April 2021

An Automated Framework for Connected Speech Evaluation of Neurodegenerative Disease: A Case Study in Parkinson's Disease

Sai Bharadwaj Appakaya
University of South Florida

Scholar Commons Citation

Appakaya, Sai Bharadwaj, "An Automated Framework for Connected Speech Evaluation of Neurodegenerative Disease: A Case Study in Parkinson's Disease" (2021). *USF Tampa Graduate Theses and Dissertations.*
https://digitalcommons.usf.edu/etd/9574

An Automated Framework for Connected Speech Evaluation of Neurodegenerative Disease:

A Case Study in Parkinson's Disease

by

Sai Bharadwaj Appakaya

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Electrical Engineering
College of Engineering
University of South Florida

Major Professor: Ravi Sankar, Ph.D.
Ismail Uysal, Ph.D.
Ehsan Sheybani, Ph.D.
Sriram Chellappan, Ph.D.
Supraja Anand, Ph.D.

Date of Approval:
March 22, 2021

Keywords: Phonatory Analysis, Pitch Synchronous Segmentation, LSTM Autoencoders,
Overfit Factor, Dysarthria, Dysphonia

Dedication

To my parents, family and friends for their unwavering support and encouragement.

Acknowledgments

I would like to sincerely thank my advisor, Dr. Ravi Sankar, for providing constant support and guidance throughout my research. I have learned many lessons under his mentorship that would be helpful forever in my life. I would like to thank Dr. Supraja Anand for providing invaluable insights into research and guidance whenever I needed it. I am very grateful for the effort spent by my dissertation committee members Dr. Uysal, Dr. Sheybani and Dr. Chellappan in helping me reach my goal.

I would like to thank my friends Harsha Vardhan and Swamy Rakesh for their steadfast support throughout my life as a graduate student. Special thanks to Prashanth Adithya and Shraddha Pandey for making me feel at home during my time in USF.

Many thanks to all my colleagues in iCONS group at USF. Muath Alsuhaibani and Sravani Kolli, thank you very much for your time.

To my parents and my family, thank you for your endless patience and encouragement throughout my life. I will forever be in your debt for your unconditional love, prayers and confidence in me.

Table of Contents

List of Tables

List of Figures

Abstract

Neurodegenerative diseases affect millions of people around the world. The progressive degeneration worsens the symptoms, heavily impacting the quality of life of the patients as well as the caregivers. Speech production is one of the physiological processes affected by neurodegenerative diseases like Alzheimer’s disease, amyotrophic lateral sclerosis (ALS) and Parkinson’s disease (PD). Speech is the most basic form of communication, and the effect of neurodegeneration degrades speech production, thereby reducing social interaction and mental well-being. PD is the second most common neurodegenerative disease affecting speech production in 90% of the diagnosed individuals. Speech analysis methods for PD in clinical methods are primarily perceptual. The acoustic analysis of speech impairments could help understand speech patterns that differ between pathological and healthy speech. This knowledge can benefit the early diagnosis and telemonitoring applications. Speech analysis is extremely beneficial because it is non-invasive, fast and easy to implement remotely. Studies using established analysis methods have been focused on speech due to these merits and steadily improving in the past decade.

The objective of most research studies working with Parkinsonian speech is the development of automatic evaluation methods. Most of these studies use sustained vowel phonations due to their simplicity in the analysis. Fewer studies have focused on using connected speech tasks like conversational speech, passage readings or monologues due to the reduced control over the data and complexity in analysis. Typical speech impairments observed in individuals with PD are identified in connected speech, and studies have shown the superiority of connected speech over sustained phonations in disease detection. The traditional processing steps adopted in existing methods have been criticized for spectral distortion and smoothing effects. In this dissertation, a new framework for the automatic evaluation of connected speech has been compared to the conventional methods.

In contrast to the existing methods, the proposed framework uses pitch synchronous segmentation of voiced components of connected speech to ensure consistency in processing, avoiding distortion and smoothing effects. Novel pitch synchronous features (PSFs) quantifying the cycle-to-cycle perturbations are extracted from each voiced segment, and the covariances of the 1st order differences in these features are used to train classifiers. This methodology helps in extracting the phonatory and articulatory deficits that are hard to quantify using traditional methods. The impact of pitch synchronous segmentation has also been tested using unsupervised feature extraction using LSTM Autoencoder.

