According to the source-filter theory of speech production, the source provides the energy for vocal fold vibration, which is then modified by the vocal tract system to produce the speech [3]. In order to capture the excitation source signal, the source signal should be separated from the vocal tract signal. This can be achieved through various methods, such as using specialized devices or employing speech signal processing methods. These methods help to differentiate the influence of the excitation source at the glottis from the resonances produced by the vocal tract system.

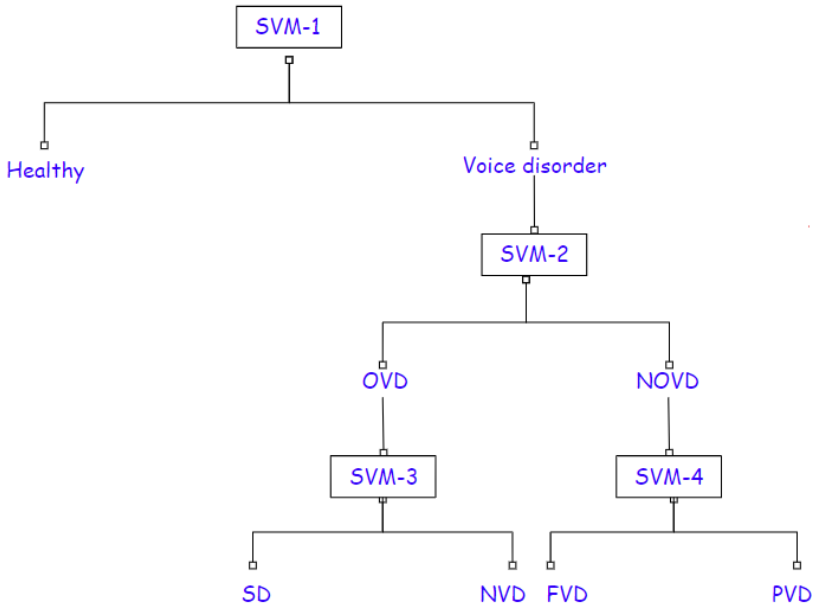



Figure 3.2: Voice disorder identification task.

### 3.2.1 EGG signal

Electroglottograph (EGG) signal represents the vocal fold vibration during the production of the voiced speech sound. It is the output of the EGG system. EGG system is a noninvasive measurement of the excitation source. It consists of two electrodes, whose one end is given the input from a high-frequency generator, and the other end is connected to the neck for measuring the impedance [154, 155]. By analyzing the impedance values, the EGG signal provides information about the opening and closing phases of the vocal folds. The high impedance value in the EGG signal indicates the opening phase of the vocal fold, while the low impedance value signifies the closing phase of vocal folds. In this way, the resultant signal from the EGG system represents the glottal flow waveform of voiced sound. Importantly, the EGG signal is not affected by vocal tract resonances. Its limitation is that it is only available with a few databases and is primarily utilized for clinical purposes. Figure 3.3 shows a schematic of the EGG during vocal fold opening and closing phase [154]. Figure 3.4 depicts the EGG signal and its first order difference signal refers to as differenced EGG (DEGG) signal.

### 3.2.2 LP residual

LP residual is one of the most widely used signals that models the excitation source information from the speech signal. LP residual signal is derived from LP analysis. LP analysis is based on the source-filter model of speech production [4, 156]. According to this, the speech signal is produced when the excitation source signal is passed through the vocal tract system. The excitation source signal

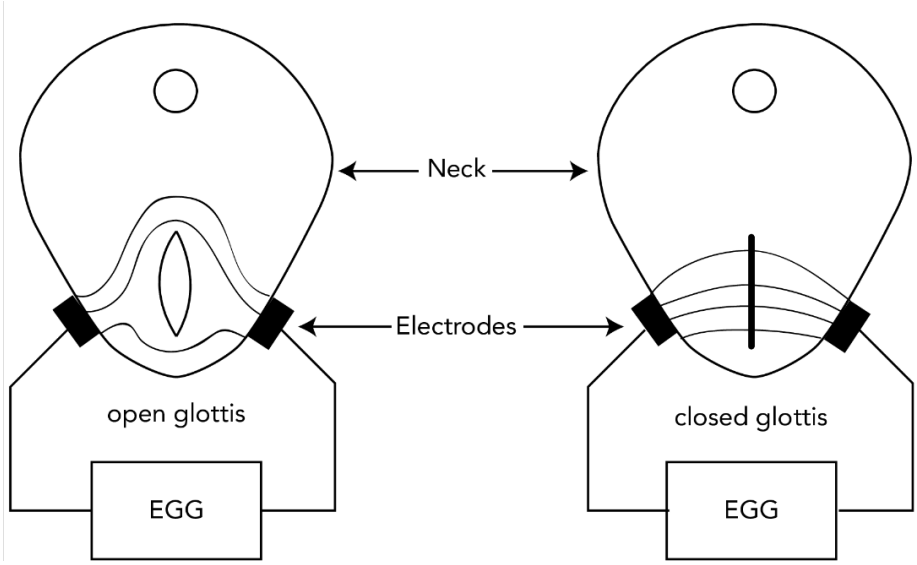



Figure 3.3: Principle of the electroglottograph device. A transverse section of the neck is shown with an open glottis (on the left) and a closed glottis (on the right). The electric field passing through the neck is represented by lines. When the vocal folds are apart, the opening distorts the electric field and impedance increases. When the vocal folds come closer, current passes through the electrodes, reducing impedance [154].

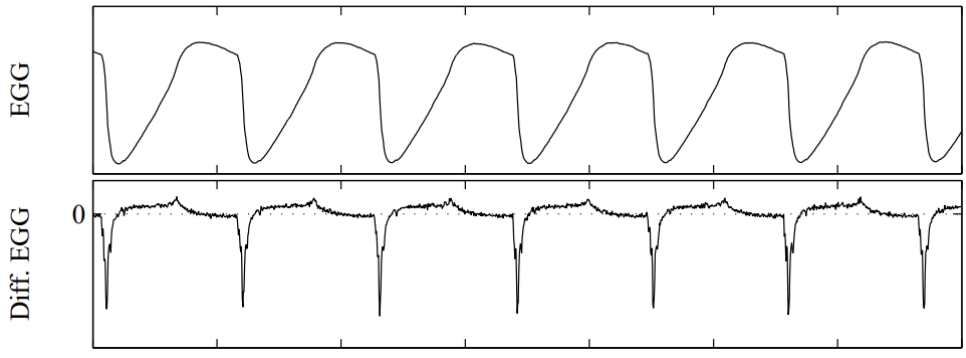



Figure 3.4: Illustration of EGG signal and its corresponding dEGG signal.

is modelled as a train of impulse signals for voiced sound and random noise for unvoiced sound. The vocal tract system is modelled as an all-pole filter system. Therefore, the excitation source signal is extracted from the speech signal by passing it through the inverse filter (the inverse of an all-pole filter). Figure 3.5 represents a block diagram for the LP model of speech production.

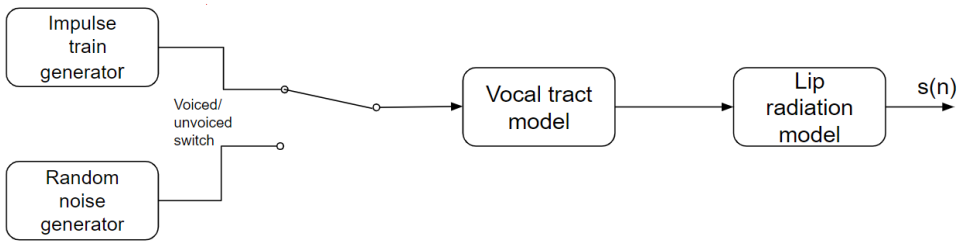



Figure 3.5: Linear prediction model of speech production [54].

### 3.2.3 Glottal inverse filtering

GVV signal is the evidence of excitation source derived from glottal inverse filtering (GIF) method [63, 157]. Figure 3.6 shows the block diagram of GIF method. In the inverse filtering method, to derive the excitation source representation, vocal tract resonances are cancelled by passing speech signal through an anti-resonance (zero) filter. The GIF method is based on the linear source-filter model of the speech production method. According to this model vocal tract system is modelled as an all-pole filter. First, the filter's response is obtained then the speech signal is passed through the inverse vocal tract filter to obtain the excitation source response. Then this signal is passed through the integrator to cancel the lip radiation effect, and the resultant signal is termed as GVV signal or glottal flow signal [158].

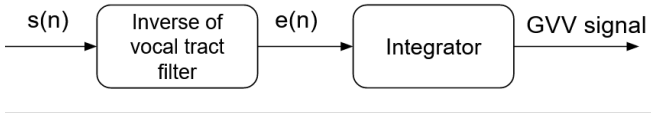



Figure 3.6: Glottal inverse filtering [63].

### 3.2.4 ZFF signal

ZFF signal is evidence of the excitation source signal obtained from the ZFF method [56]. This method is based on the assumption that for the voiced sound, the vocal tract system is excited by a sequence of impulse trains of varying strength. The effect of impulse-like excitation is present at all frequencies, including at zero frequency. In contrast, the effect due to the resonance of the vocal tract

filter is present at a much higher frequency than zero frequency. Hence, to extract the excitation source information, the speech signal is passed twice through a zero-frequency resonator (as shown in Figure 3.7). This process attenuates the higher-order harmonics corresponding to the vocal tract system and emphasises the excitation source characteristics. The output of the ZFF filter shows the polynomial growth/decay, which is due to the fact that time domain equivalent of the ZFF filter is an integrator. In order to compensate for the trend introduced in the signal, the filtered signal is passed through a moving average filter with a window size of 5 to 10 ms.

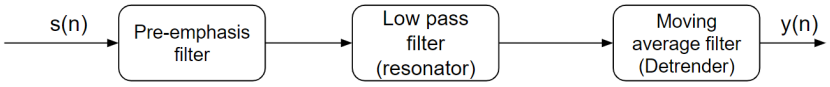



Figure 3.7: Block diagram of ZFF method [56].

## 3.3 Experimental setup

This section discusses the excitation source features explored in this chapter for automatic detection and identification of voice disorders. It also discussed the details about baseline features, database and classifier used for performing the experiments.

### 3.3.1 Features derived from the excitation source evidences

This subsection discusses the features derived from the excitation source evidences like GVV signal, ZFF signal, and LP residual signal, which are explored in this chapter for automatic detection and identification of voice disorder in a clinical way.

#### 3.3.1.1 Glottal features

The glottal flow waveform which is estimated from the inverse filtering method is used to compute glottal parameters as in [62]. The method we used to derive the GVV waveform is quasi-closed-phase (QCP) analysis method. QCP analysis is a state-of-the-art technique to estimate the glottal flow waveform [63]. Figure 3.8 depicts the block diagram of QCP method. It is based on closed-phase analysis in which the vocal tract model was estimated from speech samples in the closed phase of the glottal cycle [158] due to the decoupling of the oral cavity, lung, and trachea during this phase. QCP estimates vocal tract resonance from speech samples by using a weighted linear prediction (WLP) analysis. The

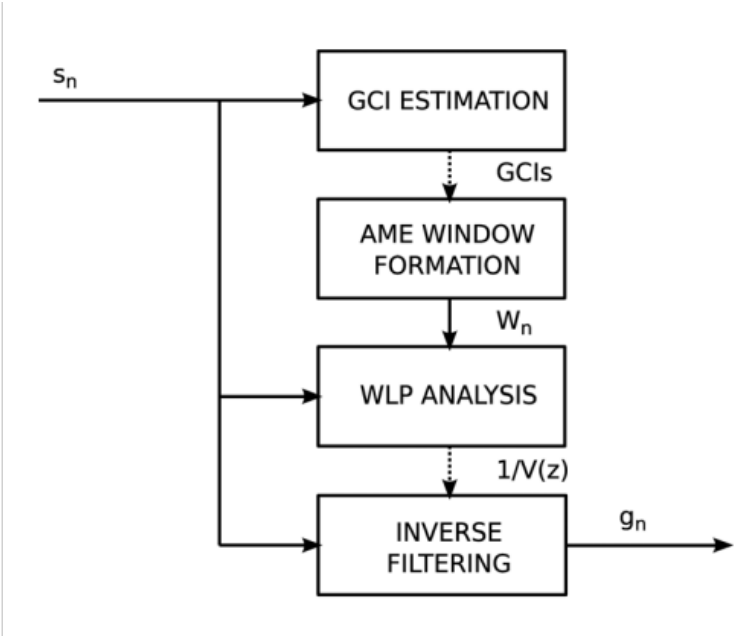



Figure 3.8: QCP method [63].

attenuated main excitation (AME) waveform was used as a weighting function, attenuating the samples of open phase region compared to the close phase samples of glottal cycles, which results in a better estimate of the vocal tract model. Finally, the glottal flow waveform was estimated by inverse filtering the speech signal with the vocal tract model. The glottal parameters include time-domain features and frequency-domain features.

1. **Time-domain features derived from glottal flow waveform**: Two sets of features are obtained directly from the time-domain representation of the glottal flow, namely time-domain and amplitude-based features. Time domain glottal flow waveform is characterised by three phases, namely closed phase ($T_c$), opening phase ($T_o$), and closing phase ($T_{cl}$) as can be seen in the Figure 3.9. During the closed phase of the glottal cycle, the vocal folds are fully in contact along their entire length, leading to the obstruction of airflow through them. The opening phase refers to the time duration in which vocal folds begin to separate, resulting in a gradual increase in airflow passing through them. During the closing phase, vocal folds start closing, which in turn results in a decrease in the airflow through them. The opening and closing phase together is referred to as the open phase ($T_o$). In general, the closed phase of glottal flow is relatively shorter than the open phase. During the opening phase the glottal flow starts increasing gently and then rapidly. Due to this, two instants are considered as opening instants, namely primary opening, $T_{o1}$ (end of the horizontal phase) and secondary opening, $T_{o2}$ (instant of abrupt increase of flow derivative). Time-domain features comprise open quotients, closing quotients, and speed quotients.

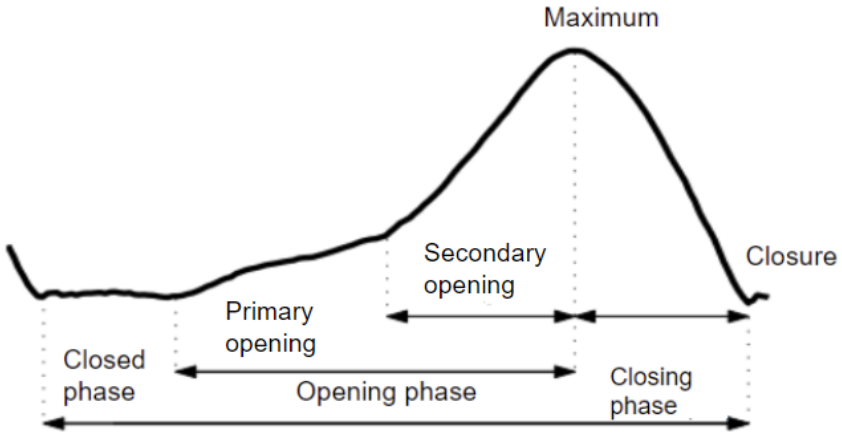



Figure 3.9: Glottal flow waveform with primary and secondary opening. The length of the glottal cycle is denoted by T. The time duration from the primary opening to the instant of maximum flow is denoted by $T_{o1}$ and the time duration from the secondary opening to the instant of maximum flow by $T_{o2}$. The closing phase length is denoted by $T_{cl}$ [62].

- **Open quotient calculated from the primary glottal opening (OQ1)**: It is defined as the ratio of the time duration of the primary open phase (sum of the primary opening and closing phase) to the total time duration of one glottal cycle.

$$OQ1 = \frac{T_{o1} + T_{cl}}{T} \tag{3.1}$$

- **Open quotient calculated from the secondary glottal opening (OQ2)**: Ratio of time duration of the secondary open phase (sum of the secondary opening and closing phase) to the total time duration is termed as OQ2. It is given by

$$OQ2 = \frac{T_{o2} + T_{cl}}{T} \tag{3.2}$$

- **Closing quotient (ClQ)**: It is defined as the ratio of the time duration of the closing phase duration to the duration of the glottal cycle.

$$ClQ = \frac{T_{cl}}{T} \tag{3.3}$$

  where $T = T_o + T_c + T_{cl}$ represents one glottal cycle.

- **Speed quotient, calculated from the primary glottal opening (SQ1)**: It is defined as the ratio of time duration of the primary opening phase to the closing phase.

$$SQ1 = \frac{T_{01}}{T_{cl}} \tag{3.4}$$

- **Speed quotient, calculated from the secondary glottal opening (SQ2)**: It is defined as the ratio of time duration of the primary opening phase to the closing phase.

$$SQ2 = \frac{T_{02}}{T_{cl}} \tag{3.5}$$

- **Quasi open quotient (QoQ)**: It is defined as ratio of the the quasi open phase of the glottis to quasi closed phase.

$$QoQ = \frac{Q_{oT}}{Q_{clt}} \tag{3.6}$$

- **Amplitude-based open quotient (OQa)**: It is Variation of open quotient derived from Liljencrants-Fant (LF-model).

$$OQa = f_{ac}(\frac{\pi}{2d_{max}} + \frac{1}{d_{min}})F0 \tag{3.7}$$

where $d_{max}$ is defined as maximum positive amplitude of differentiated glottal pulse derived from LF-model. $F0$ is fundamental frequency of vocal fold vibration. Figure 3.10 show the glottal flow and its derivative waveform used for the calculation of amplitude quotients. .

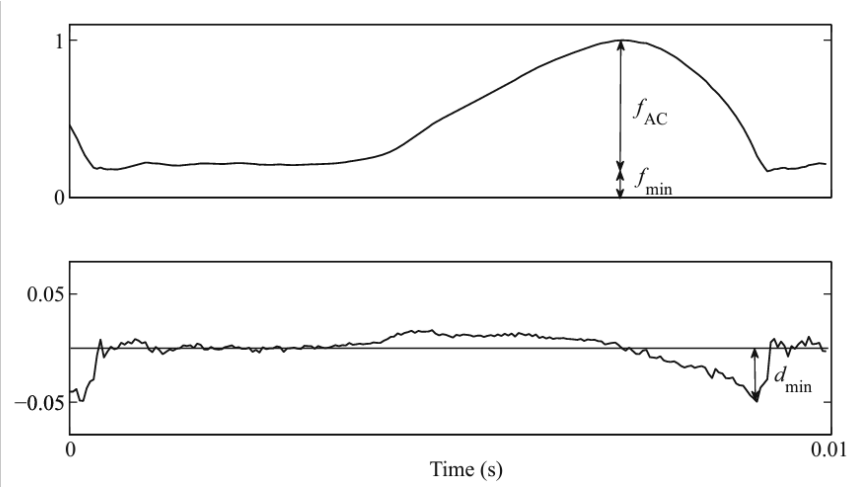



Figure 3.10: Glottal flow (at the top) and its derivative waveform (bottom). $f_{AC}$ is the AC amplitude of the glottal flow waveform, and $d_{min}$ is the negative peak amplitude of the glottal flow derivative [62].

- **Amplitude quotient (AQ)**: It is defined as the ratio between the AC-amplitude of the glottal flow signal and the amplitude of the minimum of the derivative of the glottal flow signal.

$$AQ = \frac{f_{AC}}{d_{min}} \tag{3.8}$$

- **Normalized amplitude quotient (NAQ)**: Amplitude quotient when normalized with respect to the length of the fundamental period of the glottal cycle termed as NAQ. It is given by

$$NAQ = \frac{f_{AC}}{d_{min}.T} \tag{3.9}$$

Table 3.1: Time-domain glottal features derived from GVV waveform.

| Feature | Description |
|---|---|
| OQ1 | Open quotient, derived from the primary glottal opening |
| OQ2 | Open quotient, derived from the secondary glottal opening |
| OQa | Open quotient, calculated from the LF model |
| QoQ | Quasi-open quotient |
| AQ | Amplitude quotient |
| NAQ | Normalized amplitude quotient |
| ClQ | Closing quotient |
| SQ1 | Speed quotient, calculated from the primary glottal opening |
| SQ2 | Speed quotient, calculated from the secondary glottal opening |

Table 3.1 shows the nine dimension features derived from the time-domain glottal waveform. If the glottal flow waveform does not show two different opening instants, then in that case OQ1 = OQ2 and SQ1 = SQ2.

2. **Frequency-domain features derived from glottal flow waveform**: Frequency domain parameters are derived by calculating the magnitude spectrum (in decibels) of the GVV signal. Spectrum is calculated by taking the fast Fourier transform (FFT) of the glottal signal. Figure 3.11 shows the frequency response (frequency versus amplitude in dB plot) of GVV signal. The most widely used features derived from the frequency response of GVV signal are harmonic richness factor (HRF), the difference in first and second harmonic (H1-H2), and parabolic spectral parameter (PSP).

   - **Harmonic richness factor (HRF)**: It is the ratio of the sum of the amplitudes of the harmonics above the fundamental frequency to the amplitude of the fundamental frequency.
   - **Difference in first and second harmonic H1-H2**: It indicates the slope of the glottal flow spectrum. It is the difference between the amplitude of the fundamental frequency and the second harmonic.

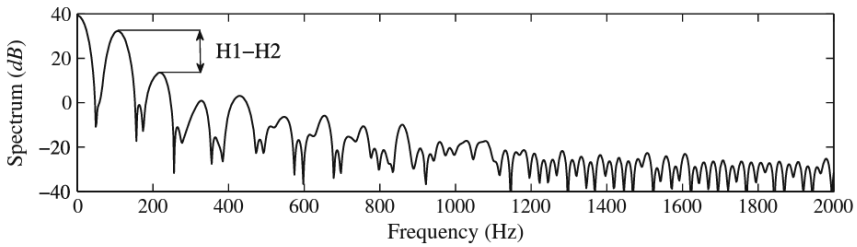



Figure 3.11: Frequency-domain representation of glottal flow waveform [62].

- **Parabolic spectral parameter(PSP)**: It is derived by matching the parabola function (second-order polynomial) to the spectrum of the GVV signal. PSP provides a single numerical value that characterizes the behaviour of the glottal flow's spectral decay compared to the maximum spectral decay theoretically achievable [159]. PSP is computed by fitting the parabola function to spectrum of the glottal flow waveform. This fitting is done by minimizing the mean square error between the discrete spectrum of glottal flow waveform denoted by $X(k)$ and the parabola function ($Y(k)$) should be minimized. The mean square error is given by:

$$E = \sum_{k=1}^{N-1} (X(k) - Y(k))^2 \tag{3.10}$$

Parabolic function is described as

$$Y(K) = ak^2 + b \tag{3.11}$$

Where 'a' and 'b' are constants that define the parabola. The constant 'a' determines the direction of parabola: if a is positive, the parabola opens upwards; if 'a' is negative, the parabola opens downwards. The constant 'b' shifts the parabola vertically.

$$E = \sum_{k=1}^{N-1} (X(k) - ak^2 - b)^2 \tag{3.12}$$

Parabolic spectral parameter is given by

$$PSP = \frac{a}{a_{max}} \tag{3.13}$$

Figure 3.12 shows examples of PSP computation derived from the glottal source spectrum: one from a male speaker with breathy phonation and the other from a female speaker with pressed phonation. It can be observed from the Figure 3.12(a), that spectral decay is large for male speaker which is matched by a parabolic function that decrease rapidly. Figure 3.12(b)) depicts slow spectral decay for female speaker which is modelled by a parabolic function with small steepness as compare to Figure 3.12(a). Table 3.2 shows the three dimension frequency-domain features derived from GVV waveform.

Table 3.2: Frequency-domain glottal features derived from GVV waveform [40].

| Feature | Description |
|---|---|
| H1-H2 | Amplitude difference between the first and second glottal harmonic |
| PSP | Parabolic spectral parameter |
| HRF | Harmonic richness factor |

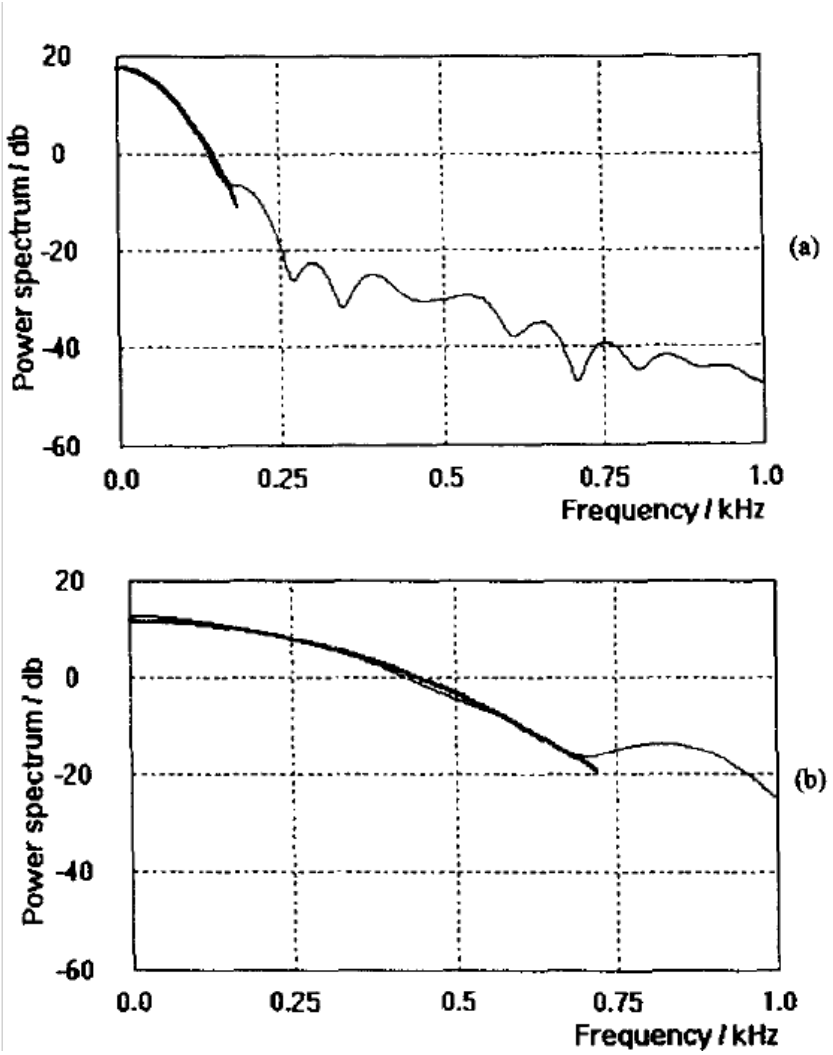



Figure 3.12: Pitch-synchronous spectrum of a glottal waveform (thin line) and the optimal parabolic match (thick line) [159].

#### 3.3.1.2 Intonation feature

Knowledge of epoch locations is important to obtain the perturbation measures corresponding to the vocal fold vibration. In this work, epoch locations are obtained from speech using zero frequency filtering (ZFF) technique [160]. This study used the epoch locations to find the fundamental frequency (F0) contour, strength of excitation (SoE) contour, and energy of excitation (EoE) contour of the ZFF signal. The F0, SoE, and EoE contours have been used to obtain 76-dimensional feature vector, which is referred to as an intonation feature vector (as in [161]) in this work.

- Fundamental frequency (F0): F0 is determined by calculating the epoch location derived from the ZFF method. The difference between the consecutive epoch location gives the measure of pitch period (T0) and the inverse of pitch period is fundamental frequency denoted by F0 [162, 141]. If $\mathcal{E}$= $\{e_1, e_2, e_3.....e_M\}$ is the number of GCI locations derived from the ZFF method, then $F_0$ is given by

$$F_0[n] = \frac{1}{T_0(n)} = \frac{f_s}{e_n - e_{n-1}}, n = 2, 3, .....M \tag{3.14}$$

  where $T_0[n]$ is the fundamental period of vocal fold vibration, $fs$ is sampling frequency and $M$ is number of epoch locations derived from ZFF method.

- Strength of excitation (SoE): The slope of ZFF signal around each epoch location is referred to as the strength of excitation which indicates the strength or intensity of GCI location. It is directly proportional to the rate at which the vocal folds close during phonation [141].

$$SoE = y[e_n + 1] - y[e_n - 1], n = 1, 2, 3, .....M \tag{3.15}$$

  where $y[n]$ is the output signal of the ZFF method.

- Energy of excitation (EoE) of ZFF signal: The mean square energy of the samples at GCI locations is defined as the energy of excitation, which gives the measure of vocal effort.

$$EoE = \sum_{i=-L/2}^{L/2} y^2[n + i], n = 1, 2, 3, .....M \tag{3.16}$$

  where y[n] is the ZFF signal, and L is the length of the window over which the energy is computed. L is taken as 10 ms for the calculation of energy.

Jitter is a cycle-to-cycle perturbation of the glottal cycle and is derived from the pitch period. Shimmer is the amplitude perturbation of the glottal cycle and is calculated from SoE and EoE. Table A.1 shows the intonation features and their corresponding feature dimension.

Table 3.3: Intonation feature and corresponding feature dimension [69].

| Feature | Dimension |
|---|---|
| Statistical measures of F0 | 5 |
| Jitter quotients of F0 | 22 |
| Shimmer quotients of strength of excitation (SOE) | 22 |
| Shimmer quotients of Energy of excitation (EOE) | 22 |
| Harmonic to noise ratio and noise to harmonic ratio | 4 |
| Pitch perturbation entropy (PPE) | 1 |

#### 3.3.1.3 Mel frequency cepstral coefficients of LP-residual, and ZFF signal

The studies in [163], revealed that Mel frequency cepstral coefficients (MFCC) of excitation source components are useful to identify the phonation type. Hence, this study explored the MFCC of LP-residual (MFCC-Residual) and ZFF signal (MFCC-ZFF) for the detection and identification of voice disorders. The MFCC-Residual and MFCC-ZFF features were obtained from segments of LP-residual and ZFF signal, respectively, with a frame-length of 20 ms and a frame shift of 5 ms. They are 39-dimensional cepstral coefficients consisting in 13 static coefficients and their first and second-order derivatives. Finally, 4 statistics, namely mean, standard deviation, kurtosis, and skewness, were calculated, resulting in 156-dimensional MFCC-Residual and MFCC-ZFF feature vectors.

### 3.3.2 Baseline features

1. **openSMILE feature set**: The open-source Speech and Music Interpretation by Large-space Extraction (OpenSMILE) is a publicly available toolkit for audio and music application designed for extracting acoustic features [164]. In our experiment, two feature sets of this toolkit are used as baseline features, namely ComParE feature set [165] and eGeMAPS feature set [166].

    - The 2013 Interspeech Computational Paralinguistics Challenge (ComParE) features set is a large-scale acoustic feature set with 6373 static paralinguistic features. These features are obtained by computing various statistical functions over low-level descriptor (LLD) contours. The ComParE feature set includes four energy-related parameters (such as zero crossing rate, RMS energy, and loudness), 55 spectral features (such as MfCC, spectral energy, spectral variance, skewness, and kurtosis), and six voicing-related features (such as jitter, shimmer, and HNR). The statistical functionals applied to the LLDs include mean, standard deviation, percentiles, quartiles, linear regression functionals, quadratic regression, and minima/maxima-related functionals.
    - extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) are small-scale (low dimension) knowledge-based acoustic feature set contains 88 parameters. Functionals are applied to 45 LLD. Frequency-related parameters are a total of (12) pitch, jitter, first three

formant frequencies and bandwidth of the first formant, their mean and standard deviations. Energy related parameters are 6, which includes loudness, shimmer, and Harmonic to noise ratios (HNR) mean and standard deviation. In total, it consists of 42 LLD on which two statistical functionals (arithmetic mean and coefficient of variations) are applied.