This study's results are encouraging. Deviations from the established methods incorporated in the proposed framework are evaluated systematically. Evaluation using existing methods using sustained phonations show impressive 92% cross-validation accuracy, but when tested on a different dataset, accuracy falls to 50%. With the proposed framework using PSF covariances, the cross-validation accuracy was 80%, and test accuracy using a different dataset was 72%. With LSTM Autoencoders, the cross-validation and test accuracies were 89% and 73%. Thus, the traditional methods deliver excellent performance with familiar data but fail with new data. The proposed framework using PSFs and Autoencoders based features deliver comparable performance with familiar and new datasets.

Chapter 1: Introduction

1.1 Digital Speech Processing

Speech is the primary form of communication mastered over years of practice through constant listening and adaptation. We often do not perceive the complexity of speech production and perception until we try to teach a computer to either synthesize sounds close to the natural speech by a human being or make the computer understand voice commands. Certain aspects of speech, like intonational variability, lexical stress and pauses, add uniqueness concerning qualities like personal identity, context and emotional state. Collectively, these aspects can be studied to identify syntactical rules followed while producing the speech. Such complex syntactic rules subject to variability between different people under different contexts and emotional states are hard to model and teach a computer.

The sense of hearing plays a vital role in our verbal communication and it is tough to understand how the brain decodes the information. It is well known that brain often on its memory to identify the words and sentences to comprehend the language. Auditory systems have been viewed as frequency analyzers that can understand speech picked up by the ear through a spectro-temporal representation [1, 2]. The field of digital speech processing tries to decode speech similarly by modeling speech under different domains targeting various aspects.

1.1.1 History

Over the years, due to advances in research and an increase in computational capabilities, digital speech processing has continuously been updating the processing methods according to the requirements. The Fourier theory shows that decomposition of any waveform is possible by using a series of sinusoids representing different frequencies at different energy levels that can reproduce the waveform very closely [3]. Thus, at a fundamental level, a vowel sound of any utterance length recorded by a microphone can be represented by its Fourier spectrum. This representation enabled the utilization of various spectral parameters like the formant frequencies for effectively coding various sounds and mitigating large memory requirements. Following this, the next step in the evolution is based on pattern recognition systems, linear predictive coding (LPC) and clustering algorithms. These tools were used to develop models that can identify isolated words, digits and even short sentences known to the system beforehand. These advances had applications in security and automation systems. Advances in pattern recognition systems and statistical modeling resulted in developing more complex statistical algorithms like Hidden Markov Models (HMM), stochastic language modeling. These algorithms took advantage of the joint probability of a larger pool of words which expanded the reach of systems and enabled the recognition of connected words from larger pools and continuous speech. By the early 90’s speech processing included stochastic language understanding, finite-state machines and statistical learning, which enabled the processing of longer sentences and understood various deeply embedded complex semantics and syntaxes. These systems also helped synthesize more natural speech, which sounded very rigid and robotic until that time. These advances also strengthened the acoustic analysis methods where features like Mel-frequency cepstral coefficients (MFCC), fundamental frequency (F0) and others were used for creating models for each word. The inter-word strengths represented by the co-occurrence of feature clusters were used to better understand the complex semantics using mathematical and statistical methods. With the advent of robust processing systems that can handle complex machine learning algorithms, language models highly evolved in the early 2000s. By then, using concatenative synthesis, advanced machine learning and mixed-initiative dialog, computers could analyze spoken dialog and identify the speech from each individual. The cocktail party problem, which has been under research for so long, had significant advances through methods like beam formation and microphone arrays [4].

From the late 2000s, advances in artificial intelligence (AI) and neural networks have immensely impacted speech processing. Primarily these advances focused on speech and speaker recognition systems. Later, they became vital for developing automated voice assistant systems like Apple’s Siri, Google Assistant, and Amazon’s Alexa with the advancements in speech synthesis mechanisms [5]. The journey of speech processing through history is fascinating, especially with the advances in the current decade (2010-2020). With the ever-increasing ability to process more complex tasks, speech processing scope also increased over the years. Though it started to find an efficient way to store and synthesize speech samples, today, the applications are very diversified and found their way into various disciplines [5].