2. **Cepstral features**: Features extracted from speech signals that model the vocal tract information are considered as another baseline feature set in this study. MFCC [50], and PLP features are computed using speech segments of 20 ms frame size with a 5 ms frame shift. First 13 dimensional static features and corresponding delta, and delta-delta features were computed, resulting in 39-dimensional features. Statistical averages such as mean, standard deviation, kurtosis and skewness were derived from these frame-level features.

### 3.3.3 Database

SVD database contains the speech recording of healthy subjects as well as subjects suffering from voice disorders. It contains almost 71 different disorder categories. Categories that contain recordings of more than 30 subjects were grouped into four classes as shown in Table 3.4. In this study, all speech samples sampling frequency was down-sampled to 8000 Hz. Structural voice disorders are mainly due to anatomic abnormalities (like growth of the lesion, swelling of vocal cords) in the larynx. Therefore, laryngitis, leukoplakia, polyp, reinke's edema, contact granuloma, vocal cord polyp, cordectomy, and frontolateral partial resection are grouped to make structural class. Neurogenic voice disorders are caused due to damage or malfunction in the central or peripheral nervous system [167]. As the nervous system interacts with the larynx, it affects the functioning of the vocal mechanism. Spasmodic dysphonia and recurrent laryngeal nerve palsy are the two disorders that are considered in this category. Functional voice disorders (commonly known as muscle tension dysphonia) are characterized by excessive laryngeal activity, tension, reduced vocal capacity, and impaired voice without any organic abnormality [86]. Functional dysphonia, and hyperfunctional dysphonia are grouped into this class. In psychogenic voice disorders, the subject will lose control over the initiation and maintenance of phonation during speech production due to disturbed psychological processes like anxiety, depression, conversion reaction, or personality disorder [168, 169]. Psychogenic dysphonia is considered in the psychogenic voice disorder category.

### 3.3.4 Classifier

Support vector machine (SVM) classifier is the most widely used classifier in pathological voice detection as it gives consistence performance even on small dataset [170]. The present study used the SVM classifier for the detection and identification of voice disorders. This study performed classification by using other classifiers like decision tree, logistic regression, k-nearest neighbour, and ensemble classifier. Among all these classifiers, the SVM classifier outperforms for most of the tasks. Moreover, different kernel functions, such as linear, polynomial and radial basis functions, were also explored. The

Table 3.4: Details of the voice disorders considered from SVD database for performing voice disorder identification task.

| Voice disorder type | Disorder name | #Speakers | Total speakers |
|---|---|---|---|
| Structural | Laryngitis | 37 | 353 |
| | Leukoplakia | 109 | |
| | Polyp | 30 | |
| | Reinke's edema | 37 | |
| | Contact granuloma | 64 | |
| | Cordectomy | 42 | |
| | Frontolateral partial resection | 32 | |
| Neurogenic | Spasmodic Dysphonia | 192 | 253 |
| | Recurrent laryngeal nerve palsy | 61 | |
| Functional | Functional dysphonia | 98 | 254 |
| | Hyperfunctional dysphonia | 154 | |
| Psychogenic | Psychogenic Dysphonia | 91 | 91 |

best performance was observed with a polynomial kernel of order 2. Further, the grid search approach is explored to select the best parameters for the quadratic kernel. In this regard, the kernel parameter (box constraint level) is changed from 0.1 to 1000 with multiples of 10 and the kernel parameters for which the classifier has the best classification accuracy are considered for further analysis. The experiments were conducted with five-fold cross-validation and the average classification accuracy of all folds is referred to as the performance of the system.

## 3.4 Results and discussion

The main objective of this work is to assess voice disorders in a clinical approach. This study explored the excitation source features (MFCC-Residual, MFCC-ZFF, Glottal, and Intonation features) for the identification of voice disorders and compared their performance with baseline features, namely vocal-tract system features (MFCC and PLP) and OpenSMILE features (ComParE and eGeMAPS) discussed in Subsection 3.3. In this regard, classification systems for the detection and identification of voice disorders are developed by using SVM classifier (discussed in Subsection 3.3) with individual excitation source feature sets and baseline feature sets. In this study, five-fold cross-validation is used so that the recordings correspond to 80% and 20% of total speakers were used as training and testing data, respectively. A total of four experiments were conducted in speaker independent approach using SVD database (discussed in Subsection 3.1). In all the experiments, binary classification systems are trained with different feature sets and corresponding results are tabulated in Table 3.5.

- Experiment 1 (Voice disorder detection) was performed to discriminate healthy voice samples from the voice disorder sample of all the classes.

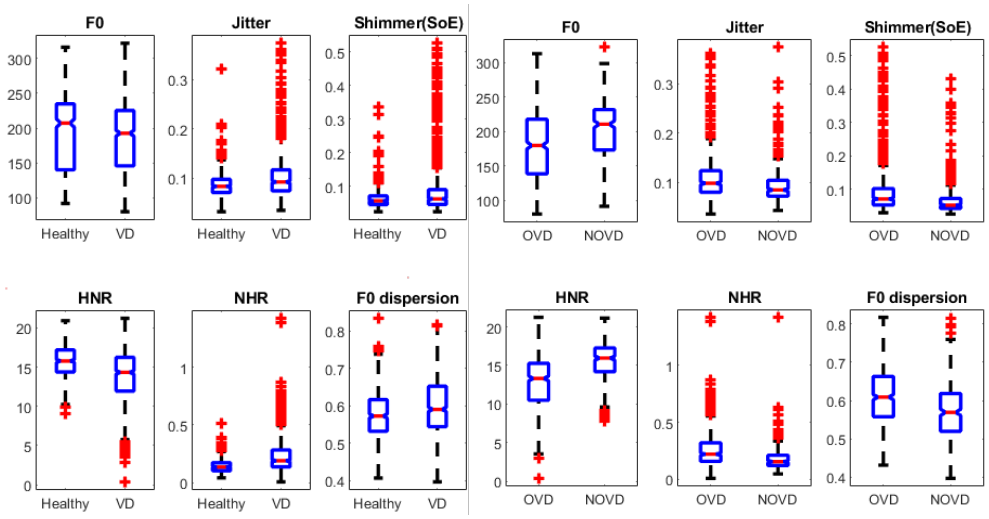
- In experiment 2, Organic voice disorder samples were classified from non-organic voice disorder samples.
- In experiment 3, Organic voice disorder samples were further classified into structural and neurogenic voice disorders.
- Experiment 4 was conducted to classify functional voice disorders from the psychogenic voice disorder category.

Table 3.5: Performance of voice disorder detection and identification systems in terms of classification accuracy (in %) for individual feature set on SVD database. Here, Exp. 1: classification of healthy and voice disorders, Exp. 2: classification of organic and non-organic voice disorders, Exp. 3: classification of structural and neurogenic voice disorders, and Exp. 4: classification of functional and psychogenic voice disorders.

| Feature type | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 |
|---|---|---|---|---|
| ComParE | **82.8** | 71.7 | **74.3** | **65.3** |
| eGeMAPS | 76.0 | 70.1 | 67.3 | 57.5 |
| MFCC | 74.4 | 72.4 | 67.8 | 63.4 |
| PLP | 74.2 | **72.7** | 70.5 | 64.1 |
| Glottal | 67.4 | 64.8 | 59.9 | 58.3 |
| Intonation | 69.3 | 66.0 | 60.2 | 52.8 |
| MFCC-Residual | 67.4 | 70.8 | 64.3 | 61.0 |
| MFCC-ZFF | 68.5 | 69.2 | 66.4 | 64.2 |

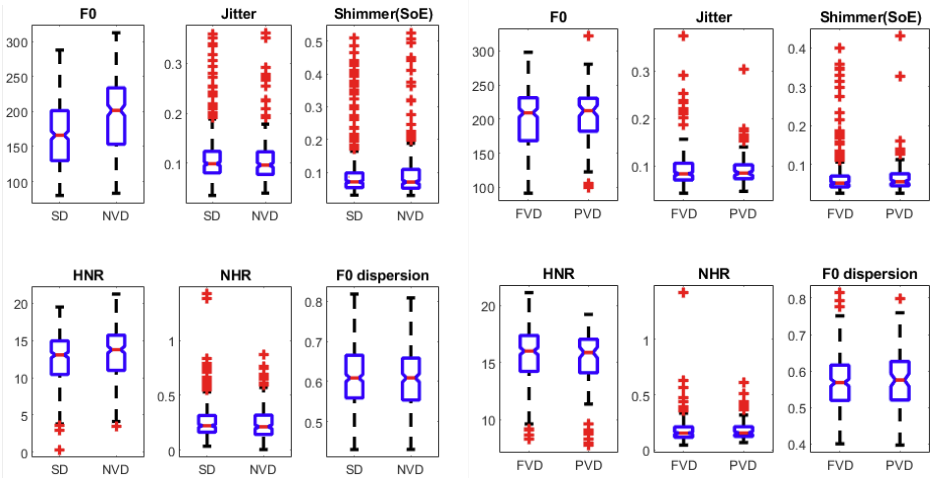
From Table 3.5, it is observed that among all individual excitation source feature sets, intonation features show the best performance for experiment 1 with a classification accuracy of 69.3%. From this, it is anticipated that perturbation parameters capture voice disorder information in a better way. On the other hand, cepstral features extracted from excitation source evidence performed best for experiments 2, 3, and 4. with the classification accuracy of 70.8%, 66.4%, and 64.2%, respectively. From this, it can be concluded that features extracted from the excitation source can capture the information that can discriminate pathological speech from healthy speech. It is observed that among all baseline feature sets ComParE feature set shows the best performance in experiments 1, 3, and 4, while PLP feature produced a better performance in experiment 2 than all other individual features. However, in all the experiments the performance of excitation source features was shown to be lower than the baseline features. Among the individual features, ComParE feature set showed the best performance in most of the experiments. However, it is a brute-forced acoustic feature set that has a very high dimension (6373) compared to the other feature sets.

From Figure 3.13, it can be seen that perturbation parameters effectively discriminate between healthy subjects and those with voice disorders due to differences in acoustic characteristics. Parameters such as jitter, shimmer, NHR, and F0 dispersion exhibit higher values in subjects with voice disorders




(a) Healthy Vs VD. (b) OVD Vs NOVD.

Figure 3.13: Distribution of intonation features for healthy and voice disorder subjects. The horizontal line within the box denotes the median, and the box covers one-quarter of the data on either side of the median. The whiskers on either side cover all points within 1.5 times the interquartile range (width of the box), and points beyond these whiskers are plotted as outliers.




(a) OVD. (b) NOVD.

Figure 3.14: Distribution of intonation features for different categories of voice disorder. The horizontal line within the box denotes the median, and the box covers one-quarter of the data on either side of the median. The whiskers on either side cover all points within 1.5 times the interquartile range (width of the box), and points beyond these whiskers are plotted as outliers.

compared to healthy subjects, likely due to the instability of vocal fold vibrations. Conversely, HNR is higher in healthy subjects, reflecting the regular vibration of their vocal folds.

The box plot depicted in Figure 3.14 illustrates the distribution of intonation features for different categories of voice disorders. It can be observed from the Figure 3.14 value of jitter, shimmer, and F0 dispersion is high for structural voice disorder and low for neurological voice disorder. This is because vocal fold vibrations are more irregular for SD than NVD. Furthermore, the box plot indicates that most of these features effectively differentiate between SD and NVD, unlike FVD and PVD. These findings suggest that distinguishing PVD requires considering both acoustic information and the subject's medical history to determine if the voice disorder is associated with psychogenic trauma.

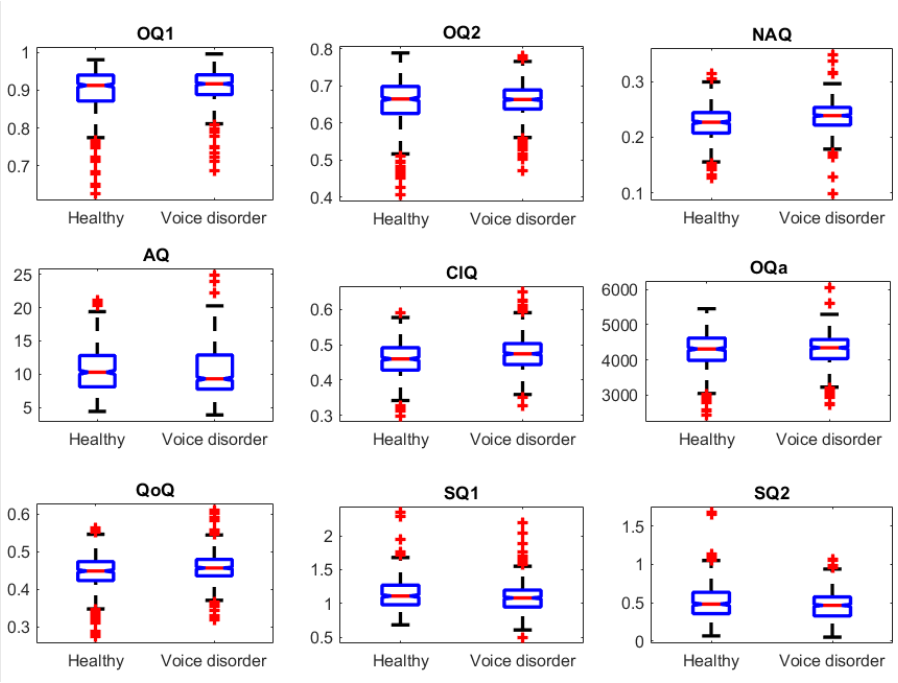



Figure 3.15: Distribution of time-domain glottal features for healthy and voice disorder subjects. The horizontal line within the box denotes the median, and the box covers one-quarter of the data on either side of the median. The whiskers on either side cover all points within 1.5 times the interquartile range (width of the box), and points beyond these whiskers are plotted as outliers.

Figure 3.15 shows the time domain glottal features derived from the QCP method for healthy and voice disorder subjects. Vocal folds do not close completely for subjects suffering from voice disorders, which results in a comparatively large OQ [78, 171] than healthy subjects. GVV signal of the healthy subject is described by a right-skewed glottal pulse, indicating that the decrease of the airflow (vocal folds close faster) is faster than the increase of airflow ( opening of vocal folds). Hence, CQ of the healthy subject is indicated by a large value and a small value of OQ, as seen in Figure 3.15. Compared to modal and pressed phonation, a comparatively large amount of glottal flow (AC amplitude) is observed for breathy phonation [172]. Time domain parameters indicate the amount of glottal flow is AQ and

NAQ. NAQ has a larger range, while AQ shows a smaller range for voice disorder as compared to healthy subjects. SQ is the ratio of the glottal opening phase to the duration of the glottal closing phase and indicates the skewness of the glottal pulse. Breathy phonation is described by symmetric glottal pulse [159]. SQ is higher for healthy subjects, as the glottal pulse is more asymmetrical compared to subjects with voice disorder.

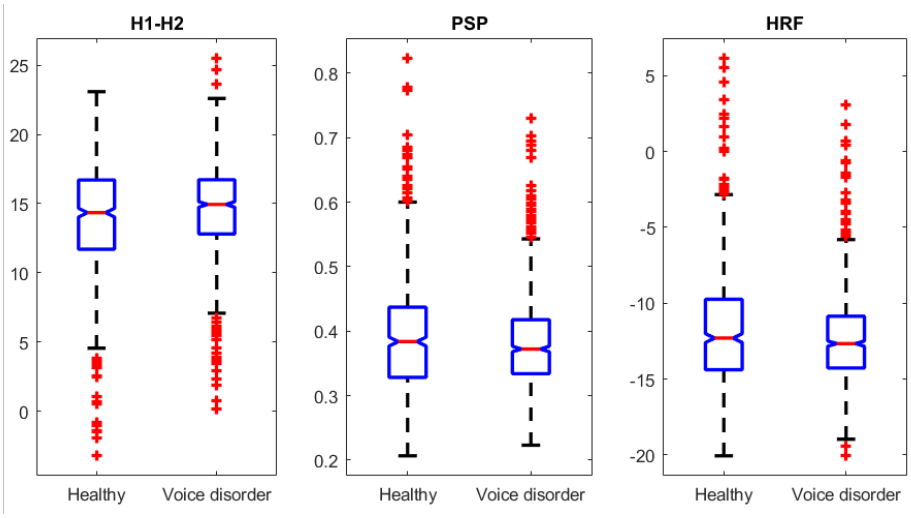



Figure 3.16: Distribution of frequency-domain glottal features for healthy and voice disorder subjects. The horizontal line within the box denotes the median, and the box covers one-quarter of the data on either side of the median. The whiskers on either side cover all points within 1.5 times the interquartile range (width of the box), and points beyond these whiskers are plotted as outliers.

For the GVV signal with a small value of OQ in the time domain, its frequency domain signal is said to have a strong second harmonic [1]. It was also shown that more skewness in the GVV signal would have a strong third harmonic in the spectrum [1]. For subjects suffering from voice disorders (due to incomplete closure), there is a large value of OQ compared to healthy subjects, resulting in a comparatively small second harmonic. Hence H1-H2 is higher for individuals with voice disorders (as shown in Figure 3.16). The presence of regular vocal fold vibration results in a large value of HRF for healthy subjects as compared voice disorder subjects. PSP indicates larger value for healthy subjects as compared to those with voice disorder. More importantly, it can be observed from all these box plots that intonation features are more effective in distinguishing between healthy subjects and individuals with voice disorders compared to glottal features. The same observation can be noticed from the results in Table 3.5 in terms of classification accuracy.

Figure 3.17, represents the speech signal, ZFF signal, F0 contour, and SoE contour derived from ZFF method for three different groups: health, OVD and NOVD, respectively, for neutral vowel /a/. It can be observed from Figure 3.17 that for the subject suffering from voice disorder, the variation in the F0 contour and SoE is more compared to a healthy subject. Moreover, these parameters show significant differences between the different categories of voice disorders.

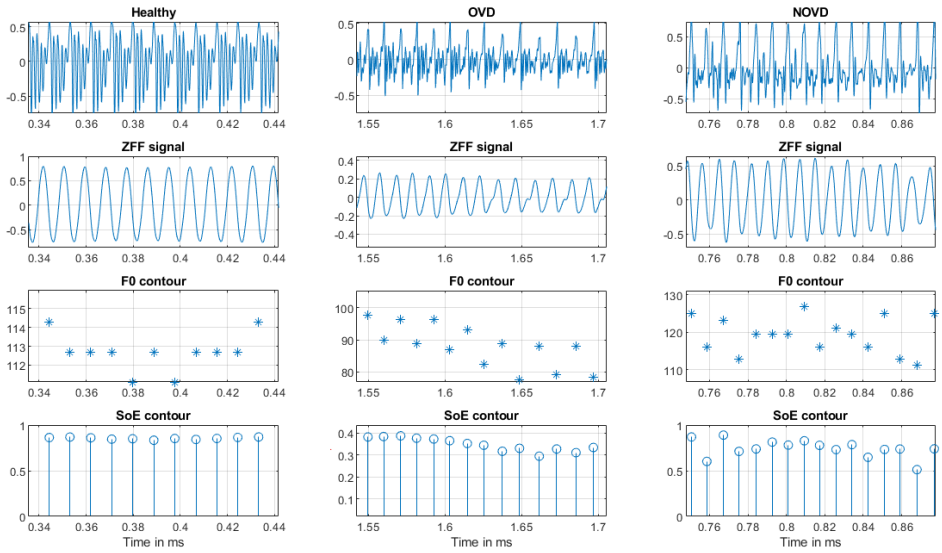



Figure 3.17: Illustration of the output signal received from the ZFF method for healthy subjects and subjects suffering from organic and non-organic voice disorders, respectively, for neutral vowel /a/.

Table 3.6: Performance of voice disorder detection and identification systems in terms of classification accuracy (in %) for combination of feature sets on SVD database. Here, Exp. 1: classification of healthy and voice disorders, Exp. 2: classification of organic and non-organic voice disorders, Exp. 3: classification of structural and neurogenic voice disorders, and Exp. 4: classification of functional and psychogenic voice disorders.

| Feature type | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 |
|---|---|---|---|---|
| Glottal + ComParE | **85.2** | 72.7 | 73.1 | 59.2 |
| Glottal + eGeMAPS | 79.0 | 70.8 | 65.5 | 60.1 |
| Glottal + MFCC | 74.4 | 71.2 | 66.7 | 64.1 |
| Glottal + PLP | 78.0 | 71.5 | 67.8 | 63.0 |
| Intonation + ComParE | 84.9 | 72.8 | 74.9 | 60.3 |
| Intonation + eGeMAPS | 81.5 | 68.5 | 68.1 | 60.1 |
| Intonation + MFCC | 77.5 | **75.0** | 65.2 | 64.4 |
| Intonation + PLP | 77.6 | 72.7 | 69.3 | 62.4 |
| MFCC-Residual + ComParE | 84.1 | 73.0 | **76.0** | 65.0 |
| MFCC-Residual + eGeMAPS | 84.3 | 70.9 | 62.6 | 63.3 |
| MFCC-Residual + MFCC | 73.1 | 74.6 | 69.6 | 66.2 |
| MFCC-residual + PLP | 74.2 | 73.0 | 68.4 | 65.3 |
| MFCC-ZFF + ComParE | 84.5 | 72.3 | 74.0 | 67.3 |
| MFCC-ZFF + eGeMAPS | 84.3 | 71.8 | 67.5 | 62.1 |
| MFCC-ZFF + MFCC | 71.7 | 72.3 | 68.7 | 63.6 |
| MFCC-ZFF + PLP | 74.4 | 70.1 | 70.5 | 65.9 |
| Glottal + Intonation + MFCC-Residual + MFCC-ZFF | 75.6 | 72.4 | 67.0 | **70.0** |

Further, experiments have been performed using combinations of feature sets to investigate the complementary nature of excitation source features and baseline feature sets. In voice disorder detection, ComParE with glottal feature combination produced the best classification accuracy of 85.2%. Intonation features with MFCC, MFCC-Residual with ComParE, and a combination of all excitation source feature sets produced the best classification accuracies 75%, 76% and 70% in experiments 2, 3 and 4, respectively. In most of the experiments, a combination of baseline features (ComParE, eGeMAPS, PLP and MFCC feature sets) with excitation source feature sets showed significant improvement in the performance of identification systems trained with individual baseline feature sets. It indicates that excitation source features capture complementary information about voice disorders compared to baseline features. Results of the present study reveal that the detection of voice disorders has a higher classification accuracy than the identification of voice disorders. Moreover, the classification of functional and psychogenic voice disorders is more challenging compared to the classification of structural and neurogenic voice disorders.

## 3.5 Conclusions

This chapter proposed a hierarchical approach using excitation source features for the automatic detection and identification of voice disorders from a clinical perspective. A more detailed analysis of voice disorders was performed to know whether the disorder is structural, neurogenic, functional or psychogenic. Excitation source features used in these experiments are intonation features, glottal features, MFCC-Residual and MFCC-ZFF. Excitation source features were compared with state-of-art MFCC, PLP, ComParE and eGeMAPs features. Among the individual features, ComParE feature set shows the best performance in most of the experiments. However, it is a brute-forced acoustic feature set that has a very high dimension (6373) compared to the other feature sets. In most of the experiments, a combination of baseline features (ComParE, eGeMAPS, PLP and MFCC feature sets) with excitation source feature sets showed significant improvement in the performance of identification systems trained with individual baseline feature sets. It indicates that excitation source features capture complementary information about voice disorders compared to baseline features. For experiment 4, when all source features were combined, the functional and psychogenic voice disorder classification system outperformed with a classification accuracy of 70%

Results of the present study reveal that the detection of voice disorders has a higher classification accuracy than the identification of voice disorders. Moreover, the classification of functional and psychogenic voice disorders is more challenging than the classification of structural and neurogenic voice disorders. From this chapter, it can be understood that features derived from the excitation source signal can discriminate different categories of voice disorder.

*Chapter 4*

# Analysis of epoch extraction methods for different categories of voice disorders

Based on the results of the studies done in Chapter 3, it can be concluded that information related to voice disorders is captured in excitation source [173]. Computation of various excitation source features such as jitter, shimmer, glottal parameters etc. involve the detection of epoch locations from speech signal. Therefore, precise determination of epoch locations plays a significant role in calculating these features for the automated detection and identification of voice disorders. This chapter analyses the different epoch extraction methods for different categories of voice disorder.

The studies in [174, 175], show that the performance of state-of-the-art epoch extraction methods is efficient in clean speech conditions. Efficacy of epoch extraction methods has been studied for telephonic quality speech [176, 177, 178], emotional speech [179, 180], and the degraded speech obtained by corrupting the clean speech with additive noise and reverberations [181, 182]. In general, the performance of these methods has been evaluated using speech utterances produced by healthy (controlled) speakers. On the other hand, the subjects suffering from voice disorders will not be able to produce normal or modal phonation [85]. Hence, the performance of the existing epoch extraction methods may vary in processing of speech associated with voice disorders due to the variations in the glottal source characteristics such as roughness, breathiness, hoarseness, abnormality in pitch and strained quality [11, 173]. In literature, the performance of epoch extraction methods was not studied for the speech associated with voice disorders. Hence, this chapter aims to compare the performance of various state-of-the-art algorithms for extracting epoch locations from speech associated with voice disorders. Moreover, the performance of a GCI detection method may vary depending on the type of voice disorder because each voice disorder can affect the phonation process in a different way. Hence, this study is also intended to investigate the performance of the epoch extraction methods for different categories of voice disorders by using SVD database [150]. It was observed from the first study that performance of the state-of-the-art epoch extraction methods degrades for different categories of voice disorders. Then the performance was also observed by applying the region-based pre-processing to the existing methods. Finally, the performance of voice disorder detection and identification system was observed with the application of region-based processing.

Rest of the chapter is organised as follows. Section 4.1 compares the performance of state-of-the-art epoch extraction methods for healthy and voice disorders subjects. Section 4.2 presents the application of the region-based processing on the state-of-the-art epoch extraction methods. Section 4.3 discusses the performance of voice disorder detection and identification system using the features extracted from the excitation source evidence after applying the region-base processing on them. Finally, the summary and conclusions of the study are described in Section 4.4.

## 4.1 Comparison of the state-of-the-art epoch extraction algorithm for different categories of voice disorders

### 4.1.1 State-of-the-art epoch extraction algorithms

In this study state-of-the-art methods of epoch extraction like Zero frequency filtering (ZFF) [56], Zero phase-zero frequency filtering (ZP-ZFF) [57], Speech event detection using the residual excitation and a mean-based signal (SEDREAMS) [174], Dynamic programming phase slope algorithm (DYPSA) [183], Yet another GCI algorithm (YAGA) [184], SEDREAMS-voice quality, Glottal closure/opening instant estimation using forward-backward algorithm (GEFBA) method, and Continuous wavelet transform-glottal closure instant (CWT-GCI) method are considered for evaluating the performance of different categories of voice disorders.

- **ZFF** method is based on the fact that vocal tract resonances are predominantly present at high frequency [184], while the discontinuity due to impulse-like nature of glottal excitation is present at all the frequencies including zero frequency. Hence, speech signal is passed through a zero frequency resonator which is low pass filter with poles located inside the unit circle. The resultant filtered signal will preserve the excitation source characteristics, at the same time high frequency resonances of vocal tract system are attenuated [58]. Output of the filter shows polynomial growth/decay, which can be removed by passing this filtered signal through trend removal (moving average) filter of length one to two pitch period. Trend removal filter effectively removes the growing/decaying trend present in the filtered signal, which in turn highlights the fluctuations caused due to impulse-like excitation. The output signal of the trend removal filter is referred to as zero frequency filtered signal. Positive to negative zero crossing of zero frequency filtered signal is marked as epoch.

- **ZP-ZFF** method is stable implementation of ZFF. ZFR used in this method, has it's poles located inside the unit circle, which make the filter stable and anti-causal infinite impulse response (IIR) filter.

- **SEDREAMS** method relies on mean based signal and residual signal for epoch extraction. First the short intervals at which epochs are expected to occur are determined from the mean-based

signal. Then LP residual signal is derived from the speech signal to capture the excitation source characteristics. And as final step, intervals extracted from the mean-based signal are combined with a peak detected from LP residual to accurately detect the GCI locations.

- **DYPSA** algorithm uses three steps to perform the epoch detection. First the candidate GCIs are detected from zero crossing of phase-slope function. Then missed GCIs are recovered from the phase-slope projection technique. In this projection technique, first it is detected that if a local minimum is followed by a local maximum without zero-crossing. Then the midpoint between these two point is projected with unit slope on the time axis to identify the GCIs which were missed out from the previous step. Then as a final step, true GCIs are detected using dynamic programming.

- **YAGA** methods is performed in two phases, candidate detection and candidate selection. In it's first phase GCIs are detected from the speech signal, and then dynamic programming is performed to select true GCIs from the candidate set. To calculate the candidate GCIs, first voice source signal is derived using iterative adaptive inverse filtering (IAIF) method. Now from this signal multi-scale product of the stationary wavelet transform (SWT) is derived to highlight the discontinuity presents in the signal followed by estimation of group delay function. Negative-going zero crossings of group delay function are marked as GCIs candidate. As last step dynamic programming algorithm is applied to detect true GCIs.

- **SE-VQ Method** algorithm was proposed to handle the different phonation type,, which is a modified form of SEDREAMS algorithm [139]. In this method, two extra steps are introduced as compare to basic SEDREAMS algorithm they are: dynamic programming and post-processing. Dynamic programming is applied to select the optimal GCI locations based on the strength of peaks in LP residual and transition cost (i.e. transition from one GCI to another GCI). Further, post-processing is applied to minimize the false positive GCIs location and to preserves the true positive GCIs. In the SEDREAMS only one peak which is the highest peak from LP residual is chosen, while in the SE-VQ, several LP residual peaks are selected in order to handle the voice quality like breathy and harsh where there are no prominent peaks.

- **GEFBA Method** is based on source signal obtained by linear prediction based inverse filtering [181]. This algorithm is performed in two phases. In the first phase, the glottal flow derivative is derived from inverse filtering based on LP analysis. Finally, in the second phase of GEFBA algorithm, a forward and move backward algorithm is performed on each voiced frame to estimate GCIs.

- **CWT-GCI Method** is based on the principle that CWT is a suitable method for determining the sharp transition from the signal [177]. In this method, to compute GCIs CWT coefficients are calculated from the analytic signal instead of speech signal. From these coefficients, the average absolute signal is obtained, and this signal is convoluted with a Gaussian filter to highlights the

peaks. The convoluted output is referred to as evidence to estimate the epoch locations. Spurious peaks are removed from the evidence signal by considering that time difference between the two consecutive peaks is not less than 2 ms. After removing the spurious peaks, positive peaks obtained from epoch evidence signal are referred to as epoch locations.

- **SPF Method** of epoch extraction is based on the estimation of time-frequency representation obtained from single pole filter (SPF) [176]. Single pole filter is a narrow band IIR filter, with pole located inside the unit circle. In this approach, first, the speech signal is passed through the bank of single-pole filters, which gives better time-frequency representation of the speech signal. From this time-frequency representation, time marginal is derived. Further, the time marginal is smoothed using a Gaussian window of 8 ms. Finally, positive crossings obtained from the smoothed time marginal, which are referred to as epoch locations.

### 4.1.2 Database

In SVD database [150], for each of the speech recordings, simultaneous EGG signals are available to obtain the ground truth epoch locations. Therefore, this study used the SVD database to evaluate the performance of epoch extraction algorithms. This is a publicly available database, can be downloaded from the site `http://www.stimmdatenbank.coli.uni-saarland.de/`. The present study considered the speech recordings from 687 healthy subjects and 679 subjects with different voice disorders from the SVD database. Each recording includes vowels /a/, /i/ and /u/ produced at a normal, low, and high pitch and also with rising-falling pitch. Also, each recording consists of a German sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). The SVD database was recorded at a sampling frequency of 50 kHz. In this study, all recordings were down-sampled to 8000 Hz. Additionally, the speech recordings correspond to 679 subjects with different voice disorders were categorized into four sub-classes, namely, structural, neurogenic, functional, and psychogenic (as discussed in Chapter 3). Further details, about each of the sub-classes are provided in Table 4.1.

### 4.1.3 Evaluation Metrics

Reference GCI locations are obtained from EGG signal. First difference EGG (dEGG) signal is obtained from EGG signal by calculating the successive sample difference and then peaks detected from this signal are marked as reference GCI locations. GCI locations obtained from speech signal (by any of the method) is termed as estimated GCI. To evaluate the performance of epoch extraction methods both GCIs (reference and estimated) are compared using the different parameters (as was done in) [184] in one larynx cycle. Identification rate(IR), miss rate (MR), false alarm rate (FAR), and identification accuracy (IA) are the popular metrics that are used for evaluating the performance of epoch extraction methods [184]. Hence, in this study, we considered these metrics for evaluating the performance of epoch extraction methods. Figure 4.1 shows the three larynx cycles of reference and estimated GCIs.

Table 4.1: Details of the voice disorders considered from SVD database for evaluating the epoch extraction algorithms. Here, FD: Functional dysphonia, PD: Psychogenic dysphonia, RLNP: Recurrent laryngeal nerve palsy, and SD: Spasmodic dysphonia,

| Voice disorder type | Disorder name | #Speakers |
|---|---|---|
| Structural | Laryngitis | 30 |
| | Leukoplakia | 41 |
| | Polyp | 45 |
| Neurogenic | SD | 30 |
| | RLNP | 188 |
| Functional | FD | 254 |
| Psychogenic | PD | 91 |

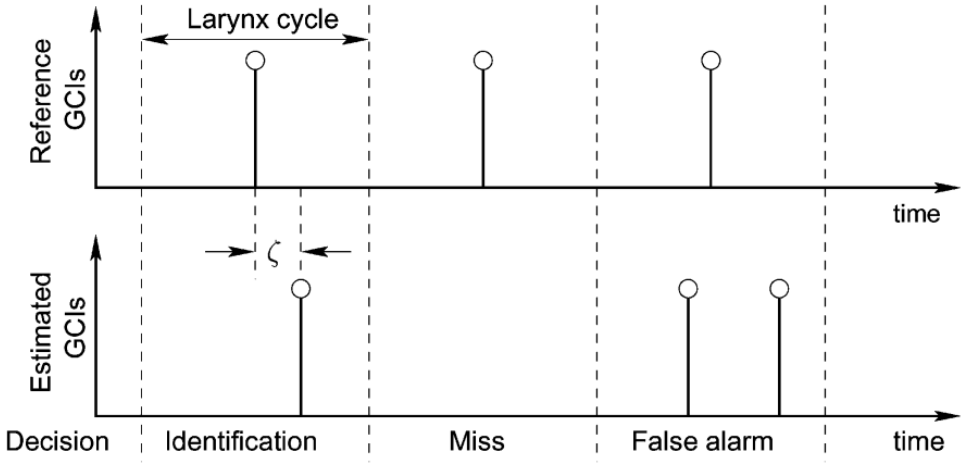



Figure 4.1: Comparison of larynx cycles of reference and estimated GCIs with possible outcomes [184].

- **Larynx cycle**: Larynx cycle (n) is defined (in terms of sample) as a range of samples, $(1/2)(n_{r-1}+n_r) < n < (1/2)(n_{r+1}+n_r)$, where $n_r$ represents the reference GCI locations while $n_{r-1}$ and $n_{r+1}$ represents the preceding and following GCI location.

- ***Identification rate***: It is defined as percentage of larynx cycle in which exactly one GCI location is identified.

- ***Miss rate***: It is defined as percentage of larynx cycle for which GCI is not detected (or missed).

- ***False alarm rate***: It is defined as percentage of larynx cycle in which multiple GCI locations are identified.

- ***Identification accuracy***: It is defined as time difference between reference GCI location and estimated GCI location for the cycle in which exact one GCI location was identified.

### 4.1.4 Results and Discussion

In this section, we compared the performance of nine state-of-the-art epoch extraction methods for the speech of healthy subjects and the speech of subjects with various voice disorders, using the SVD database, which provides simultaneous EGG recordings. The performance of each method is evaluated in terms of IDR, MR, FAR and IDA. The performance evaluation measures of different epoch extraction algorithms from healthy speech and speech associated with voice disorders on SVD dataset is reported in Table 4.2. In addition, the performance of the epoch extraction algorithms was studied for each of the four broad categories of voice disorders (structural, neurogenic, functional, and psychogenic), and the evaluation measures were reported in Table 4.3.

From the results presented in Table 4.2, it is evident that most of the epoch extraction methods (except SE-VQ, CWT-GCI, SPF, and GEFBA) work well for the healthy scenario, in which speech is produced under modal phonation. However, all epoch extraction methods show significant degradation in their performance for speech associated with voice disorders compared to healthy speech. Compared to the healthy scenario, in the voice disorder scenario, all the epoch extraction algorithms shown approximately 5 to 8% absolute reduction in IDR and (0.05 to 0.15) ms absolute increase in IDA. Among all epoch extraction methods, SEDREAMS and ZP-ZFF methods performed better in both healthy and voice disorder scenarios, in terms of IDR, FAR, and IDA. In the healthy scenario, SEDREAMS method shows the best performance in terms of IDR of 97.69%, whereas, ZP-ZFF method shown to be second best with an IDR of 97.63%. In the voice disorder scenario, the ZP-ZFF method showed the best performance in terms of IDR of 90.37% and IDA of 0.34 ms, while the ZFF and SEDREAMS methods showed IDR of 89.96% each one, which is almost equivalent to the IDR of ZP-ZFF. On the other hand, the DYPSA and YAGA methods showed comparable results in terms of IDR.

From the results reported in Table 4.3, it can be understood that among all the categories of voice disorders for the structural and neurogenic categories, the performance of all epoch extraction algorithms

Table 4.2: Performance evaluation of different epoch extraction methods for speech of healthy speakers and speech of speakers with voice disorder on SVD dataset. IDR–Identification rate, MR–Miss rate, FAR–False Alarm Rate, IDA–Identification Accuracy.

| Class | Method | IDR (%) | MR (%) | FAR (%) | IDA (ms) |
|---|---|---|---|---|---|
| Healthy | ZP-ZFF | 97.63 | 1.16 | **1.21** | 0.26 |
| | ZFF | 96.94 | **0.75** | 2.31 | 0.42 |
| | DYPSA | 95.45 | 1.42 | 3.13 | **0.23** |
| | YAGA | 96.22 | 1.03 | 2.75 | 0.66 |
| | SEDREAMS | **97.69** | 0.87 | 1.44 | 0.28 |
| | SE-VQ | 78.36 | 16.12 | 5.52 | 0.85 |
| | CWT-GCI | 92.01 | 6.35 | 1.65 | 0.45 |
| | SPF | 87.19 | 10.47 | 2.34 | 0.43 |
| | GEFBA | 72.77 | 22.09 | 5.14 | 0.54 |
| Voice disorders | ZP-ZFF | **90.37** | 4.03 | 5.6 | **0.34** |
| | ZFF | 89.96 | 3.79 | 6.25 | 0.46 |
| | DYPSA | 88.06 | 4.57 | 7.37 | 0.36 |
| | YAGA | 88.1 | **3.62** | 8.28 | 0.68 |
| | SEDREAMS | 89.96 | 4.44 | 5.59 | 0.39 |
| | SE-VQ | 74.01 | 19.05 | 6.93 | 0.91 |
| | CWT-GCI | 85.77 | 9.64 | **4.59** | 0.56 |
| | SPF | 81.27 | 13.79 | 4.93 | 0.59 |
| | GEFBA | 64.96 | 27.02 | 8.01 | 0.58 |

Table 4.3: Performance evaluation of different epoch extraction methods for speech associated with different types of voice disorders on SVD dataset. IDR–Identification rate, MR–Miss rate, FAR–False Alarm Rate, IDA–Identification Accuracy.

| Class | Method | IDR (%) | MR (%) | FAR (%) | IDA (ms) |
|---|---|---|---|---|---|
| Structural Disorders | ZP-ZFF | **87.84** | 5.72 | 6.44 | 0.42 |
| | ZFF | 87.79 | 5.37 | 6.84 | 0.52 |
| | DYPSA | 84.53 | 5.86 | 9.61 | **0.41** |
| | YAGA | 85.48 | **4.91** | 9.61 | 0.84 |
| | SEDREAMS | 87.51 | 6.24 | 6.25 | 0.47 |
| | SE-VQ | 74.63 | 17.97 | 7.40 | 1.02 |
| | CWT-GCI | 86.10 | 8.30 | 5.59 | 0.62 |
| | SPF | 83.57 | 10.9 | **5.54** | 0.66 |
| | GEFBA | 70.77 | 20.47 | 8.76 | 0.64 |
| Neurogenic Disorders | ZP-ZFF | **84.04** | 7.04 | 8.92 | **0.42** |
| | ZFF | 83.33 | 6.90 | 9.77 | 0.55 |
| | DYPSA | 81.40 | 7.32 | 11.28 | 0.47 |
| | YAGA | 81.76 | **6.22** | 12.03 | 0.71 |
| | SEDREAMS | 83.32 | 8.24 | 8.44 | 0.49 |
| | SE-VQ | 71.02 | 20.88 | 8.10 | 1.00 |
| | CWT-GCI | 80.83 | 11.90 | 7.27 | 0.65 |
| | SPF | 77.35 | 15.60 | **7.04** | 0.69 |
| | GEFBA | 61.06 | 30.89 | 8.05 | 0.62 |
| Functional Disorders | ZP-ZFF | **95.54** | 1.40 | 3.07 | **0.27** |
| | ZFF | 95.28 | **1.17** | 3.56 | 0.40 |
| | DYPSA | 93.80 | 2.16 | 4.03 | 0.28 |
| | YAGA | 93.10 | 1.40 | 5.50 | 0.62 |
| | SEDREAMS | 95.41 | 1.23 | 3.36 | 0.30 |
| | SE-VQ | 76.18 | 17.72 | 6.10 | 0.84 |
| | CWT-GCI | 88.97 | 8.59 | **2.44** | 0.49 |
| | SPF | 82.99 | 13.54 | 3.47 | 0.50 |
| | GEFBA | 65.92 | 26.16 | 7.92 | 0.55 |
| Psychogenic Disorders | ZP-ZFF | **94.34** | 2.00 | 3.66 | **0.27** |
| | ZFF | 93.77 | **1.67** | 4.56 | 0.37 |
| | DYPSA | 92.49 | 3.06 | 4.44 | 0.29 |
| | YAGA | 92.70 | 1.92 | 5.38 | 0.56 |
| | SEDREAMS | 93.81 | 2.00 | 4.18 | 0.28 |
| | SE-VQ | 74.36 | 19.76 | 5.88 | 0.77 |
| | CWT-GCI | 88.28 | 8.86 | **2.86** | 0.46 |
| | SPF | 82.93 | 13.85 | 3.21 | 0.51 |
| | GEFBA | 64.23 | 28.53 | 7.24 | 0.50 |

was very poor in terms of identification rate. Compared to the healthy scenario, for the structural, neurogenic, functional, and psychogenic voice disorder scenarios the epoch extraction algorithms showed an absolute reduction in IDR of approximately 10%, 15%, 3%, and 5%, respectively. The IDA refers to standard deviation of error, and therefore it should be lower for better performance of an epoch extraction method [56]. The IDA of the epoch extraction methods in neurogenic and structural voice disorder scenarios was increased approximately by 20 ms. More interestingly, the performance of epoch extraction methods degraded more for organic voice disorders (structural and neurogenic) than for non-organic voice disorders (functional and psychogenic). The results of this study indicate that existing epoch extraction methods need to be improved for accurate detection of epoch locations from the speech in the context of voice disorders.

## 4.2 Application of Region-wise approach for state-of-the-art epoch extraction algorithm

From the previous study, it was found that performance of the state-of-the-art epoch extraction methods, degrades for the voice disorder scenario. Some of the state-of-the-art epoch extraction methods depend on average value of pitch period for accurate estimation of GCI locations. These methods perform well for conditions in which variation of F0 is not very significant. For a healthy speaker, variation of F0 will not be significant in one single utterance. Hence, state-of-the-art epoch extraction algorithm perform well in these conditions. Voice disorders are associated with aperiodic and irregular vibration of vocal folds [11, 30], which in turn results in large variation of F0 compared to healthy speaker [24]. We have applied region-wise processing on the state-of-the-art epoch extraction methods, for extraction of GCI locations. According to this approach, GCI locations are computed for each region. Figure 4.2 shows the block diagram of region-wise epoch extraction approach. In this case F0 is extracted for each region, so that large variation of F0 will not affect performance of the overall method, specially in case of voice disorders.

Block diagram of region-wise approach used fo detection of GCI is shown in Figure 4.2. First speech activity regions are detected from speech signal. Then parameters like pitch period, fundamental frequency, maximum energy value, minimum energy value, or any other parameters, required for extracting the epochs from speech signal are computed from each region. Finally, this region, and parameters estimated in the region together are used for epoch extraction. All the regions are processed in the above mentioned approach to compute the epoch locations from the complete speech signal. The application of region-based approach is studied for the state-of-the-art epoch extraction methods.

### 4.2.1 Speech activity detection

For the speech activity or voiced activity detection, summation of residual harmonics (SRH) method [185] is used. For this, first residual signal (e(t)) is calculated using inverse filtering method. Then for each

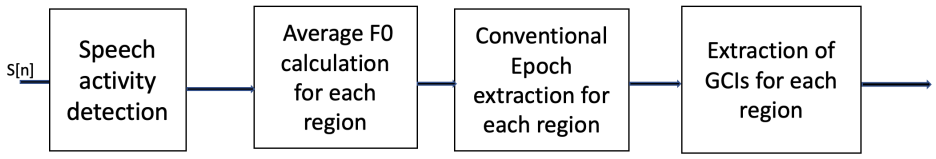



Figure 4.2: Region-wise approach for extraction of GCI location.

hanning windowed frame, the amplitude spectrum of residual signal $E(f)$ is obtained. For the voiced frame, $E(f)$ shows the peaks at the harmonics of the fundamental frequency (F0). From the spectrum of $E(f)$, for each $f$ the sum of residual harmonics is calculated.

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k.f) - E((k - \frac{1}{2}).f)] \tag{4.1}$$

$SRH(f)$ shows the maximum value at $F0$ for a frame. Using this method, a frame is marked as voiced if $SRH(f)$ is greater than the threshold value. In this way the voiced activity regions are determined for a given speech signal and parameters like pitch period, fundamental frequency, maximum energy value, and minimum energy value are calculated for each region.

### 4.2.2 Experimental results and discussion

Five state-of-the-art methods ZFF, ZP-ZFF, SEDREAMS, YAGA, and DYPSA are considered in our study to evaluate the performance. Moreover, the performance was also compared by applying the region-wise approach on the state-of-the-art epoch extraction methods in voice disorder scenario. The performance of each epoch extraction method is evaluated in terms of IR, MR, FAR, and IA is shown in Table 4.4. Figure 4.3 illustrates the epoch locations derived from the state-of-the-art epoch extraction methods on voice disorder scenario. Figure (a)-(g) shows speech signal, its corresponding ground truth signal (dEGG signal), epoch locations derived from ZFF, ZP-ZFF, SEDREAMS, DYPSA and YAGA methods before applying the region-wise approach. Figure (h)-(n) shows speech signal, its corresponding ground truth signal (dEGG signal), epoch locations derived from ZFF, ZP-ZFF, SEDREAMS, DYPSA, and YAGA methods after applying the region-wise approach. It can be observed from the Figure 4.3 that performance of ZFF, ZP-ZFF, and SEDREAMS methods improved in terms of FAR (FAR is reduced) after applying the region-wise approach, while the performance remains same for dynamic-based programming based methods.

Table 4.4 shows the performance of state-of-the-art methods without and with applying the region based process for different categories of voice disorders. Column 1 in the Table 4.4 indicates different

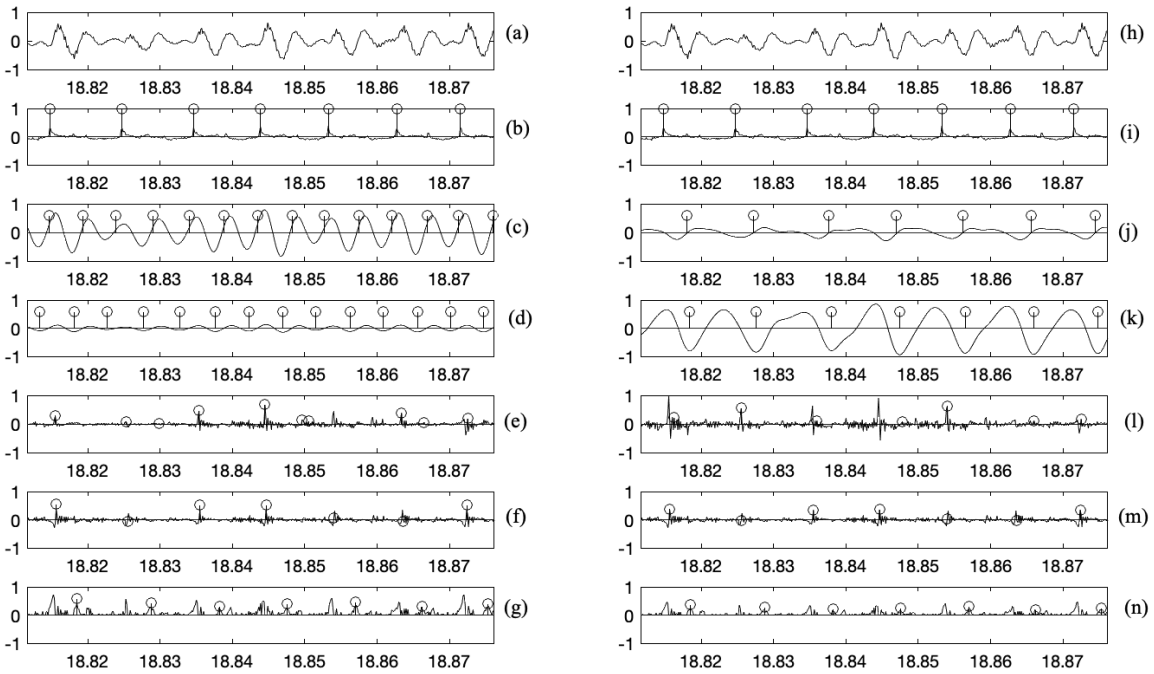



Figure 4.3: Epoch extraction from Voice disorder scenario using the state-of-the-art epoch extraction methods before and after applying the region-wise approach (a),(h) Speech segment for speech utterance associated with voice disorder. (b),(i) Differenced EGG signal. (d),(j) ZFF signal with identified GCI locations without and with region-wise approach, respectively. (e),(k) ZP-ZFF signal with identified GCI locations without and with region-wise approach, respectively. (f),(l) LP residual signal from SEDREAMS Method with identified GCI locations without and with region-wise approach, respectively. (g),(m) LP residual signal with Identified GCI location using DYPSA method without and with region-wise approach, respectively. (h),(n) LP residual signal with Identified GCI location from YAGA method without and with region-wise approach, respectively.

categories of the voice disorder. Epoch extraction methods are indicated in column 2. Column 3 to 6 indicate the performance of state-of-the-art epoch extraction methods and column 7 to 10 indicate the performance of epoch extraction methods after applying the region-wise approach. The general observation from the results is that performance of state-of-the epoch extraction methods like ZFF, ZP-ZFF, and SEDREAMS methods is degraded for all the categories pf voice disorder. It may be due to degradation in voice quality for voice disorder (presence of breathiness, creakiness, and harshness). Degradation in the performance may be due to the reason that these methods depend on average value of pitch period for extraction of epoch locations. The performance is improved in terms of FAR, for ZFF, ZP-ZFF, and SEDREAMS methods by processing the methods in region-wise approach. As in the region-wise approach, parameters used for identification of epoch locations are extracted for each region, hence variation in pitch period does not affect the performance in the for these methods. On the other hand dynamic programming based methods like DYPSA and YAGA, are robust for different categories of voice disorder. The performance of both the methods, after applying the region-wise approach is similar. It may be due to dynamic programming algorithm used for extraction of epoch locations.

Table 4.4: Performance evaluation of different epoch extraction methods for the different categories of voice disorder scenario. IR–Identification rate, MR–Miss rate, FAR–False Alarm Rate, IA–Identification Accuracy.

| Category | Epoch extraction method | Without region | | | | With region | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | IR(%) | MR(%) | FAR(%) | IA(ms) | IR(%) | MR(%) | FAR(%) | IA(ms) |
| Structural VD | ZFF | 92.99 | 2.16 | 4.85 | 0.46 | 94.14 | 2.64 | 3.21 | 0.51 |
| | ZP-ZFF | 91.45 | 2.38 | 6.17 | 0.34 | 93.10 | 2.55 | 4.35 | 0.36 |
| | SEDREAM | 88.89 | 4.04 | 7.07 | 0.39 | 91.18 | 3.10 | 5.72 | 0.41 |
| | YAGA | 91.84 | 2.23 | 5.93 | 0.83 | 91.32 | 2.30 | 6.37 | 0.86 |
| | DYPSA | 89.25 | 4.30 | 6.45 | 0.40 | 89.32 | 4.33 | 6.36 | 0.40 |
| Neurological VD | ZFF | 90.39 | 1.71 | 7.90 | 0.36 | 91.32 | 3.25 | 5.42 | 0.42 |
| | ZP-ZFF | 91.24 | 1.65 | 7.10 | 0.27 | 92.26 | 2.88 | 4.86 | 0.28 |
| | SEDREAM | 91.42 | 1.91 | 6.68 | 0.34 | 90.13 | 3.19 | 6.68 | 0.30 |
| | YAGA | 90.65 | 2.34 | 7.01 | 0.54 | 88.71 | 2.82 | 8.47 | 0.60 |
| | DYPSA | 89.03 | 4.26 | 6.71 | 0.33 | 87.98 | 4.71 | 7.32 | 0.35 |
| Functional VD | ZFF | 91.20 | 3.13 | 5.67 | 0.47 | 92.09 | 3.31 | 4.60 | 0.48 |
| | ZP-ZFF | 92.67 | 2.21 | 5.13 | 0.31 | 92.88 | 2.70 | 4.42 | 0.32 |
| | SEDREAM | 92.38 | 1.92 | 5.70 | 0.33 | 92.44 | 2.16 | 5.40 | 0.42 |
| | YAGA | 91.07 | 1.83 | 7.10 | 0.59 | 90.47 | 1.75 | 7.78 | 0.59 |
| | DYPSA | 91.43 | 3.75 | 4.82 | 0.34 | 91.08 | 3.78 | 5.14 | 0.34 |
| Psychogenic VD | ZFF | 88.98 | 1.02 | 10.00 | 0.32 | 90.28 | 1.94 | 7.78 | 0.37 |
| | ZP-ZFF | 90.13 | 0.98 | 8.89 | 0.27 | 92.05 | 1.39 | 6.56 | 0.28 |
| | SEDREAM | 94.14 | 1.41 | 4.45 | 0.33 | 92.16 | 2.28 | 5.56 | 0.37 |
| | YAGA | 91.33 | 2.07 | 6.60 | 0.52 | 91.43 | 1.90 | 6.67 | 0.54 |
| | DYPSA | 92.01 | 3.37 | 4.62 | 0.33 | 91.74 | 3.54 | 4.72 | 0.31 |

From the results, it can be observed that performance is improved in terms of IR (approximately 2%), for ZFF and ZP-ZFF methods after applying the region-wise approach, for all the categories of voice disorders. For structural voice disorder, region-wise ZFF method showed the best performance in terms of IR of 94.14% and FAR of 3.32%. For neurogenic and functional voice disorder, ZP-ZFF showed the

best performance in terms of IR of approximately 92% compared to all other epoch extraction methods. SEDREAMS method showed the best performance in terms of IR of 94.14%, MR of 1.41%, and FAR of 4.45%.

## 4.3 Extraction of excitation source based features from the region-based approach for voice disorder detection and identification

From the previous studies, it can be concluded that if GCI locations are detected with the application of region-based processing on state-of-the-art epoch extraction algorithm then performance is improved for voice disorder scenario. Hence, if the excitation source features are extracted from this method, it might improve the performance of voice disorder detection and identification system. This section compares the performance of excitation source based features with and without applying the region-wise process on the state-of-the-art methods.

### 4.3.1 Experiment setup

The experiments have been performed on SVD database for both detection and identification task. Six features are used in this regard to carry out experiments as in line with our previous study.

- MFCC-Residual-WR, and MFCC-ZFF-WR: To derive these features first residual signal and epoch locations are derived from each region for complete speech signal. MFCC features are derived from a frame-length of 20 ms and a frame-shift of 5 ms. These features consist of 39 dimensions, comprising 13 static coefficients, as well as their first and second-order derivatives. Additionally, four statistical measures—mean, standard deviation, kurtosis, and skewness—are calculated from the MFCC coefficients. This results in 156 dimension feature vector referred as MFCC-Residual-WR and MFCC-ZFF-WR, respectively.

- Intonation-WR: Intonation features are derived by applying the region-based processing to ZFF method. 76 dimension feature vector is derived which consist of perturbation parameters as discussed in Chapter 3.

- MFCC-Residual, MFCC-ZFF, and intonation: These features are considered as baseline feature for this experiment. MFCC features are derived from LP residual signal and ZFF signal as discussed in the previous chapter.

SVM classifier shows the consistent performance for small database used in the pathological application. Therefore SVM classifier with polynomial kernel of order 2 is used to train and test the voice disorder detection and identification system in this chapter.

### 4.3.2 Results and discussion

The primary aim of this study is to improve the performance of detection and identification systems by utilizing features obtained from epoch locations resulting from region-based processing. The state-of-the-art epoch extraction algorithms, namely LP residual and ZFF, are employed for this purpose. The voice disorder detection and identification system is developed using MFCC features extracted from these algorithms and classified using an SVM classifier. MFCC-Residual and MFCC-ZFF features are considered as baseline features to compare the performance. Hierarchical approach was followed in order to perform the detection and identification experiments in a clinical way. Total of four experiments were performed to know category of voice disorder and results are shown in the Table 4.5.

Table 4.5: Performance of voice disorder detection and identification systems in terms of classification accuracy (in %) for individual feature set on SVD database. Here, Exp. 1: classification of healthy and voice disorder, Exp. 2: classification of organic and non-organic voice disorders, Exp. 3: classification of structural and neurogenic voice disorders, and Exp. 4: classification of functional and psychogenic voice disorders.

| Features | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 |
|---|---|---|---|---|
| MFCC-Residual | 71.7 | 69.3 | 61.7 | 64.5 |
| MFCC-Residual-WR | 72.7 | 69 | 63.6 | **69.4** |
| MFCC-ZFF | 69.9 | 69.4 | 63.6 | 65.3 |
| MFCC-ZFF-WR | **74.4** | **72.7** | **67.1** | 65 |
| Intonation | 64.8 | 63.8 | 60.4 | 55.9 |
| Intonation-WR | 67.8 | 61.5 | 67 | 56.5 |

Table 4.5 indicate that the performance of features extracted with region-based processing is improved as compared to the baseline features. It can be observed from the Table 4.5 that for the features extracted after applying the region-based processing, the performance of voice disorder detection and identification task is improved in almost all cases. MFCC-Residual-WR feature exhibits improved performance compared to the MFCC-Residual feature in experiments 1, 3, and 4 upto 1%, 2% and 5% respectively. In experiment 2, the MFCC-Residual-WR feature achieves a comparable classification accuracy of 69.3%. The performance for the voice disorder detection task is improved by 5% for MFCC-ZFF-WR feature as compared to MFCC-ZFF. Classification accuracy increases from 63.6% to 72.7% for experiment 2 by using the MFCC-ZFF-WR feature. Similar improvements are observed for intonation features obtained by applying region-based pre-processing to the ZFF method.

## 4.4 Conclusion

In this chapter, we conducted a comparative analysis of epoch extraction methods to evaluate their performance in both healthy and voice disorder scenarios. Our study revealed that most of these methods exhibited better performance in healthy scenarios compared to voice disorder scenarios. The reason for

this performance degradation could be attributed to the higher variation in fundamental frequency ($F0$) observed in subjects with voice disorders.

To address this issue, we explored the approach of calculating epoch locations region-wise, which resulted in an improvement in the performance of the state-of-the-art epoch extraction algorithm. Subsequently, we utilized the excitation source features derived from this algorithm, both with and without the region-based processing, for voice disorder detection and identification.

The results showed that incorporating the region-based processing approach led to enhanced performance across all experiments when compared to the baseline features. Therefore, we can conclude that accurately identifying epoch locations can significantly improve the performance of automatic voice disorder detection and identification systems utilizing features derived from the excitation source signal.

*Chapter 5*

# Detection and identification of voice disorders using the features derived from long-term average spectrum

Voice disorders are characterized by abnormal voice production, change in voice quality, pitch, and loudness inappropriate to age and gender [11]. Perceptually, these voice disorders are often associated with symptoms such as roughness, breathiness, strain, and harshness in the voice. These voice qualities from the speech signal are perceived in the long term [66]. Hence these features can be captured by Long Term Average Spectrum (LTAS). LTAS captures the static characteristic of the speaker's voice instead of the short time variation present in the speech. Many researchers used LTAS in clinical application, as well as in quantification of voice quality. Some studies claim that LTAS can be used for voice classification [127]. Some researchers used LTAS as a good acoustic measure to differentiate the male and female speakers [67]. In [128], LTAS is also used to study voice quality changes before and after surgery. Other works related to LTAS were finding differences related to age [68], professional singers, different styles of singing [129], speaking and singing [130], and quantifying the quality of voice [131].

For extraction of the LTAS features, speech signal should be decomposed into multiple frequency components using filter banks. In the literature, LTAS features were extracted using the critical band filter bank [67, 68]. Recently, the author in [69] explored the single frequency filter bank for hypernasality detection. SLPs make decisions regarding the presence of voice disorders by carefully listening to the subject's entire utterance. To replicate the human basilar membrane, auditory filter banks are commonly used in the literature. The bandwidth of the auditory filter is designed such that it is narrow for lower frequencies and wider for higher frequencies. We hypothesized that auditory filters might better capture perceptual characteristics related to voice disorders compared to other filter banks. This motivated us to explore various auditory filter banks, such as constant-Q and gammatone filter banks, for automatic detection and identification of voice disorders. The performance of the voice disorder detection and identification system is then compared with other filter banks that have been previously used in the literature.

Rest of the chapter is organised as follows. In Section 5.1, filter banks used for LTAS feature extractions are discussed. The experimental setup which describes feature extraction, database and classifier is

discussed in the Section 5.2. Results obtained are presented in the Section 5.3. Conclusion and summary of this study are described in Section 5.4.