1.1.2 Applications

Digital speech processing has applications that enable various systems to understand, communicate and interpret their users. These actions help increase the autonomy of the systems and decrease the users’ workload when implemented correctly. Some of the major applications are:

- *Speech coding*: This is an application of data compression where signals containing speech need to be compressed so that the memory requirements for storage. For communication, coding can mitigate the larger bandwidth requirement and decrease the transmission latency. It uses parameter estimation methods specific to speech compressed using compression algorithms, thereby enabling high-quality reconstruction [6].

- *Speech recognition*: Speech-to-text is one of the oldest speech recognition applications targeted at improving the human-machine interface. Speech recognition enables a computer to identify the speech and convert it into text has many applications like medical transcription, security systems, biometric identification and home automation.
- *Speaker Recognition*: Major applications of speaker recognition systems involve the security systems for authentication and surveillance. In the modern era, with numerous mobile devices, speaker recognition has forensic recognition applications wherein the voiceprint will be used to identify a person of interest.
- *Speech Understanding*: While speech recognition concentrated on identifying the exact words being spoken, speech understanding aims at understanding the comprehensive meaning of the spoken text, which is crucial for automating any task using speech.
- *Speech synthesis*: Synthesis of natural speech using the parametric representation as a response during human-machine interaction. Popular voice assistant systems are examples of this application where they synthesize responses to the user based on his/her query. These voice assistants perform both understanding and synthesis operations intending to act as a real assistant.
- *Language processing*: This includes speech recognition and understanding. The recent advances in Natural Language Processing (NLP) take advantage of powerful tools like deep neural networks (DNN), recurrent neural networks (RNN), long short-term memory (LSTM) neural networks for performing various tasks. These tasks are usually aimed at analyzing and understanding language by checking grammar and style. Predict the words before they are uttered and even summarize the content of a recording.

In addition to these conventional applications, speech analysis is also popular in some multi-disciplinary applications directly or indirectly. Music classification and separation are applications where speech processing techniques have been used to identify and classify vocal music and instrumental music. It is also being applied for track separation or editing the music recordings [7]. Research in the field of biomedicine makes use of speech processing to identify and evaluate speech samples from pathological speech samples for voice quality and intelligibility. Disease severity detection, classification and voice quality improvement are some of the major applications.

1.2 Biomedical Applications of Speech Processing

Speech production and perception are complex manifestations of the cognitive, neurological and physiological states of the speaker. Acquisition, storage and processing of speech samples is a simple undertaking compared to other bio signals. Physiological or neurological ailments impacting prosody, phonation, articulation and fluency domains of speech production are studied using perceptual and acoustic methods. These four domains contain the neuromotor and cognitive information responsible for the comprehension and naturalness of speech, subject to variations specific to the ailment under research [8]. Neurological functions involved in speech production are initialized in the Broca's area, present in the left hemisphere. It is responsible for planning the functions of various muscle groups present in the larynx and vocal tract. Hence, the impact on speech can result from dysfunction in one or more of the three systems: neuromotor, cognitive, or psychological.

Dysfunction in the neuromotor system can manifest as physiological impacts on speech due to damage in the muscles or the underlying nerves carrying signals from the brain to muscles. Muscular functional abnormalities in any part of the vocal tract or diaphragm manifests as dysarthria, causing slurred and/or slowed speech and abnormalities in pitch, causing reduced voice quality, articulation difficulties and labored speech [9]. Some causes for this kind of disorder are stroke, tumors and multiple sclerosis. Voice disorders can also arise due to laryngeal cancer, Reinke's edema, subglottic stenosis, vocal fold nodules and polyps. Diseases or conditions that can cause an impact on the larynx can result in voice disorders, and research studies are targeted towards voice quality in such cases.