## 5.1 Filter banks for LTAS feature extraction

For the extraction of LTAS features the speech signal should be decomposed into multiple frequency bands using the filter bank. Filter bank is set of band pass filters which passes the selected range of frequencies of the signal, while attenuates the other frequencies. Auditory filter banks like gammatone and constant-Q are used in this study in order to effectively capture voice quality-related information in individuals with voice disorders. This section describes state of the art filter banks used in this study for voice disorder detection and identification in a clinical way, along with the extraction of the LTAS features.

### 5.1.1 State of the art filter banks

The filter banks considered in this study, namely critical band, gammatone, Constant-Q, and single frequency filter banks, are described as follows.

#### 5.1.1.1 Critical band filter bank

Critical band filter bank (CBFB), also referred to as octave band filter bank, is used to mimic human perception. Octave band filters are set of bandpass filters in which highest frequency is twice of the lowest frequency [67]. Octave band is mainly used in music, in which one octave is difference between same notes with double its frequency. Critical band filter is Butterworth band pass filter with center frequency of 30, 60, 120, 240, 480, 960, 1920, 3840, and 7680 Hz designed for the signal with sampling frequency $f_s$ of 8 kHz. Frequency-domain response of critical band filter is shown in Figure 5.1.

#### 5.1.1.2 Gammatone filter bank

The gammatone filters are the most widely used auditory filters to model the human auditory system. In the term gammatone, gamma is referred to function mostly used in probability, and tone refers to the cosine term. Gammatone filter bank (GFB) models the cochlea by overlapping bandpass filter with impulse response given by the product of a rising polynomial, a decaying exponential function, and a cosine wave [186]. Figure 5.2 and 5.3 shows the time and frequency domain response of gammatone filter. The impulse response of a gammatone filter $g(t)$ is given by,

$$g(t) = at^{(N-1)}e^{-2\pi bt}cos(2\pi f_c t + \phi) \qquad for\ \ t \geq 0. \tag{5.1}$$

Here, $N$ is the order of the filter which determines the slope of the filter's skirts, $b$ is the bandwidth of the filter, $f_c$ is center frequency, $a$ and $\phi$ are the scaling factor and phase of the cosine wave, respectively.

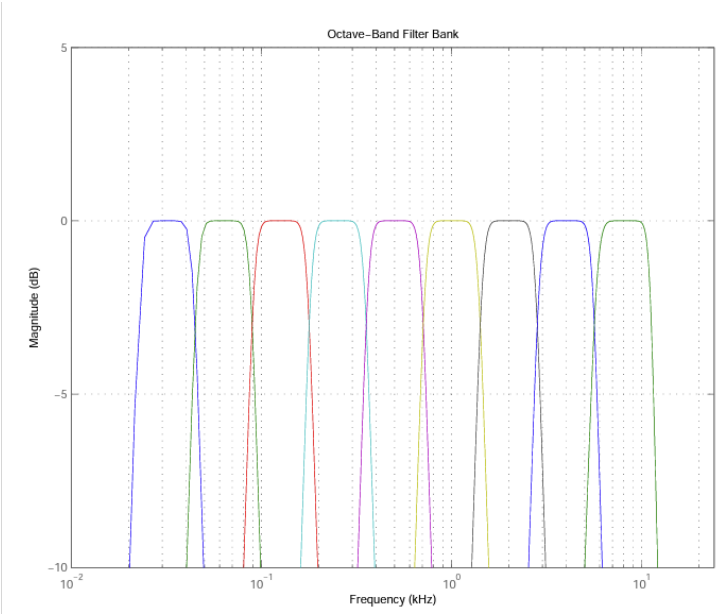



Figure 5.1: Frequency response of Critical band filter bank[67].

In general, the order of the gammatone filter is chosen in-between 3 to 5, to model the human auditory system [187]. The bandwidth $b$ correspond to each $f_c$, is obtained using the Equivalent Rectangular Bandwidth (ERB) scale which is given by [188],

$$b = ERB(f_c) = 24.7(4.37f_c + 1) \tag{5.2}$$

where, $b$ is in Hz and $f_c$ is in kHz.

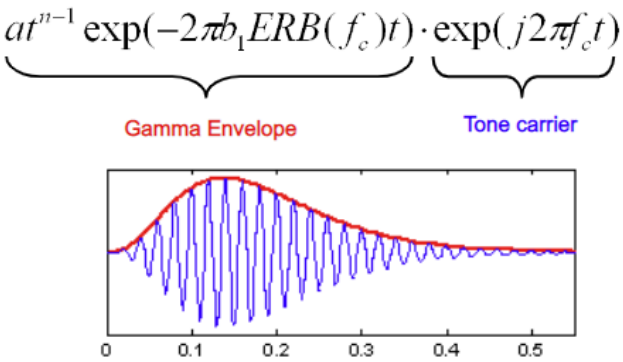



Figure 5.2: Time domain response of Gammatone function [188].

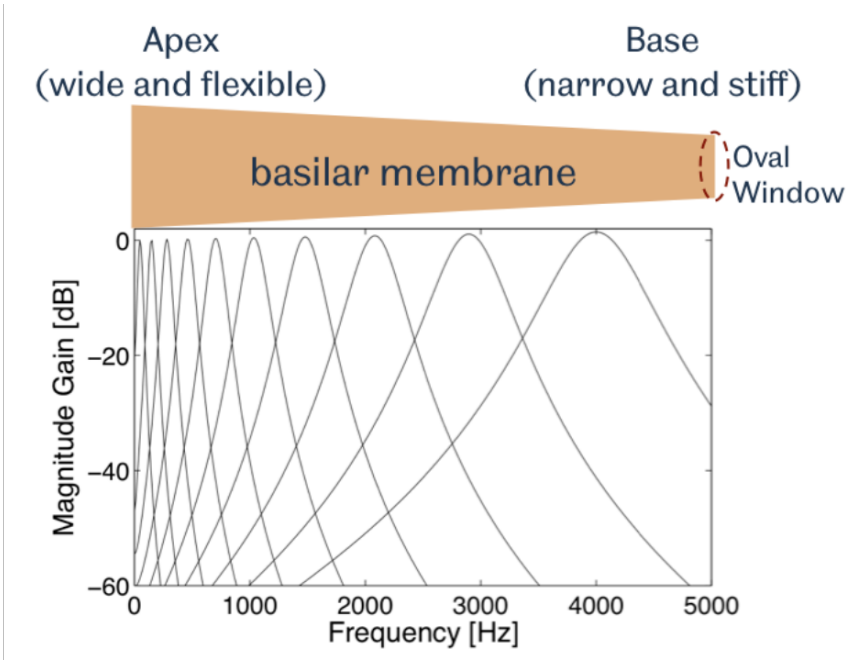



Figure 5.3: Frequency domain response of Gammatone function[188].

#### 5.1.1.3 Constant-Q filter bank

The Constant-Q filter bank (CQFB) is based on the Constant-Q transform (CQT) and utilizes a filter bank with geometrically spaced filters. These filters maintain a constant-Q factor, meaning that the ratio of the center frequency to the resolution remains constant. This unique design allows the resolution of the filters to approximate musical notes. [189]. CQT provides variable time and frequency resolution. For the discrete time signal $x[n]$ the CQT is given by

$$X[k,n] = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x[j] a_k^*(j - n + N_k/2) \tag{5.3}$$

where $k$ is frequency bin index, $N_k$ is the window length, and $a_k$ is the complex time frequency atoms which is defined as

$$a_k(n) = \frac{1}{C}\frac{n}{N_k} exp[i(2\pi \frac{f_k}{f_s} + \phi_k)] \tag{5.4}$$

$f_k$ is the center frequency of the $k$th bin, $f_s$ is the sampling frequency and $\phi_k$ is the phase offset and C is the scaling factor which is given by

$$C = \sum_{l=-\lfloor N_k/2 \rfloor}^{\lfloor N_k/2 \rfloor} w\left(\frac{l + N_k/2}{N_k}\right) \tag{5.5}$$

In order to maintain constant-Q factor, length of the window is defined as

$$N_k = \frac{f_s}{f_k} Q \tag{5.6}$$

The $k^{th}$ center frequency of constant-Q transform is given by

$$f_k = f_0 \, 2^{k/B} \tag{5.7}$$

where, $f_0$ is minimum frequency, and $B$ is number of bins per octave that determines trade-off of time-frequency resolution provided by the filter. The bandwidth of the filter $b$ is given by

$$b = f_{k+1} - f_k = f_k \, (2^{1/B} - 1). \tag{5.8}$$

Quality factor (Q-factor) is given by

$$Q = \frac{f_k}{b} = (2^{1/B} - 1)^{-1} \tag{5.9}$$

This constant-Q factor leads to high temporal resolution at high frequency and high frequency resolution at low frequency.

#### 5.1.1.4 Single frequency filter bank

The single frequency filter bank (SFFB) (as discussed in [190]), is based on single frequency filtering (SFF) which is time-frequency analysis method [191]. SFF provides amplitude envelope of the speech signal at each selected frequency as a function of time.

1. Speech signal $s[n]$ is passed through pre-emphasis filter

$$x[n] = s[n] - s[n-1] \tag{5.10}$$

2. The signal $x[n]$ is frequency shifted by multiplying it with complex exponential

$$\tilde{x}_k[n] = x[n] e^{-j\frac{2\pi \tilde{f}_k n}{fs}} \tag{5.11}$$

   $\tilde{f}_k$ is normalized frequency and is given by

$$\tilde{f}_k = \frac{fs}{2} - f_k \tag{5.12}$$

   where $f_k$ is the $k^{th}$ desired frequency and $f_s$ is the sampling frequency

3. The frequency shifted signal is passed through a single pole filter, whose pole is located on to the negative real axis (at $z = -r$).

$$H(z) = \frac{1}{1 + rz^{-1}} \tag{5.13}$$

4. The output of the filter is given by

$$y_k[n] = -r y_k[n-1] + x_k[n] \tag{5.14}$$

5. The amplitude envelope of signal $y_k[n]$ is given by

$$e_k[n] = \sqrt{(y_{kr}^2[n] + y_{ki}^2[n])} \tag{5.15}$$

where $y_{kr}[n]$ and $y_{kr}[n]$ are the real and imaginary components of $y_k[n]$, respectively.

The value of $r$ which can be selected in between 0 to 1, determines the bandwidth of the filter. The narrow filters are designed to provide high spectral resolution by choosing the value of $r$ between 0.95 to 0.995. Figure 5.4 represents the frequency-domain response of SFF.

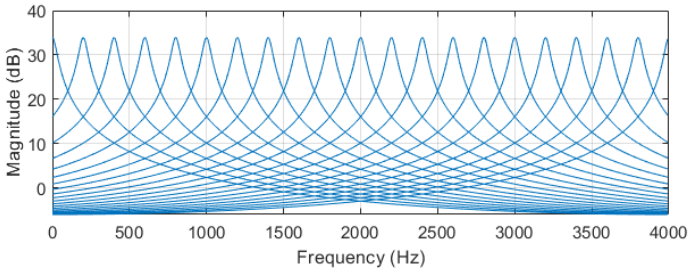



Figure 5.4: Frequency domain response of single frequency filter bank [190].

### 5.1.2 Extraction of Long term average spectral features

The long term average spectrum features capture the static information like voice quality, gender information and age-related features from the speech signal [67]. To extract these features, first, the speech signal $s[n]$ is passed through the bank of filter to decompose it into multiple time-frequency components (as shown in Figure 5.5). If $h_i[n]$ is filter's impulse response then the output of the filter is given by

$$s_i[n] = h_i[n] * s[n] \quad i = 1, 2.....N \tag{5.16}$$

where $N$ is the number of filters. All the $N$ band signals along with original full-band signal in total $N + 1$ components are framed using a non-overlapping rectangular window of 20 ms. Then root mean square energy is calculated for each frame denoted by $s_{RMSi[k]}$ correspond to the $k^{th}$ frame of $i^{th}$ band. Finally, 10 statistical averages like normalized mean, standard deviation, range, skewness and kurtosis are calculated, the resulting $((N + 1) * 10 - 1)$ dimension feature vector is denoted as LTAS feature.

## 5.2 Experimental setup

This section describes the method to extract the various features used for studying voice disorder detection and identification. Further details of the database, baseline features, and classifier used for this study are presented in the following section.

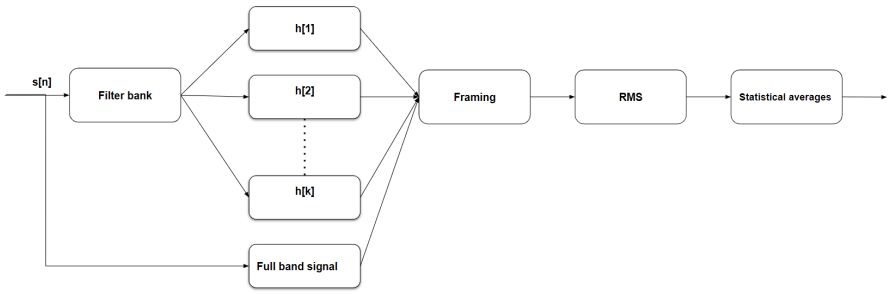



Figure 5.5: LTAS feature extraction [67].

### 5.2.1 Feature Extraction

The features explored in this study include the LTAS features obtained by using the state of the art filter banks, statistical averages of the short time features (LPCC, MFCC, PLP, etc.) and state of the art openSMILE features such as eGEMAPS and ComParE. The extraction of these features is presented as follows.

#### 5.2.1.1 LTAS based features

The parameters of each filter bank considered for extracting the LTAS features are described in the following subsection.

- CBFB-LTAS feature is calculated using 9-octave band signals and one full band speech signal. To get the time-frequency decomposition of the speech signal, first, the signal is passed through 9-octave band filters with the minimum centre frequency of 30 Hz and a maximum frequency of 7680 Hz. Finally, 99 (10*10-1) dimension CBFB-LTAS vector is obtained.

- For extraction of CQFB-LTAS feature vector, the speech signal is passed through the CQFB with 106 constant-Q spaced filters. The CQFB is realized using $f_{min}$ of $10Hz$, $f_{max}$ of $4000Hz$, and number of bins per octave $b$ of 12 [192]. In total, 107 components are used, resulting in 1069 (107*10-1) dimension LTAS feature vector.

- In case of GFB-LTAS feature extraction, the speech signal is decomposed by passing it through the 32 gammatone-tone filters [193]. The minimum and maximum frequency are selected as 0 Hz and 4000 Hz, respectively, which results in 329 (33*10-1) dimension feature vector.

- To extract the SFFB-LTAS feature vector, the speech signal is passed through 201 SFF. The pole location $r$ of 0.98 and frequency resolution of 20 Hz were used to realize the SFFB (as in [69]). Total of 202 components (201 filter responses and speech signal) are used, results in 2019 (202*10-1) dimension LTAS feature vector.

#### 5.2.1.2 Statistical averages of the state of the art features

To compute the statistical averages, first, frame-level features were computed using a Hamming window of size 25 ms with 10 ms frame shift. First $m$ static cepstral coefficients and their delta, and delta-delta features were computed yielding in $d = 3 * m$ dimension feature vector. Finally, statistical averages such as mean, standard deviation, kurtosis and skewness were derived from these frame-level features resulting in $D = (d * 4)$ dimension feature vector named as STAT features as in [69]. Conventional MFCC, LPCC, PLP, and CQCC features, which captures the vocal tract information are used to compute corresponding STAT features, namely MFCC-STAT, LPCC-STAT, PLP-STAT, and CQCC-STAT. CQCC features were calculated from the CQT-transform with $f_{min}$ of 100 Hz, $f_{max}$ of 4000 Hz and bins per octave of 192 [189].

Along with the system features, we also explored the excitation source evidence such as LP-residual and zero frequency filtered (ZFF) signal to compute the STAT features. In this regard, MFCC features were computed from LP-residual and ZFF-signal as in [40]. Then corresponding STAT features were computed and are named as MFCC-Residual-STAT and MFCC-ZFF-STAT, respectively. MATLAB implementation of the features used along with supporting material is provided in `https://github.com/Purva-Barche/LTASfilterbankcodes`.

#### 5.2.1.3 OpenSMILE features

This work explored two state of the art feature sets obtained from openSMILE tool kit [164] as baseline features. The first feature set is extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) which is low dimension knowledge-based acoustic feature [166]. It is 88 dimension feature set mainly used for extraction of emotion. The second set used is Computational Paralinguistic Challenge (ComParE) feature set which is brute-forced set [194]. It has a dimension of 6373 feature which are usually designed to extract paralinguistic information from the acoustic signal.

### 5.2.2 Database

Databases used in this chapter are Saarbruecken voice disorder (SVD) dataset [150], and Hospital Universitario Prıncipe de Asturias (HUPA) database [151].

- The SVD dataset is used for performing the experiments. In this study, the speech samples corresponding to voice disorders from SVD database were grouped into four classes as used in our previous chapter [41], namely, *Structural, Neurogenic, Functional and Psychogenic*. In this regard 625 samples were considered from healthy class and total of 950 voice samples were considered from different voice disorders category for vowel /a/, /i/, and /u/ in normal, high, low and rising-falling pitch.

- The HUPA database contains recordings of the vowel /a/ for a total of 440 subjects. Out of total 366 recordings, 239 recordings are from pathological subjects, and 201 recordings are from nor-

mal subjects. It contains organic pathologies like Bilateral Reinke's edema, Polyp, Cyst, Bilateral nodule, Recurrent nerve paralysis etc. Auditory-perceptual ratings according to GRBAS scale is available for HUPA database. It contains the five different components, Grade of hoarseness (G), Roughness (R), Breathiness (B), Asthenia (A), and Strain (S). Each component is rated as 0, 1, 2, or 3, where 0 indicates normal, 1 mild, 2 moderate and 3 indicates more severe degree of voice disorder.

### 5.2.3 Classifier

The classifier used in our study for detection and identification of voice disorders is the support vector machine (SVM) which is a supervised binary classifier. The detection and identification of voice disorders were also done by using several other classifiers like decision tree, k-nearest neighbour, ensemble classifier and logistic regression. SVM is selected among all other classifiers due to its best classification accuracy. Among all different kernels like linear, radial basis functions, and polynomial, polynomial kernel with a polynomial degree of 2 outperformed in this study. Moreover, the grid search algorithm was performed to select the optimum value of kernel parameters. Further, five-fold cross-validation was performed to find the classification accuracy.

## 5.3 Results and discussion

In the previous study [41], we have performed identification of voice disorder in clinical way by using excitation source evidences. Among the individual excitation source features the intonation features derived from ZFF signal and MFCC-Residual provided best classification accuracy of 69.3% and 70.8% for detection and identification task, respectively. In continuation to our previous studies, the present work explored the significance of long term average spectral features (supra-segmental features) using state of the art filter banks for voice disorder detection and identification tasks in the similar way to improve the performance of both the tasks. Also, the performance of the detection and identification system is compared with state of the art openSMILE features and statistical averages of frame-level features. The detection system performs a binary classification to discriminate the speech samples corresponding to healthy and voice disorders. On the other hand, identification is a multi-level classification problem in which three binary classifiers were used to identify the type of voice disorder. Total of four experiments were carried out in our thesis. Further, the relation between the LTAS features and perceptual scale was evaluated using N-way analysis of variances (ANOVA).

### 5.3.1 Performance analysis of voice disorder detection and identification system

Voice disorder detection and identification experiments were performed on the SVD dataset, whereas only detection task was performed on HUPA dataset as samples of different categories of voice disorders are not available for HUPA database. All the experiments were performed using the SVM classifier.

Performance of the detection and identification systems with individual baseline features and LTAS features obtained from various filter banks is reported in Table 5.1 for SVD database. Table 5.2 shows the voice disorder detection (Exp. 1) result for HUPA database. In addition, the performance of detection and identification systems was evaluated using the combination of filter bank features with the state of the art openSMILE features, and the results are presented on SVD and HUPA database in Table 5.3.

Table 5.1: Performance of voice disorder detection and identification systems in terms of classification accuracy (in %) for individual feature set on SVD database. Here, Exp. 1: classification of healthy and voice disorder, Exp. 2: classification of organic and non-organic voice disorders, Exp. 3: classification of structural and neurogenic voice disorders, Exp. 4: classification of functional and psychogenic voice disorders, S1 Statistical average feature set, S2 openSMILE feature set, S3 LTAS features.

| | Feature | Exp.1 | Exp.2 | Exp.3 | Exp.4 |
|---|---|---|---|---|---|
| S1 | MFCC-STAT | 76.1 | 71.6 | 69.9 | 68.2 |
| | PLP-STAT | 78.4 | 71.2 | 74.7 | 66.2 |
| | LPCC-STAT | 75.6 | 68.6 | 70.4 | 65.3 |
| | CQCC-STAT | 74.4 | 70.3 | 71.2 | **70.8** |
| | MFCC-Residual-STAT | 72 | 70.1 | 66.3 | 65.9 |
| | MFCC-ZFF-STAT | 71.3 | 69.3 | 70.6 | 69.1 |
| S2 | eGeMAPS | 80.7 | 71 | 70.6 | 64.5 |
| | ComParE | **85.9** | **75.7** | **76.5** | 69.4 |
| S3 | CBFB-LTAS | 74.3 | 69.9 | 68.6 | 66.2 |
| | GFB-LTAS | 76.9 | 71.4 | 69.9 | 67.9 |
| | CQFB-LTAS | 78 | 70.8 | 71.2 | 65.9 |
| | SFFB-LTAS | 76.8 | 69 | 69.1 | 65.9 |

From Table 5.1, it is evident that, among all STAT features PLP-STAT features shows better classification accuracy of 78.4% and 74.7% for Exp. 1 and 3 respectively . Further, ComParE feature set outperformed for all the experiments. Among all LTAS features, CQFB-LTAS performed better for Exp. 1 and 3, while GFB-LTAS performed better for Exp. 2 and 4. Moreover, the performance of the CQFB-LTAS features (78%, 70.8% and 71.2%) is comparable to the baseline eGeMAPS features (80.7%, 71% and 70.6%) for three experiments.

Table 5.2 shows the voice disorder detection (only Exp. 1) results on HUPA dataset using the different baseline features and LTAS based features. From the table it is evident among all the STAT features PLP-STAT features shows better classification accuracy of 73.7% for HUPA datset. Further, the best performance is obtained in term of classification accuracy of 82.1% for ComParE feature sets. Among the filter bank based LTAS features, CQFB-LTAS performed best with a classification accuracy of 81.4%.

Among baseline feature sets, the openSMILE features showed better classification accuracy compared to statistical feature sets; hence, the performance was also observed by combining the LTAS feature sets with openSMILE feature sets as reported in Table 5.3 for SVD (all the experiments) and HUPA (only Exp. 1) database. It can be observed from the Table 5.3 for the detection task best classification

Table 5.2: Performance of voice disorder detection systems in terms of classification accuracy ( in %) for HUPA database. Here, S1 Statistical average feature set, S2 openSMILE feature set, S3 LTAS features.

| | Features | Accuracy (%) |
|---|---|---|
| S1 | MFCC-STAT | 69.2 |
| | LPCC-STAT | 69.2 |
| | PLP-STAT | 73.7 |
| | CQCC-STAT | 62.3 |
| | MFCC-Residual-STAT | 70.2 |
| | MFCC-ZFF-STAT | 69.9 |
| S2 | eGeMAPS | 76.1 |
| | ComParE | **82.1** |
| S3 | CBFB-LTAS | 75.9 |
| | CQFB-LTAS | 81.4 |
| | GFB-LTAS | 79.2 |
| | SFFB-LTAS | 74.9 |

accuracy of 89.6% is obtained when CBFB-LTAS features combined with eGeMAPS features for SVD database. For HUPA database the best classification accuracy of 86.6% is observed when constant-Q based LTAS features were combined with ComParE feature sets. SFFB-LTAS features when combined with ComParE performed best among all other (for SVD samples) combinations for Exp. 2 and 3. It can also be observed that even by combining the different features, classification accuracy for the Exp. 4 is not increased significantly, as psychogenic voice disorder samples mostly confused with functional voice disorder.

Table 5.3: Performance of voice disorder detection and identification systems in terms of classification accuracy (in %) for combination of feature sets on SVD and HUPA database. Here, Exp. 1: classification of healthy and voice disorder, Exp. 2: classification of organic and non-organic voice disorders, Exp. 3: classification of structural and neurogenic voice disorders, and Exp. 4: classification of functional and psychogenic voice disorders.

| Features | SVD | | | | HUPA |
|---|---|---|---|---|---|
| | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 | Exp. 1 |
| CBFB-LTAS+eGeMAPS | **89.6** | 71.9 | 72.9 | 69.1 | 80 |
| CBFB-LTAS+ComParE | 86 | 76.1 | 77.2 | 67.1 | 85.4 |
| GFB-LTAS+eGeMAPS | 87.5 | 73 | 70.2 | 64.5 | 82.1 |
| GFB-LTAS+ComParE | 85.8 | 77.2 | 77 | **69.7** | 81.1 |
| CQFB-LTAS+eGeMAPS | 84.2 | 72.9 | 69.9 | 67.6 | 83 |
| CQFB-LTAS+ComParE | 87.2 | 78.3 | 75 | 67.9 | **86.6** |
| SFFB-LTAS+eGeMAPS | 84.1 | 68.7 | 69.9 | 64.5 | 78.2 |
| SFFB-LTAS+ComParE | 86.9 | **78.9** | **77.4** | 68.5 | 81.3 |

### 5.3.2 ANOVA analysis

To assess the relationship with the perceptual scale used by SLPs, statistical analyses were computed with N-way analysis of variance (N-way ANOVA). The ANOVA test determines whether or not any statistically significant difference exist between means of two or more groups by measuring the probability value (p-value). The p-value in the ANOVA test used to decide whether null hypothesis should be accepted or rejected. If p-value is very smaller than 0.05, it signifies that there is a significant difference among the means of the groups.

This analysis was performed by considering the LTAS feature as a dependent variable and perceptual ratings of Grade of hoarseness, roughness, breathiness, asthenia and strain as independent variables. ANOVA was computed on the HUPA dataset which has a perceptual rating according to the GRBAS scale. Out of 99 LTAS features, 35 features show significant interaction with the perceptual scale of roughness, while 31 features indicate significant interaction with asthenia (p ¡ 0.05). Remaining 14 features out of 99 LTAS features indicate the least value of p (much smaller than 0.05) for overall degree of hoarseness, while 11 LTAS features and 8 LTAS features shows the minimum value of p for perceptual scale of breathiness and strain, respectively. Moreover, N-way ANOVA was also obtained for different frequency ranges. Two frequency ranges were considered, one from 0 to 1 KHz and other above 1 KHz. It was observed that for the frequency range below 1 KHz, 31 and 27 LTAS features out of 69 features indicate the minimum value of p for perceptual scale R (Roughness) and Asthenia respectively. For the frequency range above 1 KHz perceptual rating, G(Overall severity) and S (Strain) indicate the minimum value of p for most of the LTAS features. Thus from this ANOVA analysis we can conclude that LTAS features indicate the stronger correlation with roughness (which might be due to degradation in the voice quality) and asthenia (indicates the degree of vocal weakness) compared to other perceptual characteristics.

## 5.4 Summary and conclusion

This study explored the state of the art filter bank-based LTAS features for the detection and identification of voice disorder. From the experimental results, it can be verified that classification accuracy for an identification system is less compared to detection system, as different disorders may share a common acoustic space. More interestingly, it was observed that the choice of filter bank in the extraction of LTAS features play an important role in the classification of voice disorders. In [69], SFFB based LTAS features showed the best performance for hyper-nasality detection, whereas, in this study, the SFFB-LTAS features showed better performance than CBFB-LTAS for the detection task. The CQFB-LTAS and GFB-LATS features showed better classification accuracy for the detection and identification of voice disorders, perhaps due to the underlying auditory filter banks (constant-Q filters and Gammatone filters). In addition, an improvement in the performance of detection and identification systems was observed with the combination of feature sets, which highlights the complementary nature of filter bank-based LTAS features. Further, we evaluated the relation between LTAS features and perceptual measure

(GRBAS scale available for HUPA database) using ANOVA analysis. The results from this experiment suggested that, most of the LTAS features have least value of the p (less than 0.5) for roughness and asthenia compared to grade, breathiness and strain. Compared to our previous study [41], significant improvement of performance for all the experiments was observed which might be due to the reason that, long term features can capture the voice disorders information in a better way as compared to the short term variations.

*Chapter 6*

# Detection and identification of voice disorders using features derived from Stockwell-Transform

From the chapter 5 it was concluded that voice quality related information associated with voice disorder is more prominently captured by in spectro-temporal domain. Hence, time-frequency methods, which can capture the glottal variations and formant variations from the speech signal, were explored for detection and analysis of voice disorders [195, 196]. This chapter explores S-Transform based time-frequency representation for voice disorder detection and identification system. In this regard, we explored different variants of S-Transform. We proposed cepstral features derived from S-Transform for the detection and identification of voice disorders. Also, by varying window-size, we studied how well the vocal tract system and excitation source information can be captured by S-Transform method. Additionally, we presented the effectiveness of S-Transform based spectral representation in capturing the acoustic correlates of different voice qualities. Performance of the proposed feature was compared with other baseline features. The complementary nature of the proposed features was explored by combining them with baseline features.

The chapter is organised as follows. Section 6.1 discusses the overview of literature studies done for analysis of voice disorders using various time-frequency methods. Section 6.2 describes the formulation of S-Transform and the method for extraction of proposed features. Section 6.3 presents a comparison of S-Transform based representation with other time-frequency representations. A detailed description of the experimental setup is given in Section 6.4.1. The experimental results of this work are presented in Section 6.5. Summary and conclusions are discussed in Section 6.6.

## 6.1 Studies in the analysis of voice disorders by time-frequency methods

Changes in acoustic characteristics due to voice disorders are reflected as variations in spectro-temporal domain [197]. In literature various time-frequency analysis methods were explored for pathological speech processing. In [198], features derived from Hilbert-Huang Transform (HHT) were used to detect voice disorders. Constant air leakage due to incomplete vocal fold closure results in noisy

components in the disordered speech signal. Therefore, average energy distribution over time in the time-frequency plane was observed to be smaller for pathological speech signals than healthy speech signals [199]. Features such as the octave max, octave mean, energy ratio, length ratio, and frequency ratio were extracted from the adaptive time-frequency transform (ATFT) algorithm for the automatic detection of voice disorders. Modulation spectral features were also investigated to detect voice disorders [200]. In [201], spectral features derived from empirical mode decomposition (EMD) were used to analyse and classify voice disorders. Different time-frequency methods like wavelet-transform [202], zero-time windowing (ZTW) [203], and single frequency filtering (SFF) [204] were also explored for the discrimination of different voice qualities. From these studies it can be concluded that better time-frequency representation is essential for classification of voice disorders and voice qualities.

S-Transform is another time-frequency analysis method, provides good spectro-temporal resolution [205, 206, 207, 208, 209]. In the literature S-Transform based time-frequency representation was explored in acoustic echo cancellation [210], automatic speech recognition systems [211], speech enhancement [212, 213] and hearing-impaired speech recognition [214]. From the above studies, it was observed that S-Transform methods provide better time-frequency localization as compared to other representations like short-time Fourier transform (STFT), wavelet-transform, etc. This study explores S-Transform method for classification of voice disorders.

The primary objective of this study is to validate the effectiveness of S-Transform in differentiating the acoustic characteristics for subjects suffering with voice disorders from healthy subjects. The current study explores S-Transform method in identifying different types of voice disorders. Moreover, we also analysed the time-frequency representations derived from the S-Transform method for different voice qualities associated with voice disorders such as breathiness, harshness, and creakiness. We also proposed use of cepstral features derived from S-Transform for the automatic detection and assessment of voice disorders. The performance of proposed system is compared with other baseline cepstral features derived from excitation source, vocal tract system, and time-frequency methods such as wavelet transform, ZTW and SFF.

## 6.2 Stockwell-Transform and cepstral feature extraction

This section describes about formulation of S-Transform, and its variants. In addition, it also explains the effect of segment length and scaling parameter on S-Transform of speech signal. Sub-section 6.2.4 discusses extraction of proposed features from S-Transform.

### 6.2.1 S-Transform

S-Transform is a time-frequency analysis method proposed by Stockwell [205, 215]. It has been used in many signal processing applications for the analysis of non-stationary signals [216, 217, 218, 219, 220]. It preserves the phase information of signal like STFT, and also provides frequency dependent

time-resolution property like wavelet-transform [205]. For a time varying signal $x(t)$ the continuous time S-Transform $S_x(\tau, f)$ is formulated as,

$$S_x(\tau, f) = \int_{-\infty}^{\infty} x(t) g(\tau - t, f) e^{-2\pi j f t} dt \tag{6.1}$$

where $g(t, f)$ represents Gaussian window and is given by,

$$g(t, f) = \frac{1}{\sigma(f)\sqrt{2\pi}} e^{\frac{-t^2}{2\sigma(f)^2}} \tag{6.2}$$

where, $\sigma(f)^2 = \frac{1}{|f|^2}$ represents variance of the Gaussian window. From Equation 6.1 and Equation 6.2, the S-Transform $S_x(\tau, f)$ of $x(t)$ can be denoted as,

$$S_x(\tau, f) = \int_{-\infty}^{\infty} x(t) \frac{|f|}{\sqrt{2\pi}} e^{\frac{-(\tau - t)^2 f^2}{2}} e^{-2\pi j f t} dt \tag{6.3}$$

The Gaussian window used in S-Transform is a function of both time and frequency. The standard deviation of the Gaussian window is reciprocal of frequency. For the low frequency, the window is wider in time domain to get better frequency resolution, while narrow window for high frequency, gives better time resolution (as shown in the Figure 6.1. Alternatively, the S-Transform can be formulated as

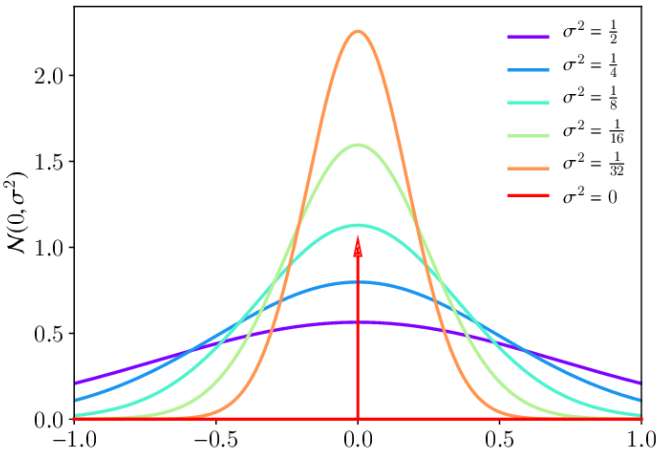


Figure 6.1: Illustration of Gaussian window by varying the variance.

in [221, 209], and it is given by,

$$S_x(\tau, f) = \int_{-\infty}^{\infty} X(\alpha + f) e^{\frac{-2\pi^2\alpha^2}{f^2}} e^{2j\pi\alpha\tau} d\alpha \tag{6.4}$$

where, $X(\alpha + f)$ and $e^{\frac{-2\pi^2\alpha^2}{f^2}}$ are frequency responses of $x(t)e^{-2\pi j f t}$ and $g(t, f)$, respectively. From the above equation, for a discrete time signal $x[n]$, $\tau = n\Delta T$, $f = m\Delta f$ and $\alpha = p\Delta f$ the discrete

S-Transform (DST) [205] $S[n\Delta T, m\Delta f]$ is given by,

$$S[n\Delta T, m\Delta f] = \begin{cases} \sum_{p=0}^{N-1} X[(p+m)\Delta f] e^{\frac{-2\pi^2 p^2}{m^2}} e^{j2\pi pn} & n \neq 0 \\ \frac{1}{N} \sum_{p=0}^{N-1} X[p\Delta f] & n = 0 \end{cases} \tag{6.5}$$

where, $X[(p+m), \Delta f]$ is DFT of a sequence $x[n]$ of length $N$. Advantage of fast Fourier transform (FFT) algorithm can be used in computing the discrete S-Transform [209]. Steps for computation of S-Transform of a discrete sequence are summarized as,

1. Select one frequency point and compute the DFT of input signal x(n) i.e. $X|p\Delta f|$.
2. Calculate DFT of N-point Gaussian function to select the frequency range i.e. $e^{\frac{-2\pi^2 p^2}{m^2}}$.
3. Shift spectrum $X|p\Delta f|$ to $X|(p+m)\Delta f|$, such that frequency of the spectrum to be selected matches with the zero-frequency of the frequency selecting Gaussian function.
4. Compute the IDFT of the resulted signal from the previous step for all the frequencies.

### 6.2.2 Effect of segment size on S-Transform of speech signal

S-Transform is computationally complex for longer sequences like speech, hence, it needs to be computed over smaller segments of a speech signal [211]. In this regard, we studied the effect of segment size on S-Transform of speech signal to understand what segment size should be chosen so it effectively captures speech signal information like excitation source and vocal tract system information, at lower computational cost.

We studied the S-Transform representation of speech signal for three different segment sizes 5 ms, 20 ms, and 100 ms, as shown in Figure 6.2. As it can be understood from the Figure 6.2 that for segment size of 5 ms, S-Transform provides good time resolution (vertical striations) but poor frequency resolution (horizontal striations). Whereas, for segment sizes 20 ms and 100 ms, S-Transform provides better time and frequency resolution. However, for longer size of segments (for example 100 ms), spectral components of speech regions with low energy are masked due to adjacent higher energy regions (region shown in rectangular box in Figure 6.2(d)). Also, for segments of longer size, computational complexity will become more. Hence, in order to get good time-frequency resolution along with minimal computation cost, in this study the segment size is chosen as 20 ms for calculation of S-Transform.

### 6.2.3 Variants of S-Transform

S-Transform proposed by Stockwell uses a Gaussian window (as discussed in the sub-section 6.2.1), whose standard deviation is inversely proportional to frequency. As frequency increases, width of the Gaussian window decreases, irrespective of analysed signal. This version of S-Transform is referred to

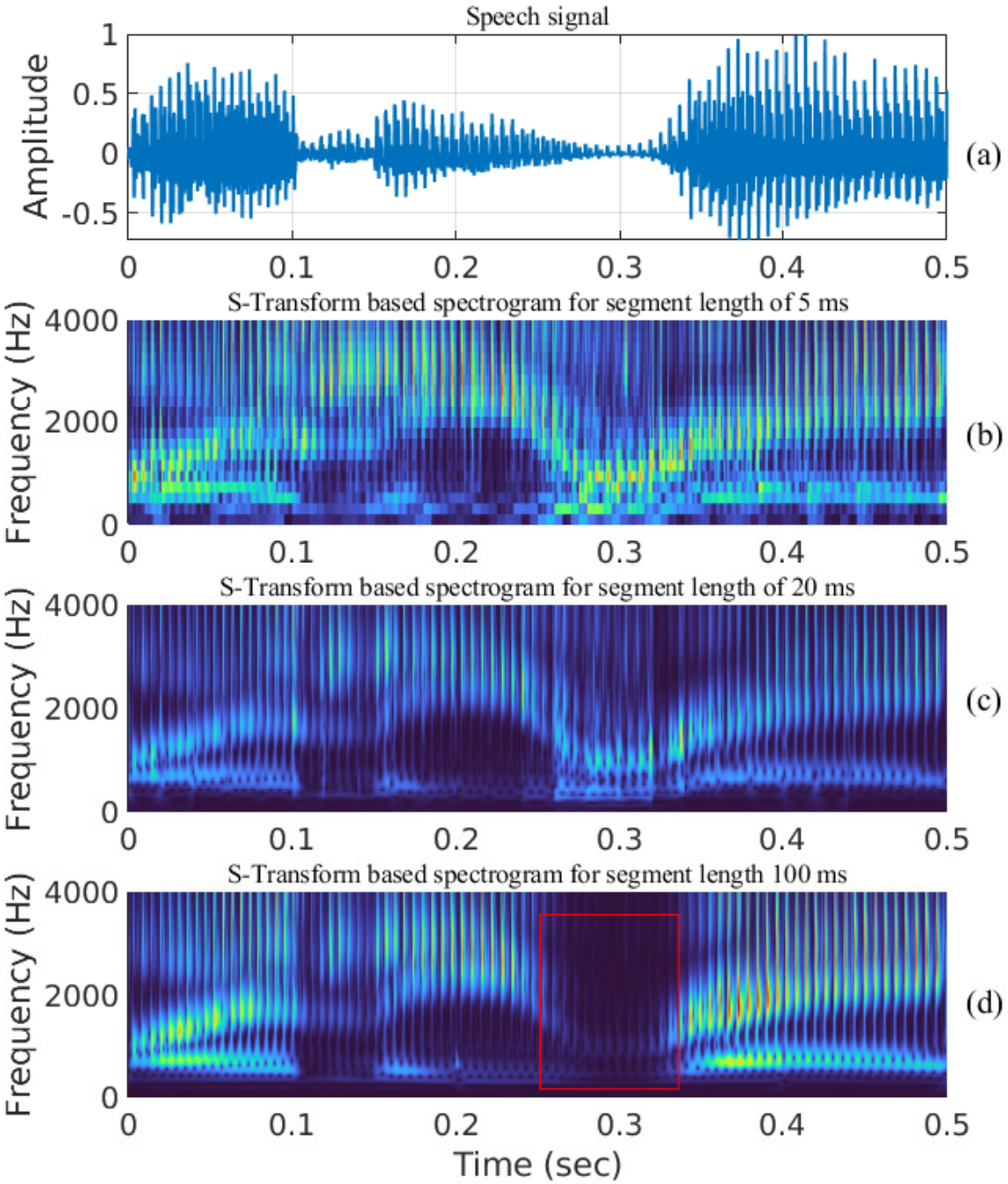



Figure 6.2: Illustration of spectrograms obtained with S-Transform for speech signal by varying the segment size. (a) Speech Signal. (b)-(d) S-Transform based spectrogram for segment length of 5 ms, 20 ms and 100 ms, respectively.

as standard S-Transform in this study. In literature many other variants of S-Transform were proposed to maximize the energy localization. Different parameters were introduced in the Gaussian window to provide better time-frequency localization. This subsection discusses the three different variants of S-Transform, namely, Sejdic's S-Transform [222], Assous's S-Transform [223], and optimized S-Transform [209] used in our study.

To improve the frequency resolution provided by S-Transform Sejdic et al. [222] introduced a new parameter $p$ which controls the shape of Gaussian window. Modified standard deviation of Gaussian window is given by,

$$\sigma(f) = \frac{1}{|f|^p} \tag{6.6}$$

The modified S-Transform is referred to as Sejdic's S-Transfrom in this study. From the experiments done in [222] it was found that the value of $p$, between 0 and 1, provides better time-frequency resolution. In our study value of $p$ is considered as 0.8, as used by [209].

Another way to control the shape of Gaussian window in S-Transform is provided in [223]. This modification of S-Transform is referred to as Assous's S-Transform and the new standard deviation of Gaussian window according to Assous et al. [223] is given by,

$$\sigma(f) = \frac{mf + k}{f} \tag{6.7}$$

where $m$ and $k$ represent constant parameters introduced to control the width of Gaussian window to provide better time and frequency resolution. Based on the experiments by Assous et al. [63], these parameters are chosen as $m = 0.05$ and $k = 0.1$ in this current study.

Recently, authors in the study [209] introduced four parameters to control the width of Gaussian window and also proposed an algorithm to select the optimal value of these parameters. In this case, the modified standard deviation used in Gaussian window is given by,

$$\sigma(f) = \frac{mf^p + k}{f^r} \tag{6.8}$$

The equation of Gaussian window can be rewritten as

$$g(\tau - t, f) = \frac{|f|^r}{(mf^p + k)\sqrt{2\pi}} e^{\frac{-(\tau - t)^2 f^{2r}}{2(mf^p + k)^2}} \tag{6.9}$$

Optimum values for these parameters $m,\ p,\ k$, and $r$ to get better time-frequency resolution were chosen by Moukadem et al. [209] as 0.3, 0.0386, 0.4276 and 0.6035, respectively. In our study, we refer this modified S-Transform as optimized S-Transform. Different variants of S-Transform and corresponding parameters are summarized in the following Table 6.1.

In order to understand which variant of the S-Transform offers better time-frequency resolution such that it can capture the excitation source and vocal tract system information effectively, we compare the spectrograms obtained from various S-Transform variants. Figure 6.3 shows the comparison of spectrograms obtained from different variants of S-Transform. From Figure 6.3 (b), we can observe

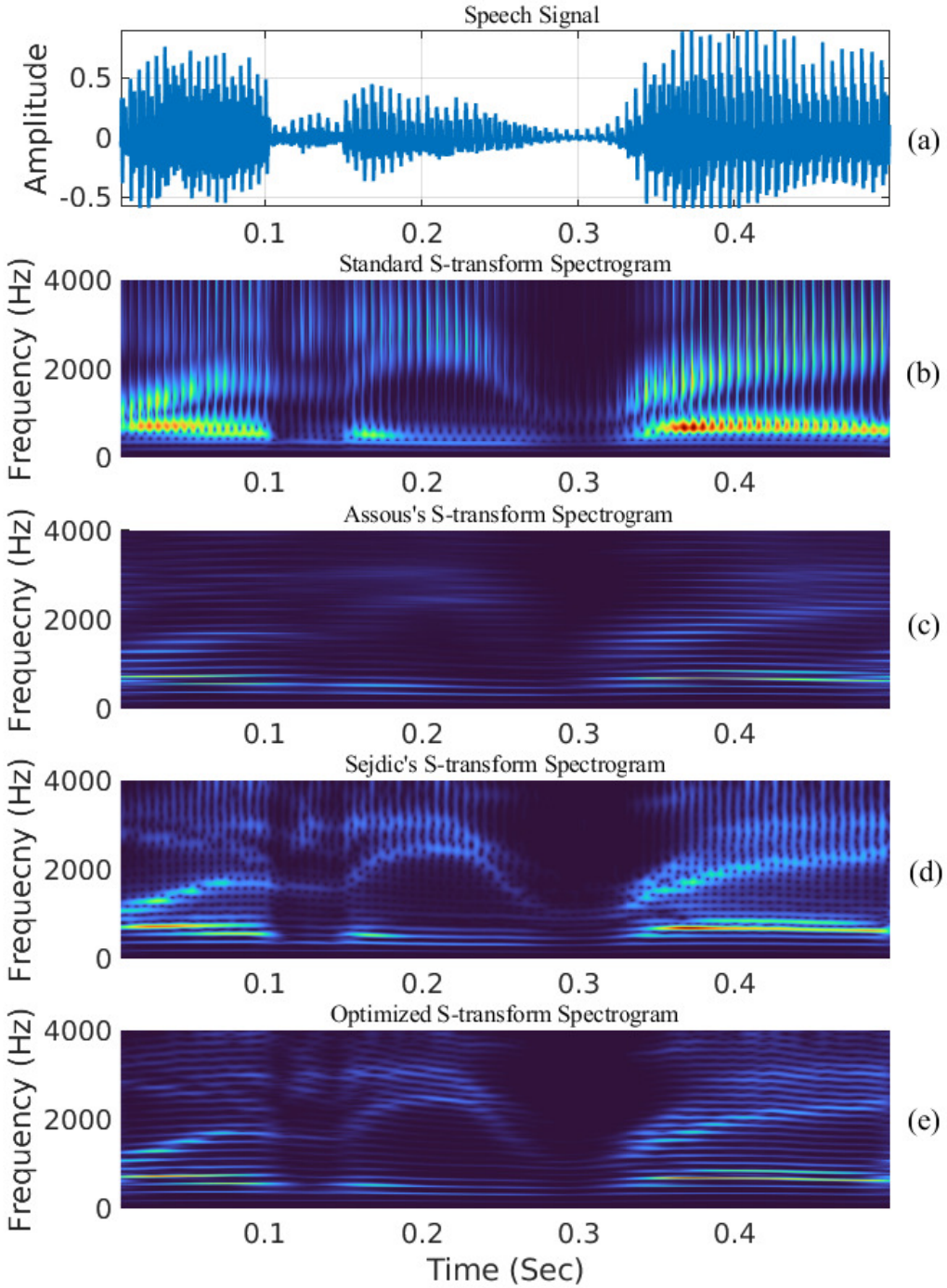



Figure 6.3: Illustration of spectrograms obtained for speech signal from different variants of S-Transform. (a) Speech signal, (b) Standard S-Transform spectrogram, (c) Assous's S-Transform spectrogram, (d) Sejdic's S-Transform spectrogram, (e) Optimized S-Transform spectrogram.

Table 6.1: Standard deviation of Gaussian window and its parameters for different variants of S-transform.

| Variants of S-Transform | Standard deviation of Gaussian window $\sigma(f)$ | Parameters |
|---|---|---|
| Standard S-Transform | $\frac{1}{f}$ | - |
| Sejdic's S-Transform | $\frac{1}{f^r}$ | r=0.8 |
| Assous's S-Transform | $\frac{mf+k}{f}$ | m=0.05, k=0.1 |
| Optimized S-Transform | $\frac{mf^p+k}{f^r}$ | m=0.3, p=0.0386, k=0.4276, r=0.6035 |

standard S-Transform provides good time resolution (at both low and high frequency bands). On the other hand, it provides good frequency resolution in low frequency bands compared to high frequency bands. From Figure 6.3 (c) and (e), we can observe Assous's and optimized S-Transform provides good frequency resolution (at both low and high frequency bands). On the other hand, it provides poor time resolution. From Figure 6.3 (d), we can observe Sejdic's S-Transform provides good frequency resolution (at both low and high frequency bands). On the other hand, it provides good time resolution in high frequency bands compared to low frequency bands.

From this graphical representation it can be observed that, formant transitions are captured efficiently by standard, Sejdic's and optimized S-Transform, as shown in Figure 6.3(b, d, & e). On the other hand, excitation source information is effectively represented by standard and Sejdic's S-Transform, as shown in Figure 6.3(b & d). It can be concluded from the above figures that, standard S-Transform can effectively capture both excitation source and vocal tract information simultaneously from speech signal.

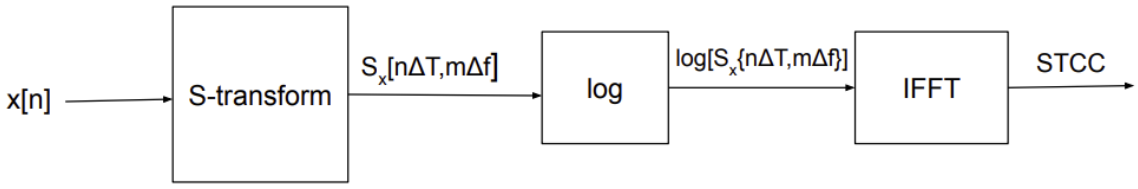



Figure 6.4: Block diagram of S-Transform cepstral coefficients (STCCs) extraction.

### 6.2.4 Extraction of cepstral coefficients from S-Transform

This subsection explains the procedure for extraction of cepstral coefficients from S-Transform. The block diagram representation of the feature extraction from S-Transform based time-frequency representation is shown in the Figure 6.4. For a discrete time speech signal $y[n]$, then pre-emphasised signal x[n] is given by

$$x[n] = y[n] - 0.9y[n-1] \tag{6.10}$$

Then, S-Transform is computed for each segment of size 20 ms. To obtain the cepstral representation from S-Transform, logarithm and discrete cosine transform (DCT) operations are applied and it is formulated as,

$$c\left(k,n\right) = IFFT\left\{log\left\{S_x\left(n\Delta T, m\Delta f\right)\right\}\right\} \tag{6.11}$$

Finally, first 13 dimensional static features and corresponding delta, and delta-delta features were computed resulting in 39-dimensional feature vector, which is referred to as S-Transform Cepstral Coefficients (STCC) in this study.

## 6.3 Importance of S-Transform in analysing the voice disorders

S-Transform uses Gaussian window whose standard deviation is inversely proportional to frequency which provides better time resolution at high frequencies along with better frequency resolution at low frequencies. Hence it is hypothesised that S-Transform method can represent the speech information in a better way. To understand, how well the S-Transform captures the phonation related information from speech signal, we have analysed the time-frequency representation obtained from S-Transform for different phonation types. Further, we compared S-Transform based representation with spectrograms obtained from STFT, SFF [191, 190], and ZTW [224] methods. In this regard modal and four non-modal phonation types, namely, breathy, creaky, harsh, and falsetto phonation are considered for analysis as illustrated in Figure 6.5-6.7.

From Figure 6.5 we can observe that formant transitions are captured in a better way by all the time-frequency representation methods. However, compared to STFT, all other methods capture phonation related information (epoch locations, energy variations with in glottal cycles, etc) in a better way. It can also observed as compared to SFF and ZTW method, S-Transform effectively captures the energy variations within the glottal cycle. Moreover, speech regions with low energy (region shown in rectangular box) are represented effectively by the S-Transform compared to other methods, as shown in Figure 6.5(e).

Figure 6.6 illustrate comparison of STFT, SFF, ZTW, and S-Transform Spectrogram for breathy and creaky phonation. In case of breathy phonation, the low muscle tension, medium longitudinal tension, and weak medical compression, results in minimum adduction of vocal folds [65]. Hence, the air is leaked through vibrating vocal folds resulting in turbulence or aspiration noise. As a result of aspiration noise, high frequency harmonics in breathy speech are significant compared to normal phonation. From

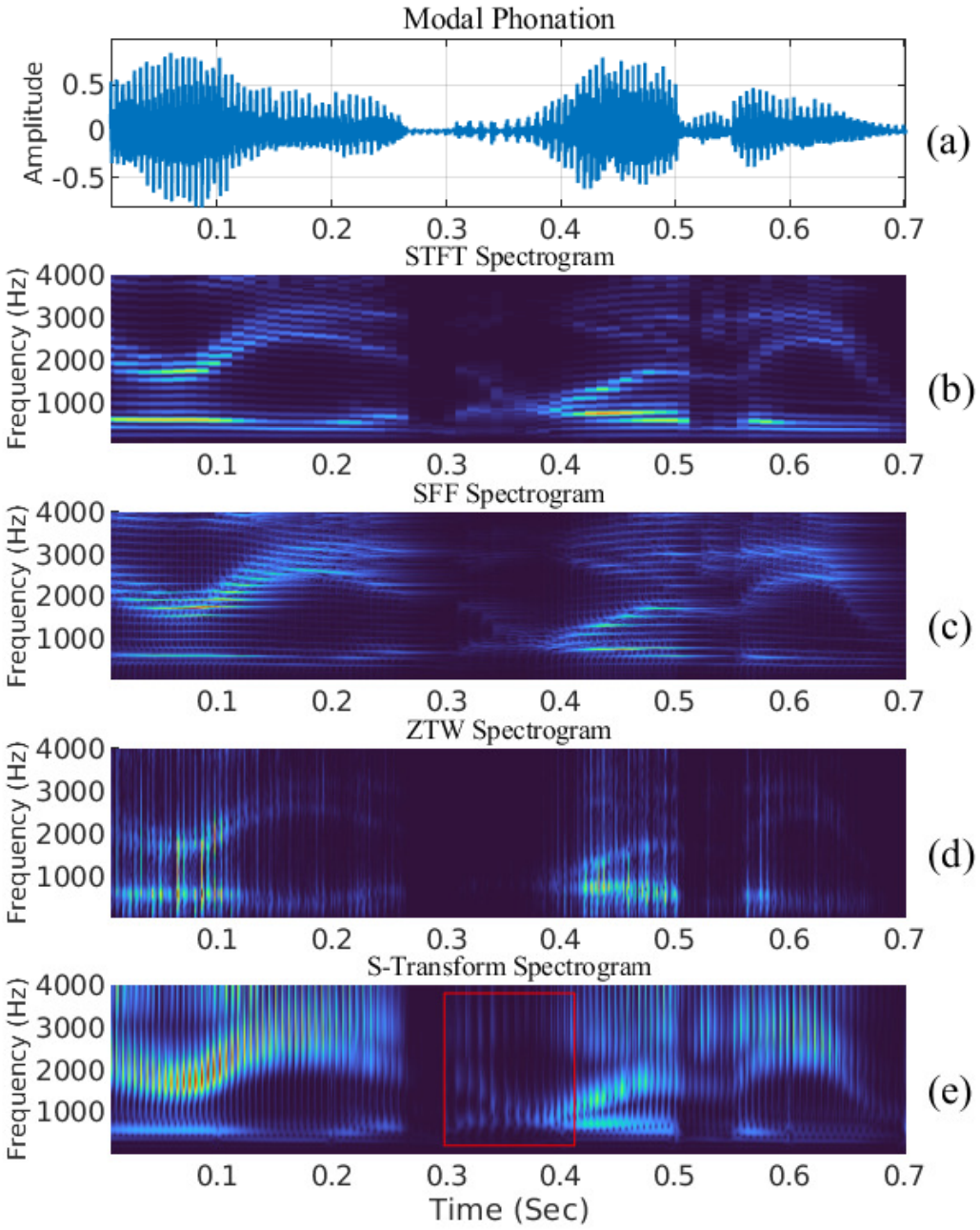



Figure 6.5: Illustration of spectrograms obtained from STFT, SFF, ZTW, and S-Transform methods for modal phonation. (a) Speech signal, (b) STFT spectrogram, (c) SFF Spectrogram, (d) ZTW spectrogram, (e) S-Transform spectrogram.

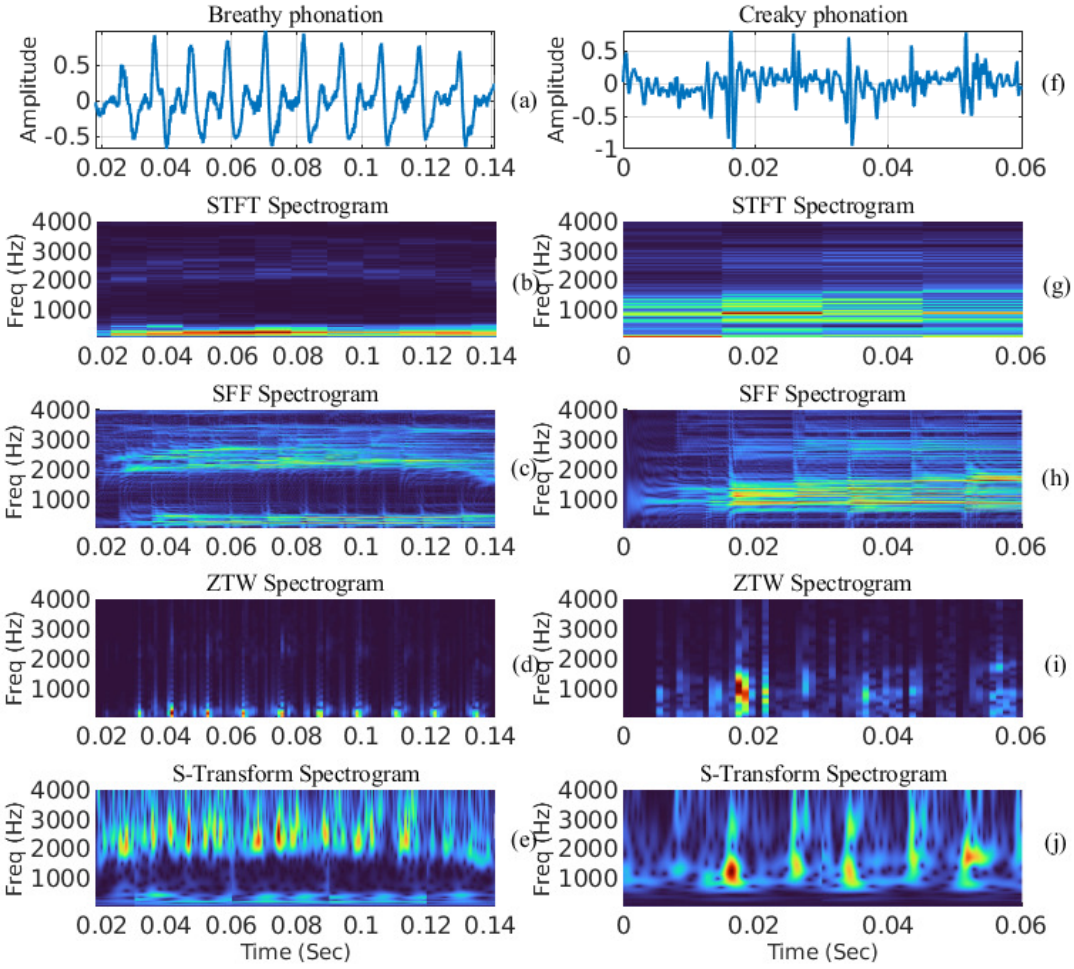



Figure 6.6: Illustration of spectrograms obtained from STFT, SFF, ZTW, and S-Transform methods for breathy and creaky phonation. (a) and (f) Speech signal, (b) and (g) STFT spectrogram, (c) and (h) SFF spectrogram, (d) and (i) ZTW spectrogram, (e) and (j) S-Transform spectrogram for breathy and creaky phonation, respectively.

the Figure 6.6, it can be observed that SFF and S-Transform based spectrograms effectively captures the high frequency harmonics present in breathy speech, compared to STFT and ZTW based spectrograms. Moreover, S-Transform method captures the high frequency harmonics and its variations present within the glottal cycle. In case of creaky phonation, vocal folds are adducted with weak longitudinal tension, and low subglottal pressure results in low and irregular fundamental (F0) frequency, and presence of secondary excitation (period-doubled vibration). Acoustic event like period-doubled vibration present in creaky speech, is represented by the S-Transform very well compared to any other method, as shown in Figure 6.6.

Figure 6.7 illustrate comparison of STFT, SFF, ZTW and S-Transform Spectrogram for harsh and falsetto phonation. Harsh phonation involves high longitudinal tension, high medical compression and strong adductive tension. These laryngeal settings result in high-frequency harmonics in speech signal. Harsh phonation has the sharpest closure with the least open quotient compared to other phonation types. In case of falsetto phonation, vocal folds are stretched longitudinally which result in thin vibrating mass. F0 is typically higher in case of falsetto phonation than modal phonation. Figure 6.7, it is understood that S-Transform can highlight glottal closer instants (GCIs) and glottal opening instants (GOIs), which are important for discriminating the different phonation types. Also, we can observe from the Figure 6.7 that the acoustic characteristics associated with harsh and falsetto phonation are observed from S-Transform representation in a better way as compared to other time-frequency representation. From all the above figures, we can draw the conclusion that the S-Transform can capture acoustic characteristics of various phonations such as speech region with low energy, high frequency harmonics, GCI, and GOI events within glottal cycle more efficiently compared to other baseline time-frequency methods.

## 6.4 Database and experimental setup

### 6.4.1 Database

SVD dataset [150], is considered in this study for performing voice disorder detection and identification task. Additionally, HUPA database [151] is considered for performing voice disorder detection experiment.

- SVD database used in this study also groups different categories of voice disorders into four classes, namely, *Structural, Neurogenic, Functional and Psychogenic*. A total of 625 samples from healthy class and 950 voice samples from different voice disorder categories, for vowel /a/, /i/, and /u/ in normal, high, low and rising-falling pitch, were considered in this study. All the speech recordings were resampled to 8000 Hz.

- The HUPA database contains speech recordings of both healthy and pathological samples of 440 speakers for the vowel /a/. These 440 recordings include 239 voice disorder samples and 201 normal samples.

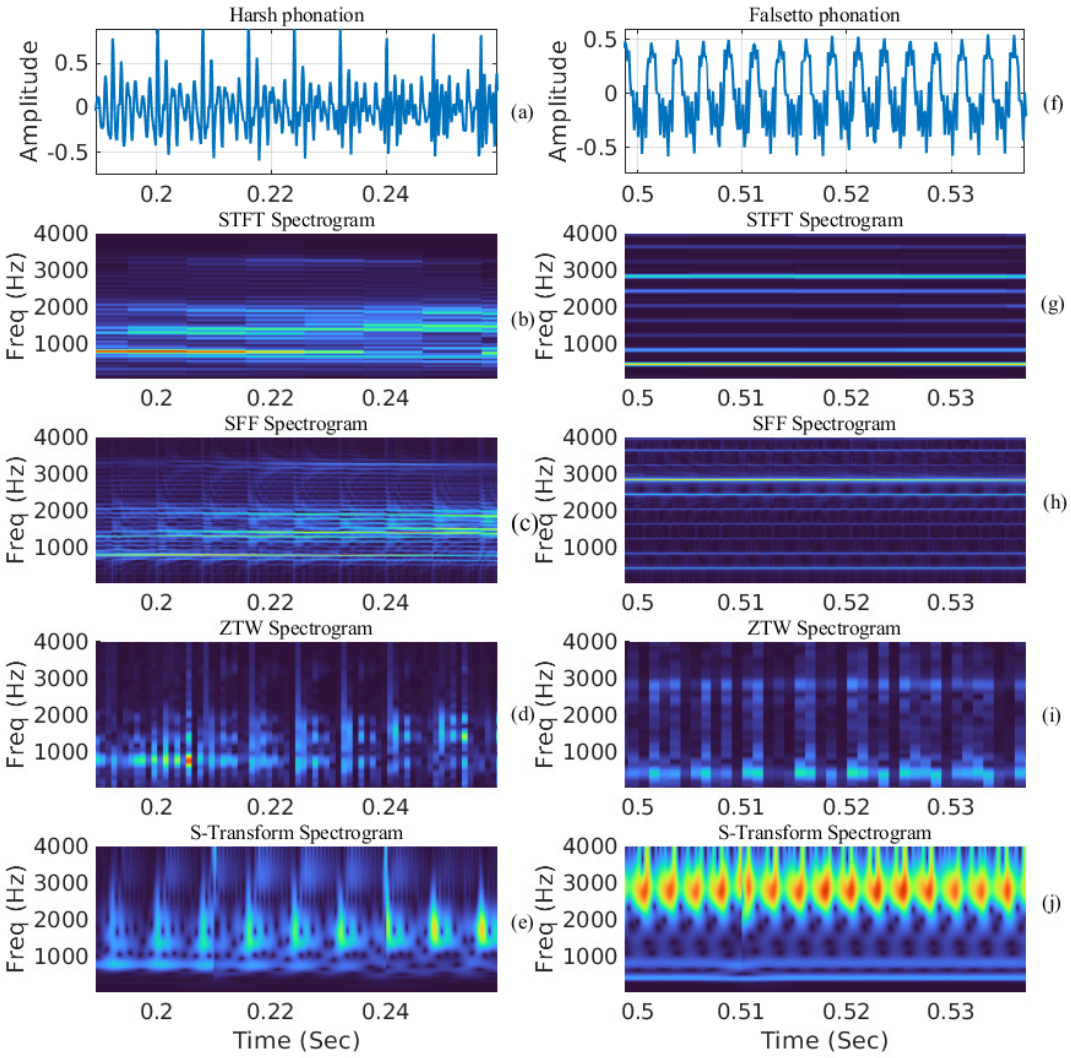



Figure 6.7: Illustration of spectrograms obtained from STFT, SFF, ZTW, and S-Transform for harsh and falsetto phonation. (a) and (f) Speech signal, (b) and (g) STFT spectrogram, (c) and (h) SFF spectrogram, (d) and (i) ZTW spectrogram, (e) and (j) S-Transform spectrogram for harsh and falsetto phonation, respectively.

Table 6.2 describes the number of healthy and voice disorders samples of SVD and HUPA database used for detection task. Table 6.3 describes the different categories of voice disorders available for SVD database used in our study for the identification task of voice disorders.

Table 6.2: Details of the number of sample used for the detection task from SVD and HUPA database.

| SVD database | | HUPA database | |
|---|---|---|---|
| Healthy | Voice Disorder | Healthy | Voice Disorder |
| 625 | 950 | 239 | 201 |

Table 6.3: Details of the different classes of SVD database and number of sample used in our experiment for the identification task. SD: Structural voice Disorder, NVD: Neurogenic Voice Disorder, FVD: Functional Voice Disorder, PVD: Psychogenic Voice Disorder.

| Disorder type | Disorder name | #Samples |
|---|---|---|
| Organic Voice Disorder (OVD) | SD | 352 |
| | NVD | 253 |
| Non-organic Voice Disorder (NOVD) | FVD | 254 |
| | PVD | 91 |

### 6.4.2 Features

In this study, three sets of features are used as baseline features for comparing the performance of voice disorder detection and identification system with proposed S-Transform based cepstral coefficients. Features extracted from the evidence of excitation source signals are considered as first set of baseline features. Cepstral coefficients extracted from speech signal that models the vocal tract system are used as second set of baseline features. Mel frequency cepstral coefficients extracted from spectro-temporal analysis methods like SFF and ZTW are considered as third set of baseline feature.

#### 6.4.2.1 Baseline feature set-1

Baseline feature set 1 contains set of features extracted from excitation source signal. Excitation source evidence like GVV [40, 63], ZFF [25, 40, 56], and LP [54, 225] residual are considered for extraction of features.

- MFCC of excitation source signal like LP residual and ZFF signal are computed using frame-length of 20 ms and a frame-shift of 5 ms [41, 204]. 13 static coefficients, and their first and second order derivatives which makes total 39 dimension feature vector. Finally 4 statistics were computed resulting in 156 dimension MFCC-Residual and MFCC-ZFF feature vector.

- Glottal flow signal is derived from QCP [63] method. From glottal flow signals time and frequency domain parameters, namely OQ1, OQ2, OQa, QoQ, ClQ, SQ1, SQ2, AQ, NAQ, H1-H2, PSP, and HRF are calculated. 16 statistics of 12-dimensional glottal parameters are calculated which results in 192 dimension glottal feature vector.

- Intonation feature vector [25, 41, 69] is set of features used to model frequency and amplitude perturbation from speech signal. It is 79 dimension feature vector contains statistics of F0, 66 perturbation parameters (jitters and shimmers estimated from SoE contour, and EoE contour), 4 HNR parameters and PPE measure.

#### 6.4.2.2 Baseline feature set-2

Features extracted from speech signal which model the vocal tract information are considered as baseline feature set 2 in this study. MFCC [50], LPCC [49, 226], and CQCC [52, 227] features are computed using speech segments of 20 ms frame size with a 5 ms frame shift. First 13 dimensional static features and corresponding delta, and delta-delta features were computed resulting in 39-dimensional features. Statistical averages such as mean, standard deviation, kurtosis and skewness were derived from these frame-level features.

#### 6.4.2.3 Baseline feature set-3

This baseline feature set contains cepstral coefficients extracted from SFF, ZTW, and wavelet transform methods.

- SFF method was proposed in [191] to derive the amplitude envelope of speech signal at each frequency. SFF methods provides good spectro-temporal resolution. From the SFF method time-frequency distribution of speech signal is obtained by passing the signal through single frequency filter-bank (SFFB). SFFB is set of complex band pass filter which decomposes the signal into number of frequency bands [190]. Cepstral coefficients are derived from the SFF spectrum are termed as single frequency cepstral coefficients (SFFCCs). Each feature contains 13 static, delta and delta-delta coefficients resulting in 39-dimension cepstral feature vector. Each utterance is represented by fixed 156-dimension feature vector by calculating four (mean, standard deviation, skewness and kurtosis) statistical averages on 39-dimension cepstral coefficients.

- To capture the high spectral resolution at each instant, ZTW method was proposed [224]. In this method speech is multiplied by high decaying (impulse-like) window to derive instantaneous spectral characteristic, moreover numerator of group delay spectrum is used to provide the good spectral resolution. Cepstral features derived from ZTW method are termed as zero-time windowing cepstral coefficients (ZTWCCs). ZTWCC is 156 dimension vector obtained by calculating four statistical averages on 39 dimension cepstral feature vector.

- Wavelet transform [228, 229] is mathematical tool to analyse the signal in both time and frequency domain, simultaneously. It is considered as scaled version of a single function called the "mother wavelet". Mother wavelet is characterised by two coefficients dilation and translation. Dilation coefficient is inversely proportional to frequency produces various versions of the wavelet, either stretched or compressed, and these are then shifted along the time axis by a translation factor to represent the movement of the analysis window in time. This approach results in good frequency resolution at low frequencies, while good time resolution at high frequency. Speech signal is first decomposed into its time-frequency component using the wavelet transform. Subsequently, from these components, 39 dimension (13 static, 13 delta, and 13 delta-delta) cepstral coefficients are derived and referred to as wavelet transform based cepstral coefficients (WTCC) [230]. Finally, 4 statistics averages, namely mean, standard deviation, kurtosis, and skewness were calculated, resulting 156 dimensional feature vector.

#### 6.4.2.4 Proposed features

For computation S-Transform based spectrum, first the speech signal is divided into non-overlapping segments of 20 ms. Then cepstral coefficients are obtained from the S-Transform spectrum (as explained in section 6.2.4). Cepstral coefficients consist of 13 static coefficients, and their first and second order derivatives. Finally, 4 statistics, namely mean, standard deviation, kurtosis and skewness were calculated, resulting in 156 dimension STCC feature vector. Along with standard S-Transform we have also explored three variants (Sejdic's, Assous's and optimized) of S-Transform for extraction of features.

### 6.4.3 Classifier

This study explores different machine learning algorithms for performing detection and assessment tasks such as support vector machine (SVM), logistic regression (LR), naive Bayes and neural network. 5-fold cross-validation experiments were performed by randomly partitioning the dataset into 5 equal sets. Out of these 5 sets one is reserved for testing the classifiers while the other four are used for training. The average classification accuracy is calculated by repeating the process for 5 times. We use grid search to select optimum values of kernel parameters. Performance is also reported in terms of area under receiver operating characteristics (ROC) curves, termed as AUC, and F1-score along with average classification accuracy. Classification accuracy is defined as the ratio of the number of correct predictions (true positive and true negative) detected by a model to the total number of predictions. F1-score measures the balance between precision and recall. AUC measures the area under the ROC curve. The value of AUC and F1-score lies between 0 and 1. If the value (of AUC and F1-score) is 0, it represents the worst classifier, while 1 represents the perfect classifier.

## 6.5 Results and Discussion

The primary objective of this study is to investigate the importance of S-Transform based time-frequency representation for voice disorder detection and identification systems. In this regards experiments have been carried out using cepstral features extracted from S-Transform representation on SVD and HUPA databases. SVD dataset is used for performing voice disorder detection and identification experiments, HUPA dataset is used for performing only detection task as speech samples of different categories of voice disorders are not available. The results of different experiments are discussed in the following sub-sections.

### 6.5.1 Performance analysis of different classifiers using S-Transform features

The main aim of these experiments is to analyze the performance of different classifiers using S-Transform based features for voice disorder detection task. Table 6.4 and Table 6.5 depicts performance of different classifiers for different variants (as referred in Section 2.2) of S-Transform based cepstral coefficient features on HUPA and SVD databases, respectively. In this regard, different classifiers like support vector machine (SVM), logistic regression (LR), Naive Bayes (Gaussian distribution (B1) and Kernel distribution(B2)), and neural network (narrow (N1) and Medium (N2)) classifiers are explored. In addition, the experiments were performed using the different kernels of SVM like linear (K1), quadratic (K2), cubic (K3), fine (K4), and coarse (K5) Gaussian.

We can observe from the Table 6.4 and Table 6.5 that in case of SVM classifier, different kernels (K1, K2, K3, and K5) performs better among all the other classifiers except for the kernel K4. Among the different classifiers, LR and Naive-Bayes classifiers performing poor for different variants of STCC features on HUPA and SVD databases, respectively. As compared to LR, B1, and B2, neural network classifiers performs better for both the database. In case of neural network, by increasing the number of neurons from 10 (N1) to 25 (N2), performance is not improved much which may be due limited amount of training data available. From this experimental results, best average classification accuracy of 79.03% and 77.3% is obtained using the K2 kernel on HUPA and SVD database, respectively. Hence, SVM classifier with quadratic kernel (K2) is chosen among the other classifiers, for performing rest of the experiments in this study.

### 6.5.2 Performance analysis of S-Transform features for voice disorder detection

Voice disorder detection is a binary classification task which distinguish the voice disorder class from the healthy class. This section discusses the performance analysis of voice disorder detection system developed with different feature set. In this regard, classification system is trained on SVM classifier with quadratic kernel (as discussed in Section 6.5.1) for both HUPA and SVD database. The results are reported in terms of classification accuracy, AUC, and F1-score. Table 6.6 shows the performance of voice

Table 6.4: Performance of voice disorder detection system using S-Transform based cepstral features on HUPA database in terms of classification accuracy (in %) for different machine learning classifiers.

| | SVM | | | | | LR | Naive bayes | | Neural Network | |
|---|---|---|---|---|---|---|---|---|---|---|
| Features | K1 | K2 | K3 | K4 | K5 | L1 | B1 | B2 | N1 | N2 |
| Standard STCC | 78.6 | **79.5** | 78.6 | 58 | 75 | 73 | 73.4 | 73.4 | 74.5 | 78.6 |
| Assos's STCC | 78.2 | **78.9** | 78 | 58.6 | 77.5 | 69.5 | 73.9 | 75.9 | 73.4 | 77 |
| Sejdic's STCC | 77 | **80.2** | 79.3 | 57 | 77 | 68.4 | 73.9 | 74.8 | 77 | 76.8 |
| Optimized STCC | 79.1 | 77.5 | **79.8** | 58.4 | 76.1 | 70.2 | 73.2 | 75 | 76.8 | 74.3 |
| Average accuracy | 78.23 | **79.03** | 78.93 | 58 | 76.4 | 70.28 | 73.6 | 74.78 | 75.43 | 76.68 |

Table 6.5: Performance of voice disorder detection system using S-Transform based cepstral features on SVD database in terms of classification accuracy (in %) for different machine learning classifiers.

| | SVM | | | | | LR | Naive bayes | | Neural Network | |
|---|---|---|---|---|---|---|---|---|---|---|
| Features | K1 | K2 | K3 | K4 | K5 | L1 | B1 | B2 | N1 | N2 |
| Standard STCC | 78 | **79.8** | 79.5 | 60.2 | 73.5 | 78 | 68.4 | 68.5 | 75.3 | 76.6 |
| Assous's STCC | 74.4 | **76.9** | 76.5 | 60.2 | 70.7 | 73.9 | 69.5 | 67.3 | 72.9 | 72.4 |
| Sejdic's STCC | 74.5 | **76.4** | 75.3 | 60.2 | 71.9 | 74.3 | 67 | 67.7 | 70.2 | 70.7 |
| Optimized STCC | 73.9 | **76.1** | 74.5 | 60.2 | 69.5 | 72.4 | 66.6 | 66.5 | 71.6 | 72.2 |
| Average accuracy | 75.2 | **77.3** | 76.45 | 60.2 | 71.4 | 74.65 | 67.88 | 67.5 | 72.5 | 72.98 |

disorder system using baseline feature set and cepstral coefficient obtained from different variations of S-Transform.

From Table 6.6, it is observed that among the baseline excitation source based features, MFCC features extracted from ZFF signal shown good classification accuracy of 71.8% for HUPA database. Performance of baseline excitation source based features is better than vocal tract system features which indicates that these features capture information related to voice disorder in a better way. Among all the baseline features SFFCC performs best in terms of classification accuracy, AUC, and F1 score of 75.2%, 0.83, and 0.73, respectively for HUPA database. In comparison to the three sets of baseline features, all the variants of S-Transform based features gave better performance for voice disorder detection system. This improvement in the performance indicates that cepstral features obtained from S-Transform contains the information which can discriminate speaker suffering from voice disorder from healthy speaker in better way. Sejdic's based STCC features gave the best performance compared to all other features in terms of classification accuracy, AUC and F1-score of 80.2%, 0.88, and 0.79, respectively for HUPA database. For SVD dataset within the baseline features, WTCC performed better in terms of classification accuracy, AUC and F1-score of 78.7%, 0.86, and 0.83, respectively. Baseline features extracted

Table 6.6: Performance of voice disorder detection system using baseline features and S-transform based cepstral features in terms of classification accuracy (Acc.), area under the ROC curve (AUC), and F1-score on HUPA and SVD database.

| | HUPA | | | SVD | | |
|---|---|---|---|---|---|---|
| Features | Acc.(%) | AUC | F1-score | Acc.(%) | AUC | F1-score |
| MFCC-Residual | 69.1 | 0.75 | 0.65 | 71.4 | 0.78 | 0.77 |
| MFCC-ZFF | 71.8 | 0.79 | 0.69 | 71.8 | 0.78 | 0.77 |
| Intonation | 67 | 0.73 | 0.59 | 72.9 | 0.79 | 0.77 |
| Glottal | 71.4 | 0.78 | 0.66 | 70.6 | 0.77 | 0.76 |
| MFCC | 68.9 | 0.78 | 0.65 | 75.6 | 0.84 | 0.80 |
| LPCC | 70.5 | 0.75 | 0.67 | 75.3 | 0.81 | 0.80 |
| CQCC | 67.3 | 0.76 | 0.63 | 74.1 | 0.81 | 0.79 |
| SFFCC | 75.2 | 0.83 | 0.73 | 74.1 | 0.80 | 0.79 |
| ZTWCC | 73.6 | 0.82 | 0.71 | 74.8 | 0.81 | 0.79 |
| WTCC | 69.8 | 0.79 | 0.68 | 78.7 | 0.86 | 0.83 |
| Standard STCC | 79.5 | 0.87 | 0.77 | **79.8** | **0.87** | **0.84** |
| Assous's STCC | 78.9 | 0.86 | 0.76 | 76.9 | 0.84 | 0.82 |
| Sejdic's STCC | **80.2** | **0.88** | **0.79** | 76.4 | 0.82 | 0.81 |
| Optimized STCC | 77.5 | 0.86 | 0.75 | 76.1 | 0.82 | 0.81 |

from vocal tract system and time-frequency methods (SFF and ZTW methods) performed in between 74% to 75.6% in terms of classification accuracy for voice disorder detection task on SVD dataset. Comparing with the baseline features, standard STCC feature set performed best with classification accuracy of 79.8%, AUC of 0.87, and F1-score of 0.84, respectively for SVD database.

S-Transform based cepstral features results in improved performance (approximately of 4% in terms of classification accuracy) as compared to baseline features for both HUPA and SVD database which may be due to better time-frequency representation captured by S-Transform. It may be due to the reason that S-Transform based time-frequency representation captures variation in the glottal cycle along with formant transition variation in a better way as compared to other time-frequency representations (as discussed in Section 6.3).

### 6.5.3 Performance analysis of S-Transform features for voice disorder identification

Voice disorder identification task was performed in clinical perspective (as it was done in our previous work [41], [227] to identify the cause of voice disorder) for SVD database due to availability of different categories of voice disorders. It is performed using three multi-level classifier. The three binary classifiers are trained on SVM classifier (using the Quadratic kernel) to identify the category of voice disorder. Experimental results of voice disorder identification system are discussed in this section.

Table 6.7: Performance of voice disorder identification system using the baseline and STCC features on SVD database in terms of classification accuracy (Acc.), Area under curve (AUC) and F1-score. Here Exp. 2: Organic voice disorder vs non-organic voice disorder, Exp. 3: Structural voice disorder vs neurogenic voice disorder, Exp. 4: Functional voice disorder vs psychogenic voice disorder.

| | Exp. 2 | | | Exp. 3 | | | Exp. 4 | | |
|---|---|---|---|---|---|---|---|---|---|
| Features | Acc.(%) | AUC | F1-score | Acc.(%) | AUC | F1-score | Acc.(%) | AUC | F1-score |
| MFCC-Residual | 69.6 | 0.72 | 0.77 | 64.1 | 0.66 | 0.71 | 68.2 | 0.49 | 0.83 |
| MFCC-ZFF | 70.2 | 0.74 | 0.77 | 71.6 | 0.73 | 0.78 | 67.6 | **0.56** | 0.82 |
| Intonation | 68.7 | 0.7 | 0.75 | 63 | 0.66 | 0.74 | **68.5** | 0.5 | **0.83** |
| Glottal | 66.1 | 0.68 | 0.73 | 61.5 | 0.64 | 0.68 | 68.8 | 0.51 | 0.83 |
| MFCC | 69.9 | 0.73 | 0.77 | 70.6 | 0.75 | 0.76 | 65 | 0.53 | 0.82 |
| LPCC | 68.1 | 0.71 | 0.75 | 71.6 | 0.77 | 0.76 | 63.9 | 0.43 | 0.82 |
| CQCC | 70 | 0.73 | 0.77 | 68.3 | 0.72 | 0.75 | 67.3 | 0.5 | 0.83 |
| SFFCC | 70.2 | 0.77 | 0.77 | 69.3 | 0.74 | 0.75 | 65.9 | 0.42 | 0.83 |
| ZTWCC | 70.2 | 0.72 | 0.77 | 66.4 | 0.68 | 0.72 | 65.9 | 0.44 | 0.83 |
| WTCC | 69.9 | 0.77 | 0.74 | 69.4 | 0.76 | 0.75 | 65.3 | 0.43 | 0.78 |
| Standard STCC | 71.7 | 0.75 | 0.78 | **76** | **0.81** | **0.80** | 64.2 | 0.48 | 0.82 |
| Assous's STCC | 70.7 | 0.75 | 0.77 | 74.7 | 0.8 | 0.79 | 66.2 | 0.5 | 0.83 |
| Sejdic's STCC | **73.2** | **0.77** | **0.79** | 72.1 | 0.79 | 0.76 | 65.6 | 0.5 | 0.83 |
| Optimized STCC | 71.3 | 0.77 | 0.77 | 70.7 | 0.76 | 0.76 | 65.3 | 0.47 | 0.83 |

Table 6.7 shows the result of voice disorder identification system using the baseline features and S-Transform based cepstral features on SVD database. Exp. 2 performs discrimination of organic voice disorders from non-organic voice disorders, to detect the cause of voice disorder. Exp. 3 discriminate structural voice disorder from neurogenic voice disorder. Exp. 4 performs binary classification task to discriminate functional voice disorder from psychogenic voice disorder. For Exp. 2 among the baseline features, SFFCC performs best with classification accuracy and AUC of 70.2% and 0.77, respectively. Among all the variants of STCC features, Sejdic's STCC performed best for this task with classification accuracy, AUC, and F1-score of 73.2%, 0.77 and 0.79, respectively. As compared to other two tasks of identification, for Exp. 3, standard STCC feature set improved the performance up to 6% from the baseline features. From Table 6.7, cepstral features derived from standard representation of S-Transform outperforms the baseline feature with classification accuracy of 76 %. Exp. 4 is more challenging than the other two identification tasks. As we can observe from Table 6.7 that performance of this identification task is not improved with the STCC features. It may be due to the reason that functional

and psychogenic voice disorders share a common acoustic space [168]. Best classification accuracy of 68.5% is obtained with intonation features for Exp. 4.

### 6.5.4 Performance analysis of S-Transform and baseline feature combination for voice disorder detection and identification

Further to investigate the complimentary nature of proposed features set, combination of STCC and baseline features was explored for voice disorder detection and identification. And the results are reported in Table 6.8 and Table 6.9, respectively. From the results reported in the Table 6.6 and Table 6.7, it can be seen that standard STCC features shown consistently better performance for most of the detection and identification tasks. Hence, the baseline features are combined with standard STCC features for performing these experiments. An improvement in the performance of detection systems was observed from the Table 6.8, when STCC features are combined with baseline feature sets, which highlights the complementary nature of cepstral features obtained from S-Transform. The performance is improved in between 8 to 10% in terms of classification accuracy for both the database. The combination of MFCC features with STCC features outperformed all the other features in terms of classification accuracy if 82% and 81.7% for HUPA and SVD database, respectively. Fusion of glottal source features with S-Transform based features shown the comparable results in terms of classification accuracy of 81.3% and best AUC (of 0.89) among all the other combined feature set for HUPA dataset.

Table 6.8: Performance of voice disorder detection system using combination of features in terms of classification accuracy (Acc.), area under the ROC curve (AUC), and F1-score on HUPA and SVD database

| | HUPA | | | SVD | | |
|---|---|---|---|---|---|---|
| Features | Acc.(%) | AUC | F1-score | Acc.(%) | AUC | F1-score |
| MFCC-Residual + STCC | 79.5 | 0.87 | 0.78 | 79 | 0.87 | 0.83 |
| MFCC-ZFF+ STCC | 81.8 | 0.88 | **0.80** | 79.9 | 0.87 | 0.84 |
| Intonation + STCC | 76.6 | 0.86 | 0.74 | 79.1 | 0.87 | 0.83 |
| Glottal + STCC | 81.3 | **0.89** | 0.78 | 79.2 | 0.86 | 0.83 |
| MFCC+ STCC | **82** | 0.87 | **0.80** | **81.7** | **0.89** | **0.85** |
| LPCC+ STCC | 78.9 | 0.87 | 0.77 | 80.1 | 0.87 | 0.84 |
| CQCC+STCC | 74.5 | 0.85 | 0.72 | 81.3 | 0.88 | **0.85** |
| SFFCC+STCC | 79.1 | **0.89** | 0.77 | 80.5 | 0.87 | 0.84 |
| ZTWCC+STCC | 79.8 | 0.86 | 0.78 | 79.4 | 0.86 | 0.83 |
| WTCC+STCC | 77.7 | 0.86 | 0.76 | 80.7 | 0.87 | 0.84 |

Table 6.9 shows the results of voice disorder identification system computed for SVD database with combination of baseline feature with STCC features. It can be observed from the table that performance is improved for all the three tasks as compared to the individual feature set. For Exp. 2 and Exp. 3 the

best classification accuracy 73% and 76.2 is obtained when MFCC features are combined with STCC features. It can also be observed that even by combining the different features, classification accuracy for the Exp. 4 is not increased significantly. It may be due to the reason that assessment of psychogenic voice disorder requires multidisciplinary diagnosis involving speech analysis, non-communicative voice analysis along with behavioural analysis [169].

Table 6.9: Performance of voice disorder identification system using the combination of features on SVD database in terms of classification accuracy (Acc.), Area under curve (AUC), and F1-score. Exp. 2: Organic voice disorder vs non-organic voice disorder, Exp. 3: Structural voice disorder vs neurogenic voice disorder, Exp. 4: Functional voice disorder vs Psychogenic voice disorder.

| | Exp. 2 | | | Exp. 3 | | | Exp. 4 | | |
|---|---|---|---|---|---|---|---|---|---|
| Features | Acc.(%) | AUC | F1-score | Acc.(%) | AUC | F1-score | Acc.(%) | AUC | F1-score |
| MFCC-Residual+STCC | 71.1 | 0.77 | 0.77 | 74 | 0.79 | 0.78 | 66.8 | **0.53** | 0.79 |
| MFCC_ZF+STCC | 72.6 | 0.78 | 0.79 | 73.2 | 0.79 | 0.78 | 65.6 | 0.51 | 0.79 |
| Intonation+STCC | 70.7 | 0.77 | 0.77 | 69.1 | 0.77 | 0.74 | 64.5 | 0.44 | 0.78 |
| Glottal+STCC | 72.4 | 0.76 | 0.78 | 71.4 | 0.75 | 0.76 | 63.6 | 0.42 | 0.77 |
| MFCC+STCC | **73** | **0.79** | **0.79** | **76.2** | **0.82** | **0.80** | 65.3 | 0.56 | 0.78 |
| LPCC+STCC | 72.7 | 0.77 | 0.78 | 73.4 | 0.80 | 0.77 | 66.5 | 0.5 | 0.79 |
| CQCC+STCC | 71.8 | 0.77 | 0.78 | 75.9 | 0.80 | 0.80 | **68.2** | 0.52 | **0.80** |
| SFFCC+STCC | 71.7 | 0.77 | 0.78 | 72.2 | 0.79 | 0.77 | 65.9 | 0.44 | 0.79 |
| ZTWCC+STCC | 71.4 | 0.77 | 0.77 | 70.2 | 0.76 | 0.75 | 65.6 | 0.45 | 0.79 |
| WTCC+STCC | 72.5 | 0.77 | 0.79 | 74.7 | 0.8 | 0.79 | 63.9 | 0.41 | 0.77 |

In a nutshell the experimental results can be summarised as

- Features extracted from S-Transform based time-frequency representation outperformed all the baseline features for voice detection and identification. The possible reason for this improvement may be that these feature capture information related to phonation in a better way as compared to all other time frequency representations (as disused in Section 6.3).

- Sejdic's STCC features showed best performance in terms of classification accuracy, AUC and F1-score of 80.2%, 0.88 and 0.79, respectively for HUPA database in case of voice disorder detection.

- For SVD database standard STCC feature perform best with the classification accuracy of 79.8% for the voice disorder detection experiment.

- S-Transform based cepstral feature also outperformed voice disorder identification task (Exp. 2 and Exp. 3) with classification accuracy of 73.2% and 76%, respectively.

- By performing the fusion of features (STCC with baseline) classification accuracy is improved up to 10% and 4% for voice disorder detection and identification, respectively. This highlights the complementary nature of the extracted feature in discrimination of voice pathology.

### 6.5.5 ANOVA analysis

ANOVA is a statistical test used to compare means among different groups and within each group. Its purpose is to assess whether there are statistically significant differences between the means of two or more groups. F-ratio is the ratio of variations between the groups to variation within group and p represents the probability of observing an F-statistic larger than the computed test-statistic value. When the two groups being compared are similar, the results of ANOVA's F-ratio will be close to 1 and the p-value will be greater than 0.005. Conversely, a higher F-ratio indicates that the two classes being compared are different from each other.

Table 6.10: Result of ANOVA analysis performed on SVD database using the individual feature set.

| Features | No of features shown value of $p<0.005$ |
|---|---|
| MFCC-Residual | 3 |
| MFCC-ZFF | - |
| MFCC | 1 |
| LPCC | 6 |
| CQCC | 5 |
| SFFCC | 13 |
| ZTWCC | 13 |
| STCC | 18 |

ANOVA analysis was conducted, considering healthy and voice disordered subjects as the independent variable and various cepstral features (MFCC-ZFF, MFCC-Residual, MFCC, LPCC, CQCC, SFFCC, ZTWCC, and STCC) as the dependent variables. The results are reported in the Table 6.10. The analysis focused on the first 39 dimensions of the features, which represent the mean of static, delta, and delta-delta coefficients. The results indicate that, for MFCC-ZFF, none of the 39 dimension features exhibited a p-value smaller than 0.005. However, in the case of SFFCC and ZTWCC, 13 features displayed significantly smaller p-values, even below 0.001. Moreover for STCC feature vector 18 features shown the value of p very smaller than 0.0001 and highest value of F-ratio is 51.65. From this analysis it can be concluded that STCC features can capture the acoustic characteristics that distinguish voice disorders from healthy speech, compared to other baseline features which may be due to its better spectro-temporal resolution.

## 6.6 Conclusion

The present work investigates the cepstral features derived from time-frequency analysis method, namely, S-Transform for discrimination of voice disorders from healthy speech. Along with detection, this study also explored assessment of voice disorders from clinical perspective using S-Transform based cepstral features. Cepstral feature extracted from standard S-Transform and it's three other variants are explored for performing the experiments. The performance of detection and assessment system is also compared with baseline features like perturbation features, cepstral features derived from speech signal, excitation source signal and from time-frequency methods (SFF, ZTW and wavelet transform methods). The performance of voice disorder detection and assessment system using STCC features is explored using different classifiers like SVM with different kernels, logistic regression, Naive Bayes, and neural networks. The best classifier from this experiment is selected to perform other experiments in this study. Among all the other classifiers SVM classifier with quadratic kernel performed best for all the S-Transform based features. The experiments are conducted on HUPA and SVD database. From the experiment results it can be verified that performance of voice disorder detection system is better than voice disorder assessment system ; possible reason being healthy speech samples are easily distinguishable from pathological speech samples due to varying acoustic characteristics, whereas distinguishing different voice disorders is difficult as these acoustic characteristics are similar for most of the categories of voice disorders. This study investigated importance of good time-frequency representation for capturing the voice quality information from speech signal. It was observed that good time-frequency representation is important in capturing the voice quality related information which in turn is essential for voice disorder detection and assessment systems. Compared to all the baseline features S-Transform based features work best for both HUPA and SVD database in case of voice disorder detection task. Best classification accuracy of 80.2 % and 79.8 % is achieved using S-Transform based cepstral features for HUPA and SVD database, respectively. Further, combination of S-Transform based features with baseline features improved the performance of voice disorder detection and assessment system which indicates that STCC features contain complementary information related to voice disorders.

Moreover, to assess the effectiveness of the S-Transform in capturing phonation related information from speech signals, we analysed time-frequency representations obtained using the S-Transform method for various phonation types. In this regard, we focused on examining modal phonation as well as four non-modal phonation types: breathy, creaky, harsh, and falsetto phonation. It can be observed that the S-Transform method effectively captures high-frequency harmonics present in breathy speech and accurately represents the period-doubled vibrations associated with creaky speech. Additionally, S-Transform based time-frequency representation efficiently capture the acoustic characteristics of harsh and falsetto phonation. This effectiveness might be due to the better time-frequency localization capabilities offered by the S-Transform when compared to other time-frequency methods.

*Chapter 7*

# Conclusions

This thesis focused on the developing an automated system for detection and identification of voice disorders in a clinical way using the various acoustic features. Detection system determines whether a given speech sample belongs to a healthy subject or a subject with a voice disorder. Identification task focused on knowing the specific category of voice disorder, which can be classified as structural, neurogenic, functional, or psychogenic. All the experiments were conducted on SVD and HUPA datatabase. Multi-class classification approach was explored using the support vector machine (SVM) classifier to perform the experiments. In Experiment 1 (voice disorder detection task), the objective was to differentiate between healthy voice samples and voice disorder samples across all classes. Experiment 2 focused on classifying organic voice disorder samples from non-organic voice disorder samples. Experiment 3 aimed to distinguish between structural voice disorders and neurogenic voice disorders. Lastly, Experiment 4 was conducted to discriminate between functional voice disorders and psychogenic voice disorders.

As the voice disorders are associated with abnormality in anatomy and function of the larynx, excitation source features were explored for detection and identification task in our first study. Excitation source features like intonation features, glottal features, and cepstral coefficients derived from excitation source signal were explored. The aim was to assess the discriminating capabilities of these features in distinguishing between various categories of voice disorders. Excitation source features were compared with state-of-art MFCC, LPCC, and openSMILE features. It was noted that among all the individual excitation source feature sets, intonation features exhibited the highest performance. The classification accuracy achieved using intonation features was 69.3%. for voice disorder detection task. This result suggested that perturbation parameters are more effective in capturing information related to voice disorders. Cepsral features derived from excitation source evidences performed best for experiments 2, 3, and 4 with the classification accuracy of 70.8%, 66.4%, and 64.2%, respectively. From the results it can be concluded that features extracted from the excitation source possess the capability to capture information that can effectively discriminate among various categories of voice disorders. It was also found that, identification task is more challenging than the detection task. This result indicates that this per-

formance degradation may be due to the similar acoustic characteristics shared by different categories of voice disorders.

Computation of various excitation source features like jitter, shimmer, and glottal parameters etc. requires the accurate detection of epoch locations from the speech signal. Therefore in our second study, compared the performance of various state-of-the-art epoch extraction algorithms for speech associated with voice disorders. The performance was also compared for different categories of voice disorders. In this study it was found that, all epoch extraction algorithms shown degradation in their performance for speech associated with voice disorders compared to healthy speech. It may be due to the reason that for the subjects suffering with voice disorder, variation of F0 is more in single utterance as compared to healthy subjects. Moreover, some state-of-the-art epoch extraction methods depend on average value of F0 for the detection of epoch locations, hence we applied the region-wise approach as pre-processing step for the calculation of epoch locations. The performance of state-of-the-art epoch extraction method was also compared with and without applying region-wise approach. From the results, it was observed that performance is improved for some of the algorithms up to 2% after applying the region-wise approach for most of the categories of voice disorder. Further, excitation source features are extracted by applying the region-based processing to the state-of-the-art epoch extraction algorithm for building the voice disorder detection and identification system. The performance of the system is also compared with the excitation source features derived without applying the region-based approach to the state-of-the-art epoch extraction algorithms. From the results of this study showed improvement in the performance for both detection and identification system, which concluded that accurately identifying epoch locations plays crucial role.

Degradation in the voice quality is one of the important characteristic used by SLPs for the assessment of voice disorders. The long term average spectrum features which captures the voice quality were also explored for the identification and detection of voice disorders. Four state-of-the-art filter banks designed with critical-band, constant-Q, gammatone, and single-frequency filtering approaches were used for the extraction of features. Moreover, the performance of the systems is compared with state-of-the-art statistical-average and openSMILE features. Voice disorder detection experiment was carried out on SVD and HUPA database, while only SVD database is used for identification task. Identification task is performed in clinical way, in which four binary classifiers were trained in our study. From the results, it was observed that constant-Q filter bank based LTAS features performed better among all LTAS features with classification accuracy of 78% and 81.4% for voice disorder detection task on SVD and HUPA database, respectively.

From the previous studies it was found that voice disorders and voice quality information can be captured in a better way in time-frequency domain as compared to the individual excitation source and system feature. Stockwell-Transform (S-Transform) provides good time-frequency localization; hence, it may efficiently capture the voice disorder related information from speech signal. With this motivation, we investigated different variants of S-Transform for the classification of voice disorders. We also proposed the S-Transform based cepstral coefficients for voice disorder detection and identifica-

tion. The performance of the proposed feature was compared with baseline features on SVD and HUPA databases. As compared to baseline features, proposed features performed best in terms of classification accuracy of 80.2% and 79.8% on HUPA and SVD databases, respectively, for voice disorder detection task. Also, the proposed features performed better in case of assessment task. Further, the experimental results reveal that the combination of cepstral coefficients derived from S-Transform with baseline features improved the performance of proposed systems by 8% and 4% for detection and identification task, respectively. This in turn, indicates the complementary nature of the proposed features in classification of voice disorders.

## 7.1 Future scope

The future scope of this thesis can be summarised as:

- ***Analysis of phase spectrum***: From the literature, it was found that the phase spectrum of speech signal captures information about voice quality. Incorporating the phase spectrum along with the magnitude spectrum, which provides complete information about the speech, might improve automatic detection and identification of voice pathology. Due to the importance of the phase component for the analysis of voice quality, it is planned to analyse the phase-based feature for voice disorder to understand its significance.

- ***Analysis of loudness***: Loudness is a crucial perceptual characteristic that SLPs rely on when assessing voice disorders. Individuals with voice disorders often struggle to produce sounds at a sufficient volume compared to those with healthy voices. Therefore, it is important to explore the features that can effectively capture information related to loudness in speech signals. Moreover, research should focus on analyzing the impact of loudness on different categories of voice disorders.

- ***Deep neural network architecture***: This thesis explored the various machine learning architecture for building the voice disorder detection system. Various deep learning neural network architectures are explored in the literature. Therefore the role of deep neural network can be explored in the detection and identification of voice disorders.

- ***Impact of gender information and age related information***: It is reasonable to consider that SLPs may exhibit bias towards certain speaker characteristics, such as gender or age, when making decisions about the presence of voice disorders. Therefore, it is important to investigate the influence of gender and age during the development of pathological detection systems.

- ***Cross database experiments***: In this thesis the performance of voice disorder detection and identification system was explored on individual dataset such as SVD and HUPA. There is necessity to assess the performance of the system across various datasets, wherein the systems are trained on one database and their performance is evaluated on other databases.

# Related Publications

**Conferences:**

1. **Barche Purva**, Krishna Gurugubelli, and Anil Kumar Vuppala. "Towards Automatic Assessment of Voice Disorders: A Clinical Approach." In INTERSPEECH, pp. 2537-2541. 2020.

2. **Barche Purva**, Krishna Gurugubelli, and Anil Kumar Vuppala. "Comparative study of different epoch extraction methods for speech associated with voice disorders." In ICASSP 2021-2021, pp. 6923-6927. IEEE 2021.

3. **Barche Purva**, Krishna Gurugubelli, and Anil Kumar Vuppala. "Comparative Study of Filter Banks to Improve the Performance of Voice Disorder Assessment Systems using LTAS Features." In 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 737-742. IEEE, 2021.

**Journal:**

1. **Barche Purva**, Krishna Gurugubelli, and Anil Kumar Vuppala. "Stockwell-Transform Based Feature Representation for Detection and Assessment of Voice Disorders." Accepted at international Journal of Speech Technology 2024.

2. **Barche Purva**, Krishna Gurugubelli, and Anil Kumar Vuppala. "Epoch extraction in real-world scenario." Submitted to Computer Speech & Language 2024.

**List of other publications**:

1. Vats Nayan Anand, **Barche Purva**, Ganesh S. Mirishkar, and Anil Kumar Vuppala. "Exploring High Spectro-Temporal Resolution for Alzheimer's Dementia Detection." In 2022 IEEE International Conference on Signal Processing and Communications (SPCOM), pp. 1-5. IEEE, 2022.

2. Syed Abdul Jabbar, **Barche Purva**, Gurugubelli Krishna, Syed AZEEMUDDIN, and Anil Kumar Vuppala, “Implementation of Zero-Phase Zero Frequency Resonator Algorithm on FPGA”, ACM IC3, Noida, 2022.

3. Monica Ponnam, **Barche Purva**, and Anil Kumar Vuppala, ”Automatic Detection of Parkinson’s Disease using Zero-Time Window based cepstral Feature”, IEEE INDICON, 2023.

4. Syed Abdul Jabbar, **Barche Purva**, Gurugubelli Krishna, Syed AZEEMUDDIN, and Anil Kumar Vuppala,” Stable Implementation of Voice Activity Detector Using Zero-Phase Zero Frequency Resonator on FPGA”, IEEE RTC 2023, USA.

5. Vamshi raghu simha narasinga, Hina Fathima, Kowshik Motepalli, Sangeetha Mahesh, Sai Akarsh C, **Purva Barche**, Sai Ganesh Mirishkar, Ajish Abraham and Anil Vuppala, “Enhancing Stuttering Detection: A Syllable-Level Stutter Dataset,” Accepted at IEEE International Conference on Signal Processing and Communications (SPCOM) 2024.

# Appendix A

This chapter discusses detail description of some of the baseline feature used in our study. Section A.1 describes the intonation feature used in this study. Section A.2 discusses the detail about openSMILE feature set. In our study, we used Computational Paralinguistics Challenge features set (ComParE) and extended Geneva Minimalistic Acoustic Parameter Set(eGeMAPS) feature sets as baseline features from the openSMILE feature set.

## A.1:Intonation feature set

The intonation feature set consists of 76-dimensional features used to model phonation-related characteristics of the speech signal [25]. These features are derived from evidence of the excitation source signal and include the fundamental frequency of vibration of vocal folds, jitter, shimmer, harmonic-to-noise ratio, and its variants. Zero frequency filtering (ZFF) method [56] is used to compute the epoch locations from the excitation source signal, which in turn used to derive the intonation features.

In ZFF method, speech signal is passed twice through zero frequency resonator. Zero frequency resonator is low pass filter, which attenuates the higher-order harmonics corresponding to the vocal tract system. Filtering the signal results in an output that grows or decay rapidly. This is because the time-domain equivalent of zero frequency resonator functions as an integrator. This trends in output is removed by passing the signal through trend removal filter. The resulting mean subtracted signal is referred to as zero frequency filtered signal. Negative to positive crossing of the ZFF signal corresponds to the epoch locations.

Figure A.1 depicts input speech signal to ZFF method, its corresponding ground truth, and output derived from ZFF method. The fundamental frequency of vocal fold vibration (F0), along with the strength and energy of excitation features, are derived from the epoch locations.

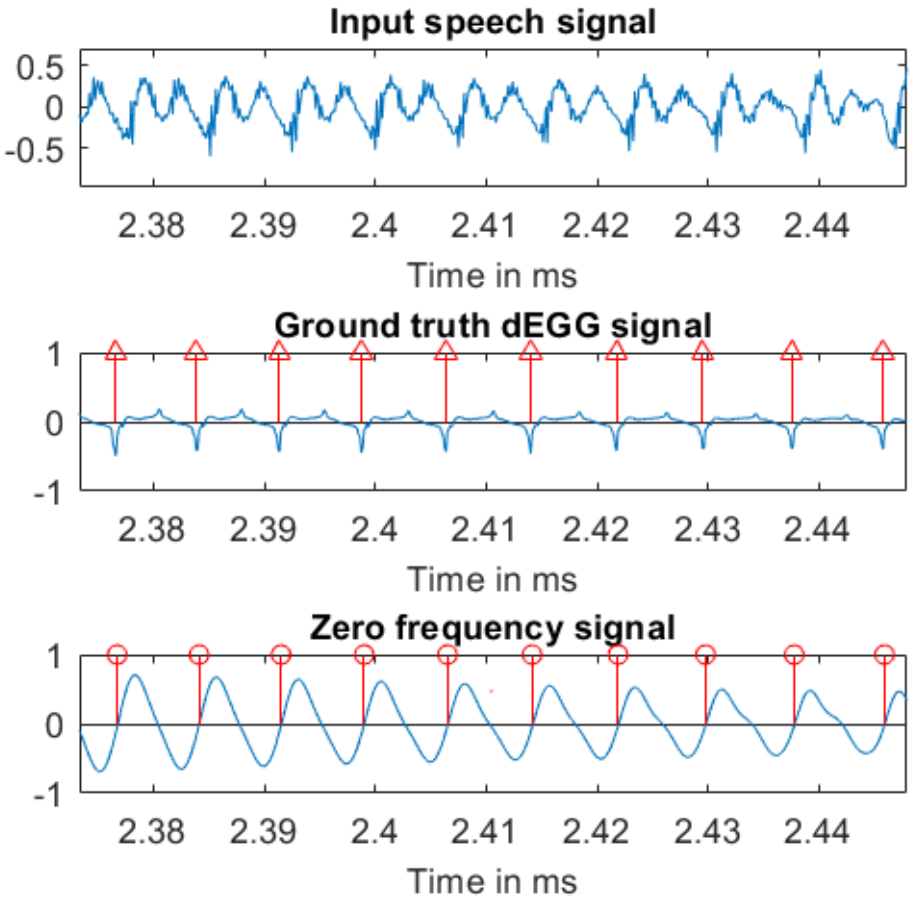



Figure A.1: Epoch extraction from ZFF method. (a) Speech signal. (b) Derivative of EGG signal. (c) ZFF signal and corresponding epoch locations.

- Fundamental frequency (F0): F0 is determined by calculating the epoch location derived from the ZFF method. The difference between the consecutive epoch location gives the measure of pitch period (T0) and the inverse of pitch period is fundamental frequency denoted by F0 [162, 141]. If $\mathcal{E}$= $\{e_1, e_2, e_3.....e_M\}$ is the number of GCI locations derived from the ZFF method, then $F_0$ is given by

$$F_0[n] = \frac{1}{T_0(n)} = \frac{f_s}{e_n - e_{n-1}}, n = 2, 3, .....M \tag{1}$$

  where $T_0[n]$ is the fundamental period of vocal fold vibration.

- Strength of excitation (SoE): The slope of ZFF signal around each epoch location is referred to as the strength of excitation which indicates the strength or intensity of GCI location. It is directly proportional to the rate at which the vocal folds close during phonation [141].

$$SoE = y[e_n + 1] - y[e_n - 1] \tag{2}$$

  where $y[n]$ is the output signal of the ZFF method.

- Energy of excitation (EoE) of ZFF signal: The mean square energy of the samples at GCI locations is defined as the energy of excitation, which gives the measure of vocal effort.

$$EoE = \sum_{i=-L/2}^{L/2} y^2[n + i] \tag{3}$$

  where y[n] is the ZFF signal, and L is the length of the window over which the energy is computed. L is taken as 10 ms for the calculation of energy.

Table A.1: Intonation feature and corresponding feature dimension [69].

| Feature | Dimension |
|---|---|
| Statistical measures of F0 | 5 |
| Jitter quotients of F0 | 22 |
| Shimmer quotients of strength of excitation (SOE) | 22 |
| Shimmer quotients of Energy of excitation (EOE) | 22 |
| Harmonic to noise ratio and noise to harmonic ratio | 4 |
| Pitch perturbation entropy (PPE) | 1 |

The detail about each feature is given as following:

1. **Statistical measures of F0** : Mean, median, standard deviation, minima and maxima are the 5 statistical measure considered as subset of intonation features.

2. **Jitter quotient**: It is a cycle-to-cycle perturbation of the glottal cycle and is derived from the pitch contour ($T_0$) or the fundamental frequency contour (F0). It contains 22 dimension feature vector, which is discussed below.

   (a) **Mean-absolute-Difference of successive cycle**: It is defined cycle to cycle variations of the fundamental frequency.

   $$Jitter_{F0,abs} = \frac{1}{N}\sum_{i=0}^{N-1}|F_i - F_{i+1}| \tag{4}$$

   Where $N$ is number of $F0$ extracted periods.

   (b) **Ration of mean absolute difference and mean of** $F0$: It is defined as ratio of mean-absolute-difference of successive cycle to mean of $F0$ and is expressed in percentage.

   $$Jitter_{F0,\%} = 100\frac{\frac{1}{N}\sum_{i=0}^{N-1}|F_i - F_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}F_{0,i}} \tag{5}$$

   (c) **Perturbation quotient measures using 3 cycles**: It is defined as absolute of difference between one fundamental frequency and average of the fundamental frequency with its two neighbours, divided by average fundamental frequency. Three different variants of this parameter are considered [30].

   $$Jitter_{F0,\%} = \frac{\frac{1}{N-1}\sum_{i=0}^{N-1}|F_i - \frac{1}{3}\sum_{n=i-1}^{i+1}F_n|}{\frac{1}{N}\sum_{i=1}^{N}F_i} \tag{6}$$

   (d) **Perturbation quotient measures using 5 cycles**: It is defined as absolute of difference between one fundamental frequency and average of the fundamental frequency with its four neighbours i.e. two previous and two subsequent periods, divided by average fundamental frequency. Three different variants of this parameter are considered [27].

   $$Jitter_{F0,\%} = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}|F_i - (\frac{1}{5}\sum_{n=i-2}^{i+2}F_n)|}{\frac{1}{N}\sum_{i=1}^{N}F_i} \tag{7}$$

(e) **Perturbation quotient measures using 11 cycles**: It is defined as absolute of difference between one fundamental frequency and average of the fundamental frequency with its 10 neighbours i.e. five previous and five subsequent periods, divided by average fundamental frequency. Three different variants of this parameter are considered.

$$Jitter_{F0,\%} = \frac{\frac{1}{N-1}\sum_{i=5}^{N-5}|F_i - \frac{1}{11}\sum_{n=i-5}^{i+5}F_n|}{\frac{1}{N}\sum_{i=1}^{N-1}F_i} \tag{8}$$

(f) **Zeroth order perturbation**: It is defines as

$$Jitter_{F_{0,p1}} = \frac{1}{N}\sum_{i=1}^{N-1}|F_{0,i} - \frac{1}{N}\sum_{j=1}^{N}F_{0,j}| \tag{9}$$

(g) **Jitter in dB**: It is defined as average absolute difference between two consecutive fundamental frequency in logarithm.

$$Jitter(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log\frac{A_{i+1}}{A_i}| \tag{10}$$

(h) **Frequency modulation (FM)**: It is defined as the ratio of the difference between the maximum and minimum values of the mean fundamental frequency to their sum. If $F0$ is a vector of length $N$ containing all the fundamental periods for a given speech signal, then:

$$FM = \frac{max(F_0) - min(F_0)}{max(F_0) + min(F_0)} \tag{11}$$

(i) **Parameters derived from Teager energy operator**: Fundamental frequency contour was also calculated by using Teager-Kaiser energy operator (TKEO). From this contour mean, standard deviation, and the $5^{th}$, $25^{th}$ $75^{th}$ and $95^{th}$ percentile values were computed [231].

3. **Shimmer quotient**: Shimmer is defined as cycle to cycle variation in amplitude between consecutive cycles of the glottal flow waveform [26]. From this amplitude, 22 shimmer quotients are derived by applying the same formulas used for jitter quotients, but replacing F0 with A.

4. **Harmonic to noise ratio (HNR) and Noise to harmonic ratio (NHR)**: HNR is defined as ratio of energy between the harmonic or periodic component extracted from the speech signal to noise or aperiodic or noise component calculated from speech signal. NHR is reverse of HNR. Two statistics measures, mean and standard deviation derived from HNR and NHR are used to represent a four-dimensional feature set [26].

5. **Pitch perturbation entropy (PPE)**:It is deviation from periodicity derived from the entropy and measures the impaired control of stable pitch during sustained vowel [231].

$$PPE = \frac{\sum_{i}^{L_{PPE}} p(i) \ln(p(i))}{\ln(L_{PPE})} \tag{12}$$

   Where $p(i)$ is Discrete probability distribution of logarithm value of pitch period and $L_{PPE}$ length of points used to calculate pitch perturbation or spread measure.

## A.2: OpenSMILE feature set

Open-source Speech and Music Interpretation by Large-space Extraction (OpenSMILE ) is open source tool used for extraction of acoustic features from speech signal and classify of music and speech signals [164]. This toolkit is capable of extracting low-level description (such as energy, loudness, pitch,voice quality, mel-spectrum, etc.) and applying various filter and functional to these descriptor. In our study two sets from openSMILE tool are used.

**A.2.1: Computational Paralinguistics Challenge (ComParE)features set**: The 2013 Interspeech ComParE features set is large-scale high-dimension brute-forced acoustic feature set contains 6373 static features resulting from the computation of various functional over low-level descriptor (LLD) contours [165]. The low-level descriptors cover a broad set of descrip-

Table A.2: ComParE acoustic feature set: 65 provided low-level descriptors(LLD)

| **4 energy related LLD** | **Group** |
|---|---|
| Sum of auditory spectrum (loudness) | prosodic |
| Sum of RASTA-filtered auditory spectrum | prosodic |
| RMS Energy, Zero-Crossing Rate | prosodic |
| **55 spectral LLD** | **Group** |
| RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz) | spectral |
| MFCC 1–14 | cepstral |
| Spectral energy 250–650 Hz, 1 k–4 kHz | spectral |
| Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9 | spectral |
| Spectral Flux, Centroid, Entropy, Slope | spectral |
| Psychoacoustic Sharpness, Harmonicity | spectral |
| Spectral Variance, Skewness, Kurtosis | spectral |
| **6 voicing related LLD** | **Group** |
| $F_0$ (SHS & Viterbi smoothing) | prosodic |
| Prob. of voicing | voice qual. |
| log. HNR, Jitter (local & $\delta$), Shimmer (local) | voice qual. |

tors (features) from the fields of speech processing, music information retrieval and general sound analysis. LLDs are feature which are related to low level description of audio information like temporal, spectrum related, voice quality related features. Supra-segmental features are calculated by applying a large set of statistical functional to acoustic LLD. There are 4 energy related parameter (like zero crossing rate, RMS energy, loudness), 55 spectral features (MFCC, spectral energy, Spectral variance, skewness, kurtosis) and 6 voicing related features (Jitter, Shimmer, HNR). The statistical functionals applied to the LLD include the mean, stan-

dard deviation, percentiles and quartiles, linear regression functionals, quadratic regression and minima/maxima related functionals.

Table A.3: Functionals applied to ComParE Feature set [1]: arithmatic mean of LLD [2]: not applied to voicing related LLD except F0 [3]: only applied to F0

| **Functionals applied to LLD / Δ LLD** | **Group** |
|---|---|
| quartiles 1–3, 3 inter-quartile ranges | percentiles |
| 1 % percentile (≈ min), 99 % pctl. (≈ max) | percentiles |
| percentile range 1 %–99 % | percentiles |
| position of min / max, range (max – min) | temporal |
| arithmetic mean[1], root quadratic mean | moments |
| contour centroid, flatness | temporal |
| standard deviation, skewness, kurtosis | moments |
| rel. dur. LLD is above 25 / 50 / 75 / 90 % range | temporal |
| relative duration LLD is rising | temporal |
| rel. duration LLD has positive curvature | temporal |
| gain of linear prediction (LP), LP Coeff. 1–5 | modulation |
| mean, max, min, std. dev. of segment length[2] | temporal |
| **Functionals applied to LLD only** | **Group** |
| mean value of peaks | peaks |
| mean value of peaks – arithmetic mean | peaks |
| mean / std.dev. of inter peak distances | peaks |
| amplitude mean of peaks, of minima | peaks |
| amplitude range of peaks | peaks |
| mean / std. dev. of rising / falling slopes | peaks |
| linear regression slope, offset, quadratic error | regression |
| quadratic regression a, b, offset, quadratic err. | regression |
| percentage of non-zero frames[3] | temporal |

**A.2.2: extended Geneva Minimalistic Acoustic Parameter Set(eGeMAPS)**: eGeMAPS are small-scale (low-dimension) knowledge-based acoustic feature set contains 88 parameters,these feature set is also designed to extract paralinguistic information from speech with small feature set compared to ComParE feature set (6373 features) [166]. Functionals are applied to 45 LLD. Frequency related parameter are total of (12) pitch, jitter, first three formant frequency and bandwidth of first formant their mean and standard deviations. In total, it consist of 42 LLD on which two statistical functionals (arithmetic mean and coefficient of variations) is applied makes total of 88 parameters.

Table A.4: eGeMAPS acoustic feature set: 42 provided low-level descriptors(LLD)

| **1 energy related LLD** | **Group** |
|---|---|
| Sum of auditory spectrum (loudness) | Prosodic |
| **25 spectral LLD** | **Group** |
| $\alpha$ ratio (50–1 000 Hz / 1-5 k Hz) | Spectral |
| Energy slope (0–500 Hz, 0.5–1.5 k Hz) | Spectral |
| Hammarberg index | Spectral |
| MFCC 1–4 | Cepstral |
| Spectral Flux | Spectral |
| **6 voicing related LLD** | **Group** |
| F0 (Linear & semi-tone) | Prosodic |
| Formants 1, 2, (freq., bandwidth, ampl.) | Voice Quality |
| Harmonic difference H1–H2, H1–A3 | Voice Quality |
| log. HNR, Jitter (local), Shimmer (local) | Voice Quality |

# Bibliography


[1] I. R. Titze and D. W. Martin, "Principles of voice production," 1998.

[2] L. R. Rabiner and R. W. Schafer, *Introduction to digital speech processing*. Now Publishers Inc, 2007.

[3] K. Johnson, *Acoustic and auditory phonetics*. John Wiley & Sons, 2011.

[4] M. Rothenberg, "Acoustic interaction between the glottal source and the vocal tract," *Vocal fold physiology*, vol. 1, pp. 305–323, 1981.

[5] J. L. Flanagan, "Some properties of the glottal sound source," *Journal of Speech and Hearing Research*, vol. 1, no. 2, pp. 99–116, 1958.

[6] N. R. Williams, "Occupational groups at risk of voice disorders: a review of the literature," *Occupational medicine*, vol. 53, no. 7, pp. 456–460, 2003.

[7] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. F. Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2017.

[8] R. T. Sataloff, *Clinical assessment of voice*. Plural Pub Incorporated, 2005.

[9] P. W. Flint, B. H. Haughey, K. T. Robbins, J. R. Thomas, J. K. Niparko, V. J. Lund, and M. M. Lesperance, *Cummings otolaryngology-head and neck surgery e-book*. Elsevier Health Sciences, 2014.

[10] K. Omori, "Diagnosis of voice disorders," *JMAJ*, vol. 54, no. 4, pp. 248–253, 2011.

[11] A. E. Aronson, "Clinical voice disorders," *An interdisciplinary approach*, 1985.

[12] M. Frohlich, D. Michaelis, and H. W. Strube, "Acoustic" breathiness measures" in the description of pathologic voices," in *Proc. ICASSP*, vol. 2. IEEE, 1998, pp. 937–940.

[13] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[14] L. Eskenazi, D. G. Childers, and D. M. Hicks, "Acoustic correlates of vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 298–306, 1990.

[15] S. B. Davis, "Acoustic characteristics of normal and pathological voices," in *Speech and language*. Elsevier, 1979, vol. 1, pp. 271–335.

[16] M. Huckvale and C. Buciuleac, "Automated detection of voice disorder in the saarbrücken voice database: Effects of pathology subset and audio materials," in *Proc. INTERSPEECH*, 2021, pp. 4850–4854.

[17] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," *in: Proc. INTERSPEECH 2018,*, pp. pp. 446–450.

[18] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. Khanapi Abd Ghani, M. S. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, and F. T. Al-Dhief, "Voice pathology detection and classification using convolutional neural network model," *Applied Sciences*, vol. 10, no. 11, p. 3723, 2020.

[19] K. Ezzine and M. Frikha, "Investigation of glottal flow parameters for voice pathology detection on SVD and MEEI databases," in *2018 4th International conference on advanced technologies for signal and image processing (ATSIP)*. IEEE, 2018, pp. 1–6.

[20] V. Gupta, "Voice disorder detection using long short term memory (LSTM) model," *ArXiv*, vol. 1812.01779, 2018.

[21] S. A. Syed, M. Rashid, S. Hussain, and H. Zahid, "Comparative analysis of CNN and RNN for voice pathology detection," *BioMed Research International*, vol. Vol.2021,2021, 2021.

[22] N. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67 745–67 755, 2020.

[23] V. Berisha, C. Krantsevich, G. Stegmann, S. Hahn, and J. Liss, "Are reported accuracies in the clinical speech machine learning literature overoptimistic?" in *Proc. INTERSPEECH*, 2022, pp. 2453–2457.

[24] J. Laver, S. Hiller, and J. M. Beck, "Acoustic waveform perturbations and voice disorders," *Journal of Voice*, vol. 6, no. 2, pp. 115–126, 1992.

[25] N. Adiga, C. Vikram, K. Pullela, and S. M. Prasanna, "Zero frequency filter based analysis of voice disorders." in *Interspeech*, 2017, pp. 1824–1828.

[26] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis–jitter, shimmer and hnr parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.

[27] F. Klingholz and F. Martin, "Quantitative spectral evaluation of shimmer and jitter," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 2, pp. 169–174, 1985.

[28] A. Al-Nasheri, Z. Ali, G. Muhammad, and M. Alsulaiman, "An investigation of mdvp parameters for voice pathology detection on three different databases," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[29] L. W. Lopes, J. D. da Silva, L. B. Simões, D. da Silva Evangelista, P. O. C. Silva, A. A. Almeida, and M. F. B. de Lima-Silva, "Relationship between acoustic measurements and self-evaluation in patients with voice disorders," *Journal of Voice*, vol. 31, no. 1, pp. 119–e1, 2017.

[30] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Sig. Process.*, vol. 2009, pp. 1–9, 2009.

[31] Y. Maryn, P. Corthals, M. De Bodt, P. Van Cauwenberge, and D. Deliyski, "Perturbation measures of voice: a comparative study between multi-dimensional voice program and praat," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 4, pp. 217–226, 2009.

[32] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, and G. S. Berke, "Comparison of voice analysis systems for perturbation measurement," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 1, pp. 126–134, 1996.

[33] G. d. Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[34] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 537–543, 1997.

[35] J.-W. Lee, S. Kim, and H.-G. Kang, "Detecting pathological speech using contour modeling of harmonic-to-noise ratio," in *Proc. ICASSP*. IEEE, 2014, pp. 5969–5973.

[36] T. Drugman, T. Dubuisson, and T. Dutoit, "On the mutual information between source and filter contributions for voice pathology detection," in *Proc. Interspeech 2009*, 2009, pp. 1463–1466.

[37] F. Klingholtz, "Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2218–2224, 1990.

[38] Y. Qi, R. E. Hillman, and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *The Journal of the Acoustical Society of America*, vol. 105, no. 4, pp. 2532–2535, 1999.

[39] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of speech, language, and hearing research*, vol. 43, no. 2, pp. 469–485, 2000.

[40] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 367–379, 2020.

[41] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Towards automatic assessment of voice disorders: A clinical approach," *Proc. Interspeech 2020*, pp. 2537–2541, 2020.

[42] M. Kohler, M. M. Vellasco, E. Cataldo, *et al.*, "Analysis and classification of voice pathologies using glottal signal parameters," *Journal of Voice*, vol. 30, no. 5, pp. 549–556, 2016.

[43] C. R. Watts and S. N. Awan, "Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts," *Journal of Speech, Language, and Hearing Research*, 2011.

[44] B. R. Kumar, J. S. Bhat, and N. Prasad, "Cepstral analysis of voice in persons with vocal nodules," *Journal of Voice*, vol. 24, no. 6, pp. 651–653, 2010.

[45] Y. D. Heman-Ackah, D. D. Michael, and G. S. Goding Jr, "The relationship between cepstral peak prominence and selected parameters of dysphonia," *Journal of Voice*, vol. 16, no. 1, pp. 20–27, 2002.

[46] R. K. Balasubramanium, J. S. Bhat, S. Fahim III, and R. Raju III, "Cepstral analysis of voice in unilateral adductor vocal fold palsy," *Journal of voice*, vol. 25, no. 3, pp. 326–329, 2011.

[47] Y. D. Heman-Ackah, D. D. Michael, M. M. Baroody, R. Ostrowski, J. Hillenbrand, R. J. Heuer, M. Horman, and R. T. Sataloff, "Cepstral peak prominence: a more reliable measure of dysphonia," *Annals of Otology, Rhinology & Laryngology*, vol. 112, no. 4, pp. 324–333, 2003.

[48] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, 2014.

[49] J. I. Godino-Llorente, S. Aguilera-Navarro, and P. Gómez-Vilda, "Lpc, lpcc and mfcc parameterisation applied to the detection of voice impairments," in *Sixth International Conference on Spoken Language Processing*, 2000, pp. 965–968.

[50] F. Javanmardi, S. R. Kadiri, M. Kodali, P. Alku, *et al.*, "Comparing 1-dimensional and 2-dimensional spectral feature representations in voice pathology detection using machine learning and deep learning classifiers," in *Interspeech*. International Speech Communication Association, 2022.

[51] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[52] A. K. Dubey, S. M. Prasanna, and S. Dandapat, "Hypernasality severity detection using constant q cepstral coefficients." in *INTERSPEECH*, 2019, pp. 4554–4558.

[53] M. Jičínskỳ and J. Mareš, "Measurable changes of voice after voice disorder treatment," in *Proceedings of the Computational Methods in Systems and Software*. Springer, 2019, pp. 295–305.

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

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

[56] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, 2008.

[57] K. Gurugubelli and A. K. Vuppala, "Stable implementation of zero frequency filtering of speech signals for efficient epoch extraction," *IEEE Sig. Process. Lett.*, vol. 26, no. 9, pp. 1310–1314, 2019.

[58] K. S. Srinivas and K. Prahallad, "An fir implementation of zero frequency filtering of speech signals," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 9, pp. 2613–2617, 2012.

[59] P. Gangamohan and B. Yegnanarayana, "A robust and alternative approach to zero frequency filtering method for epoch extraction." in *INTERSPEECH*, 2017, pp. 2297–2300.

[60] D. Veeneman and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 369–377, 1985.

[61] A. Paavo, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[62] P. Alku, "Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[63] 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*, vol. 22, no. 3, pp. 596–607, 2013.

[64] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[65] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[66] T. Leino, "Long-term average spectrum in screening of voice quality in speech: untrained male university students," *J. of Voice*, vol. 23, no. 6, pp. 671–676, 2009.

[67] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (ltas)," *Journal of voice*, vol. 10, no. 1, pp. 59–66, 1996.

[68] D. Sergeant and G. F. Welch, "Age-related changes in long-term average spectra of children's voices," *J. of Voice*, vol. 22, no. 6, pp. 658–670, 2008.

[69] M. H. Javid, K. Gurugubelli, and A. K. Vuppala, "Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech," in *Proc. ICASSP*. IEEE, 2020, pp. 6754–6758.

[70] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2002.

[71] J. A. Seikel, D. G. Drumright, and D. W. King, *Anatomy & physiology for speech, language, and hearing*. Cengage Learning, 2015.

[72] G. P. Moore, "Observations on laryngeal disease, laryngeal behavior and voice," *Annals of Otology, Rhinology & Laryngology*, vol. 85, no. 5, pp. 553–564, 1976.

[73] H. Hirose, "Laryngeal adjustments in consonant production," *Phonetica*, vol. 34, no. 4, pp. 289–294, 1977.

[74] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[75] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London*, vol. 31, pp. 1–186, 1980.

[76] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia phoniatrica et logopaedica*, vol. 48, no. 5, pp. 240–254, 1996.

[77] C. T. Ishi, K.-I. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 47–56, 2007.

[78] D. N. Gowda and M. Kurimo, "Analysis of breathy, modal and pressed phonation based on low frequency spectral density." in *INTERSPEECH*, 2013, pp. 3206–3210.

[79] Y. Swerdlin, J. Smith, and J. Wolfe, "The effect of whisper and creak vocal mechanisms on vocal tract resonances," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2590–2598, 2010.

[80] D. P. Cantwell and L. Baker, *Developmental speech & language disorders.* Guilford Press, 1987.

[81] W. Lanier, *Speech disorders*. Greenhaven Publishing LLC, 2010.

[82] R. W. Bastian and J. P. Thomas, "Do talkativeness and vocal loudness correlate with laryngeal pathology? a study of the vocal overdoer/underdoer continuum," *Journal of Voice*, vol. 30, no. 5, pp. 557–562, 2016.

[83] J. Baker, D. I. Ben-Tovim, A. Butcher, A. Esterman, and K. McLaughlin, "Development of a modified diagnostic classification system for voice disorders with inter-rater reliability study," *Logopedics Phoniatrics Vocology*, vol. 32, no. 3, pp. 99–112, 2007.

[84] C. L. Ludlow, "Spasmodic dysphonia: a laryngeal control disorder specific to speech," *Journal of neuroscience*, vol. 31, no. 3, pp. 793–797, 2011.

[85] M. E. Freeman and M. E. Fawcus, *Voice disorders and their management*. Whurr Publishers, 2000.

[86] J. Baker, "Functional voice disorders: clinical presentations and differential diagnosis," in *Handbook of clinical neurology*. Elsevier, 2016, vol. 139, pp. 389–405.

[87] E. Seifert, "Stress and distress in non-organic voice disorder," *Swiss medical weekly*, vol. 135, no. 2728, 2005.

[88] L. Flood, "Laryngeal function and voice disorders: Basic science to clinical practice cr watts," *The Journal of Laryngology & Otology*, vol. 133, no. 9, pp. 833–833, 2019.

[89] M. Fawcus, *Voice disorders and their management*. Springer, 2013.

[90] A. L. Rosenthal, S. Y. Lowell, and R. H. Colton, "Aerodynamic and acoustic features of vocal effort," *Journal of Voice*, vol. 28, no. 2, pp. 144–153, 2014.

[91] E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *The Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 511–529, 1988.

[92] J. R. Smitheran and T. J. Hixon, "A clinical method for estimating laryngeal airway resistance during vowel production," *Journal of Speech and Hearing Disorders*, vol. 46, no. 2, pp. 138–146, 1981.

[93] E. U. Grillo, K. Perta, and L. Smith, "Laryngeal resistance distinguished pressed, normal, and breathy voice in vocally untrained females," *Logopedics Phoniatrics Vocology*, vol. 34, no. 1, pp. 43–48, 2009.

[94] B. Barsties and M. De Bodt, "Assessment of voice quality: current state-of-the-art," *Auris Nasus Larynx*, vol. 42, no. 3, pp. 183–188, 2015.

[95] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 1, pp. 21–40, 1993.

[96] J. Oates, "Auditory-perceptual evaluation of disordered voice quality," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[97] M. Hirano, "Grbas" scale for evaluating the hoarse voice & frequency range of phonation," *Clinical examination of voice*, vol. 5, pp. 83–84, 1981.

[98] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, and V. Woisard, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques," *European Archives of Oto-rhino-laryngology*, vol. 258, no. 2, pp. 77–82, 2001.

[99] D. D. Mehta and R. E. Hillman, "Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods," *Current opinion in otolaryngology & head and neck surgery*, vol. 16, no. 3, p. 211, 2008.

[100] B. H. Jacobson, A. Johnson, C. Grywalski, A. Silbergleit, G. Jacobson, M. S. Benninger, and C. W. Newman, "The voice handicap index (vhi) development and validation," *American Journal of Speech-Language Pathology*, vol. 6, no. 3, pp. 66–70, 1997.

[101] W. E. Halawa, S. S. Perez, and C. G. Antonio, "Measurement of vocal handicap in patients with vocal nodules and functional dysphonias," *Egyptian Journal of Ear, Nose, Throat and Allied Sciences*, vol. 12, no. 2, pp. 121–124, 2011.

[102] F. B. Madeira and S. Tomita, "Voice handicap index evaluation in patients with moderate to profound bilateral sensorineural hearing loss," *Brazilian journal of otorhinolaryngology*, vol. 76, no. 1, pp. 59–70, 2010.

[103] N. D. Hogikyan and G. Sethuraman, "Validation of an instrument to measure voice-related quality of life (v-rqol)," *Journal of voice*, vol. 13, no. 4, pp. 557–569, 1999.

[104] C. A. Rosen and T. Murry, "Diagnostic laryngeal endoscopy," *Otolaryngologic Clinics of North America*, vol. 33, no. 4, pp. 751–757, 2000.

[105] S. D. Rajput and M. J. Poriya, "Stroboscopy: an evolving tool for voice analysis in vocal cord pathologies," *Int J Otorhinolaryngol Head Neck Surg*, vol. 3, no. 04, pp. 927–931, 2017.

[106] V. Reynolds, A. Buckland, J. Bailey, J. Lipscombe, E. Nathan, S. Vijayasekaran, R. Kelly, Y. Maryn, and N. French, "Objective assessment of pediatric voice disorders with the acoustic voice quality index," *Journal of Voice*, vol. 26, no. 5, pp. 672.e1–672.e7, 2012.

[107] B.-F. Zaidi, M. Boudraa, S.-A. Selouani, D. Addou, and M. S. Yakoub, "Automatic recognition system for dysarthric speech based on mfcc's, pncc's, jitter and shimmer coefficients," in *Science and Information Conference*. Springer, 2019, pp. 500–510.

[108] M. Vasilakis and Y. Stylianou, "Voice pathology detection based eon short-term jitter estimations in running speech," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 153–170, 2009.

[109] N. B. Pinto and I. R. Titze, "Unification of perturbation measures in speech signals," *The Journal of the Acoustical Society of America*, vol. 87, no. 3, pp. 1278–1289, 1990.

[110] P. Lieberman, "Perturbations in vocal pitch," *The Journal of the Acoustical Society of America*, vol. 33, no. 5, pp. 597–603, 1961.

[111] Y. Horii, "Vocal shimmer in sustained phonation," *Journal of Speech, Language, and Hearing Research*, vol. 23, no. 1, pp. 202–209, 1980.

[112] V. L. Heiberger and Y. Horii, "Jitter and shimmer in sustained phonation," in *Speech and language*. Elsevier, 1982, vol. 7, pp. 299–332.

[113] M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding, "Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task," *Journal of voice*, vol. 25, no. 1, pp. 44–53, 2011.

[114] P. Gómez-Vilda, R. Fernández-Baillo, V. Rodellar-Biarge, V. N. Lluis, A. Álvarez-Marquina, L. M. Mazaira-Fernández, R. Martínez-Olalla, and J. I. Godino-Llorente, "Glottal source biometrical signature for voice pathology detection," *Speech Communication*, vol. 51, no. 9, pp. 759–781, 2009.

[115] L. M. Kopf, C. Jackson-Menaldi, A. D. Rubin, J. Skeffington, E. J. Hunter, M. D. Skowronski, and R. Shrivastav, "Pitch strength as an outcome measure for treatment of dysphonia," *Journal of Voice*, vol. 31, no. 6, pp. 691–696, 2017.

[116] G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, A. Al-nasheri, and M. A. Bencherif, "Voice pathology detection using interlaced derivative pattern on glottal source excitation," *Biomedical signal processing and control*, vol. 31, pp. 156–164, 2017.

[117] Y. Koike and J. Markel, "Application of inverse filtering for detecting laryngeal pathology," *Annals of Otology, Rhinology & Laryngology*, vol. 84, no. 1, pp. 117–124, 1975.

[118] F. Klingholz, "The measurement of the signal-to-noise ratio (snr) in continuous speech," *Speech Communication*, vol. 6, no. 1, pp. 15–26, 1987.

[119] Y. Zhang and J. J. Jiang, "Acoustic analyses of sustained and running voices from patients with laryngeal pathologies," *Journal of Voice*, vol. 22, no. 1, pp. 1–9, 2008.

[120] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1329–1334, 1986.

[121] K. Shama, A. Krishna, and N. U. Cholayya, "Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–9, 2006.

[122] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[123] J. I. Godino-Llorente and et al., "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *J. of Voice*, vol. 24, no. 1, pp. 47–56, 2010.

[124] F. Rubén, G.-L. Juan, Ignacio, S.-L. Nicolás, O.-R. Víctor, and M. G.-A. Juana, "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: sustained vowels and running speech," *Journal of Voice*, vol. 27, no. 1, pp. 11–23, 2013.

[125] A. Al-Nasheri, Z. Ali, G. Muhammad, and M. Alsulaiman, "Voice pathology detection using auto-correlation of different filters bank," in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2014, pp. 50–55.

[126] G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio, and J. Révis, "Characterization of the pathological voices (dysphonia) in the frequency space," 2007.

[127] B. Hammarberg, B. Fritzen, J. Gauffin, and J. Sundberg, "Acoustic and perceptual analysis of vocal dysfunction," *Journal of phonetics*, vol. 14, no. 3-4, pp. 533–547, 1986.

[128] K. Tanner and et al., "Spectral moments of the long-term average spectrum: sensitive indices of voice change after therapy?" *J. of Voice*, vol. 19, no. 2, pp. 211–222, 2005.

[129] G. Kovačić, P. Boersma, and H. Domitrović, "Long-term average spectra in professional folk singing voices: A comparison of the klapa and dozivački styles," *Proc. Inst. of Phonetic Sciences, Univ. of Amsterdam*, vol. 25, pp. 53–64, 2003.

[130] T. F. Cleveland, J. Sundberg, and R. Stone, "Long-term-average spectrum characteristics of country singers during speaking and singing," *Journal of voice*, vol. 15, no. 1, pp. 54–60, 2001.

[131] K. Peter, "LTAS criteria pertinent to the measurement of voice quality," *J. of Phonetics*, vol. 14, no. 3-4, pp. 477–482, 1986.

[132] G. Muhammad, "Voice pathology detection using vocal tract area," in *2013 European Modelling Symposium*. IEEE, 2013, pp. 164–168.

[133] J.-W. Lee, H.-G. Kang, J.-Y. Choi, and Y.-I. Son, "An investigation of vocal tract characteristics for acoustic discrimination of pathological voices," *BioMed Research International*, vol. 2013, 2013.

[134] B. Halberstam, "Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels," *ORL*, vol. 66, no. 2, pp. 70–73, 2004.

[135] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Analysis and classification of phonation types in speech and singing voice," *Speech Communication*, 2020.

[136] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[137] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2649–2658, 1998.

[138] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Computer Speech & Language*, vol. 27, no. 4, pp. 1028–1047, 2013.

[139] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Communication*, vol. 55, no. 2, pp. 295–314, 2013.

[140] R. Glave and A. Rietveld, "Is the effort dependence of speech loudness explicable on the basis of acoustical cues?" *The Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 875–879, 1975.

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

[142] S. Baghel, S. M. Prasanna, and P. Guha, "Exploration of excitation source information for shouted and normal speech classification," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 1250–1261, 2020.

[143] J. Sundberg, E. Fahlstedt, and A. Morell, "Effects on the glottal voice source of vocal loudness variation in untrained female and male voices," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 879–885, 2005.

[144] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, and P. Alku, "Analysis and synthesis of shouted speech." in *INTERSPEECH*, 2013, pp. 1544–1548.

[145] R. Wayland and A. Jongman, "Acoustic correlates of breathy and clear vowels: The case of khmer," *Journal of Phonetics*, vol. 31, no. 2, pp. 181–201, 2003.

[146] G. d. Krom, "Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 4, pp. 794–811, 1995.

[147] S. Narasimhan and K. Vishal, "Spectral measures of hoarseness in persons with hyperfunctional voice disorder," *Journal of Voice*, vol. 31, no. 1, pp. 57–61, 2017.

[148] B. R. Kumar, J. S. Bhat, and P. Mukhi, "Vowel harmonic amplitude differences in persons with vocal nodules," *Journal of Voice*, vol. 25, no. 5, pp. 559–561, 2011.

[149] M. Eye and E. Infirmary, "Elemetrics disordered voice database (version 1.03)," *Voice and Speech Lab, Boston, MA*, 1994.

[150] W. J. Bogdan, "Saarbruecken voice database," 2007.

[151] J. I. Godino-Llorente and et al., "Acoustic analysis of voice using wpcvox: a comparative study with multi dimensional voice program," *European Archives of Oto-Rhino-Laryngology*, vol. 265, no. 4, pp. 465–476, 2008.

[152] T. A. Mesallam, M. Farahat, K. H. Malki, M. Alsulaiman, Z. Ali, A. Al-Nasheri, and G. Muhammad, "Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *Journal of healthcare engineering*, vol. 2017, 2017.

[153] J. I. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, S. Aguilera-Navarro, and P. Gomez-Vilda, "An integrated tool for the diagnosis of voice disorders," *Medical engineering & physics*, vol. 28, no. 3, pp. 276–289, 2006.

[154] S. Coretta, "Modelling electroglottographic data with wavegrams and generalised additive mixed models," 2019.

[155] H. Pulakka *et al.*, "Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography," Master's thesis, Helsinki University of Technology, 2005.

[156] A. O'Cinneide, D. Dorran, and M. Gainza, "Linear prediction: The problem, its solution and application to speech," 2008.

[157] E. Holmberg, "Aerodynamic measurements of normal voice," Ph.D. dissertation, Department of Linguistics, University of Stockholm, 1993.

[158] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[159] P. Alku, "Parameterisation methods of the glottal flow estimated by inverse filtering," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

[160] K. Gurugubelli and A. K. Vuppala, "Stable implementation of zero frequency filtering of speech signals for efficient epoch extraction," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1310–1314, 2019.

[161] M. H. Javid, K. Gurugubelli, and A. K. Vuppala, "Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech," in *Proc. ICASSP*, 2020, pp. 6754–6758.

[162] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[163] S. R. Kadiri, P. Alku, *et al.*, "Mel-frequency cepstral coefficients of voice source waveforms for classification of phonation types in speech," *Proc. INTERSPEECH*, pp. 2508–2512, 2019.

[164] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[165] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, *et al.*, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, 2013, pp. 148–152.

[166] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[167] D. R. Boone, S. C. McFarlane, S. L. Von Berg, and R. I. Zraick, "The voice and voice therapy," 2005.

[168] E. Seifert and J. Kollbrunner, "An update in thinking about nonorganic voice disorders," *Archives of Otolaryngology–Head & Neck Surgery*, vol. 132, no. 10, pp. 1128–1132, 2006.

[169] P. Clarós, A. Karlikowska, A. Clarós-Pujol, A. Clarós, and C. Pujol, "Psychogenic voice disorders literature review, personal experiences with opera singers and case report of psychogenic dyspho-nia in opera singer," *Int J Depress Anxiety*, vol. 2, p. 015, 2019.

[170] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *Journal of Voice*, vol. 33, no. 6, pp. 947.e11–947.e33, 2018.

[171] S. R. Kadiri and P. Alku, "Glottal features for classification of phonation type from speech and neck surface accelerometer signals," *Computer Speech & Language*, vol. 70, p. 101232, 2021.

[172] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing," *Journal of New Music Research*, vol. 42, no. 2, pp. 171–186, 2013.

[173] Y. Maryn, M. De Bodt, and N. Roy, "The acoustic voice quality index: toward improved treatment outcomes assessment in voice disorders," *Journal of Communication Disorders*, vol. 43, no. 3, pp. 161–174, 2010.

[174] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 20, no. 3, pp. 994–1006, 2011.

[175] A. I. Koutrouvelis, G. P. Kafentzis, N. D. Gaubitch, and R. Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, no. 2, pp. 316–328, 2016.

[176] C. M. Vikram and S. R. M. Prasanna, "Epoch extraction from telephone quality speech using single pole filter," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 25, no. 3, pp. 624–636, 2017.

[177] Y. M. Keerthana, M. K. Reddy, and K. S. Rao, "Cwt-based approach for epoch extraction from telephone quality speech," *IEEE Sig. Process. Lett.*, vol. 26, no. 8, pp. 1107–1111, 2019.

[178] K. Gurugubelli, H. M. Javid, K. R. Alluri, and A. K. Vuppala, "Toward improving the performance of epoch extraction from telephonic speech," *Circuits, Systems, and Sig. Process.*, pp. 1–15, 2020.

[179] P. Gangamohan and S. V. Gangashetty, "Epoch extraction from speech signals using temporal and spectral cues by exploiting harmonic structure of impulse-like excitations," in *Proc. ICASSP*. IEEE, 2019, pp. 6505–6509.

[180] S. R. Kadiri, A. Paavo, and B. Yegnanarayana, "Comparison of glottal closure instants detection algorithms for emotional speech," in *Proc. ICASSP*. IEEE, 2020, pp. 7379–7383.

[181] A. I. Koutrouvelis, G. P. Kafentzis, N. D. Gaubitch, and R. Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, no. 2, pp. 316–328, 2015.

[182] V. Khanagha, K. Daoudi, and H. M. Yahia, "Detection of glottal closure instants based on the microcanonical multiscale formalism," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, no. 12, pp. 1941–1950, 2014.

[183] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. ICASSP*. IEEE, 2002, pp. 349–352.

[184] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 15, no. 1, pp. 34–43, 2006.

[185] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," *arXiv preprint arXiv:2001.00459*, 2019.

[186] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception*. Elsevier, 1992, pp. 429–446.

[187] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[188] B. C. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

[189] J. C. Brown, "Calculation of a constant q spectral transform," *JASA*, vol. 89, no. 1, pp. 425–434, 1991.

[190] K. Gurugubelli and A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6410–6414.

[191] G. Aneeja and Y. Bayya, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 23, no. 4, pp. 705–717, 2015.

[192] P. Yannis and et al., "Music classification by low-rank semantic mappings," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2013, no. 1, p. 13, 2013.

[193] H. K. Maganti and et al., "Auditory processing-based features for improving speech recognition in adverse acoustic conditions," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2014, no. 1, p. 21, 2014.

[194] B. Schuller and et al., "The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition," in *Proc. INTERSPEECH*, 2015.

[195] C.-W. Jo and D.-H. Kim, "Analysis of disordered speech signal using wavelet transform," in *Fifth International Conference on Spoken Language Processing*, 1998.

[196] G. Gidaye, J. Nirmal, K. Ezzine, and M. Frikha, "Wavelet sub-band features for voice disorder detection and classification," *Multimedia Tools and Applications*, vol. 79, no. 39, pp. 28 499–28 523, 2020.

[197] M. Markaki and Y. Stylianou, "Normalized modulation spectral features for cross-database voice pathology detection," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[198] L. Chen, C. Wang, J. Chen, Z. Xiang, and X. Hu, "Voice disorder identification by using hilbert-huang transform (HHT) and k nearest neighbor (KNN)," *Journal of Voice*, pp. 932.e1–932.e11, 2021.

[199] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 3, pp. 421–430, 2005.

[200] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocology*, vol. 36, no. 2, pp. 60–69, 2011.

[201] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "Pathological voice analysis and classification based on empirical mode decomposition," in *Development of multimodal interfaces: active listening and synchrony*. Springer, 2010, pp. 364–381.

[202] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1170–1179, 2013.

[203] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ZTWCCs)." in *Proc. INTERSPEECH*, 2018, pp. 232–236.

[204] K. S. Reddy and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC)." in *Proc. INTERSPEECH*, 2018, pp. 441–445.

[205] R. G. Stockwell, L. Mansinha, and R. Lowe, "Localization of the complex spectrum: the s transform," *IEEE transactions on signal processing*, vol. 44, no. 4, pp. 998–1001, 1996.

[206] E. Sejdic, L. Stankovic, M. Dakovic, and J. Jiang, "Instantaneous frequency estimation using the S-Transform," *IEEE signal processing letters*, vol. 15, pp. 309–312, 2008.

[207] W. Lin and M. Xiaofeng, "An adaptive generalized S-transform for instantaneous frequency estimation," *Signal Processing*, vol. 91, no. 8, pp. 1876–1886, 2011.

[208] I. Djurović, E. Sejdić, and J. Jiang, "Frequency-based window width optimization for S-transform," *AEU-International Journal of Electronics and Communications*, vol. 62, no. 4, pp. 245–250, 2008.

[209] A. Moukadem, Z. Bouguila, D. O. Abdeslam, and A. Dieterlen, "A new optimized Stockwell transform applied on synthetic and real non-stationary signals," *Digital Signal Processing*, vol. 46, pp. 226–238, 2015.

[210] M. Hamidia and A. Amrouche, "A new robust double-talk detector based on the Stockwell transform for acoustic echo cancellation," *Digital Signal Processing*, vol. 60, pp. 99–112, 2017.

[211] H. K. Vydana and A. K. Vuppala, "Detection of fricatives using S-transform," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. 3896–3907, 2016.

[212] S. Saoud, S. Bousselmi, M. B. Naser, and A. Cherif, "New speech enhancement based on discrete orthonormal stockwell transform," *International Journal of Advanced Computer Science and Applications 7 (10) 2016*.

[213] M. Zhu, Z. Jiang, X. Zhang, and Y. Qi, "A S-transform based spectrum enhancement method for complex noise environment," in *2014 International Conference on Audio, Language and Image Processing*. IEEE, 2014, pp. 382–385.

[214] A. Revathi and N. Sasikaladevi, "Hearing impaired speech recognition: Stockwell features and models," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 979–991, 2019.

[215] R. G. Stockwell, "A basis for efficient representation of the S-transform," *Digital Signal Processing*, vol. 17, pp. 371–393, 2007.

[216] G. Livanos, N. Ranganathan, and J. Jiang, "Heart sound analysis using the S transform," in *Computers in Cardiology 2000. Vol. 27*. IEEE, 2000, pp. 587–590.

[217] A. Moukadem, A. Dieterlen, N. Hueber, and C. Brandt, "A robust heart sounds segmentation module based on S-transform," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 273–281, 2013.

[218] C. R. Pinnegar, H. Khosravani, and P. Federico, "Time–frequency phase analysis of ictal eeg recordings with the S-transform," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 11, pp. 2583–2593, 2009.

[219] C. Beuter and M. Oleskovicz, "S-transform: from main concepts to some power quality applications," *IET Signal Processing*, vol. 14, no. 3, pp. 115–123, 2020.

[220] M. Geng, W. Zhou, G. Liu, C. Li, and Y. Zhang, "Epileptic seizure detection based on stockwell transform and bidirectional long short-term memory," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 3, pp. 573–580, 2020.

[221] S. Ventosa, C. Simon, M. Schimmel, J. J. Dañobeitia, and A. Mànuel, "The S-transform from a wavelet point of view," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2771–2780, 2008.

[222] E. Sejdić, I. Djurović, and J. Jiang, "A window width optimized S-transform," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–13, 2007.

[223] S. Assous and B. Boashash, "Evaluation of the modified S-transform for time-frequency synchrony analysis and source localisation," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–18, 2012.

[224] 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.

[225] B. S. Atal, "Speech analysis and synthesis by linear prediction of the speech wave," *The journal of the acoustical society of America*, vol. 47, no. 1A, pp. 65–65, 1970.

[226] J. C. Saldanha and et al., "Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features," *J. of Medical Imaging and Health Informatics*, vol. 4, no. 2, pp. 168–173, 2014.

[227] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Comparative study of different epoch extraction methods for speech associated with voice disorders," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6923–6927.

[228] P. M. Bentley and J. McDonnell, "Wavelet transforms: an introduction," *Electronics & communication engineering journal*, vol. 6, no. 4, pp. 175–186, 1994.

[229] J. Crowe, N. Gibson, M. Woolfson, and M. G. Somekh, "Wavelet transform as a potential tool for ecg analysis and compression," *Journal of biomedical engineering*, vol. 14, no. 3, pp. 268–272, 1992.

[230] S. Waldekar and G. Saha, "Wavelet transform based mel-scaled features for acoustic scene classification." in *INTERSPEECH*, vol. 2018, 2018, pp. 3323–3327.

[231] A. Tsanas, "Accurate telemonitoring of parkinson's disease symptom severity using non-linear speech signal processing and statistical machine learning," Ph.D. dissertation, Oxford University, UK, 2012.