1.2.1 Neurodegenerative Language Disorders

Malfunction in the neuromotor system can result in a slow degeneration of the secondary neurons responsible for carrying information to various muscular groups. Parkinson's Disease (PD) is one such neurodegenerative disease that results in a slow degeneration of the dopaminergic neurons. Amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS) and Huntington's disease also cause neuronal degeneration, showing its impact on speech production. Cognitive degeneration observed in disorders like Alzheimer's disease causes a progressive decline of the primary neurons in the cortex, which leads to deterioration of short-term memory, leading to reduced communication abilities of the patient. Psychological pathologies arise due to malfunctioning of mutual interaction of neuron subsets, causing improper social behavior. Autism, depression or psychotic diseases are some of the well-documented psychological pathologies causing abnormalities in speech production and social behavior [10].

All the pathologies mentioned here result in speech disorders at various levels in some (or all) speech domains (prosody, phonation, articulation and fluency). Hence, clinical diagnosis of many neurodegenerative diseases involves language assessment as a crucial step [11]. The study of extended speech output provides a valuable source of knowledge encompassing the phonological, phonetic, morpho-syntactic, lexico-semantic, and pragmatic levels of language organization. Clinical research methods identify various distinctive linguistic variables that can measure various language deficits and extract them from different speech tasks designed to elicit various deficits through perceptual studies. The variables studied under this method are studied as scores that can be used with various tools to identify statistically significant differences among the subject groups. Advances in computational analysis methods are used in speech technology to evaluate anomalies detected in the prosody, phonation, articulation and fluency domains at each level. These methods add statistical robustness, making them reliable and objective solutions for clinical and rehabilitative applications [10].

1.2.2 Perceptual and Acoustic Studies

Perceptual methods study the variations in the speech using a set of 'linguistic variables' or 'features' correlating to the symptoms specific to the pathology under focus. These methods rely upon trained or regular listeners who understand the descriptions of the features they must rate on a predefined scale. Veronica *et al.* have reviewed 61 perceptual studies focusing on the connected speech from patients with neurodegenerative language disorders [11]. They identified 61 most relevant features grouped under five different linguistic levels for this application. Sub-groups of these 61 features were used for studying nine different pathologies, including Parkinson’s disease, Alzheimer’s disease, Huntington’s disease and amyotrophic lateral sclerosis. These studies used various speech tasks, each one explicitly targeting five linguistic-levels. This method of extracting information, though time-consuming at times, can take advantage of the power of intuition and experience that are hard to model and instill into a computer. Some researchers claim perceptual methods or naked ear evaluation of speech and vocal perturbations can assist in the differential diagnosis. It is rather hard to compare these methods between research studies due to the heterogeneity at various stages of analysis. Though similar features are analyzed in different studies, factors like listeners' experience in the field, the semantics of the feature descriptions, differences between control groups from different studies can produce contradicting conclusions [12].

Acoustic analysis of neurodegenerative speech disorders is defined as the analysis of features extracted using speech technology methods. In the case of many neurodegenerative diseases, early research using speech processing methods was aimed at rehabilitative studies and vocal disorder intensity measurements. Vocal features like the pitch variations, studied under perceptual methods [13], were replicated using fundamental frequency measures extracted using speech processing methods [14]. Studies over the impact of various speech therapy/treatment methods have also used acoustic features extracted from pathological speech before and after the treatment to find any significant group differences [15]. In the past decade, numerous studies [16-22] have targeted the use of speech to extract various prevailing and novel features like fundamental frequency and recurrence period density entropy (RPDE), respectively, for classification of various neurodegenerative diseases. Remote and passive data collection and processing mechanisms for diagnosis and monitoring applications have gained traction in the past five years due to the increasing prevalence of cloud-based computing and the availability of wearable sensors [23]. Speech analysis has been one of the main focal points for remote applications due to its stress-free collection methods and robustness in data collection methods [24].

A brief literature search shows that acoustic analysis methods can potentially impact diagnostic methods of pathologies like Parkinson’s disease, which causes significant fluctuations in speech production. Current state-of-the-art methods in this area have been using advanced neural network architectures to model pathological speech for classification applications.

Advanced networks like recurrent neural networks (RNN) [25], long short term memory (LSTM) neural networks [26] and convolutional neural networks (CNN) [27] have been showing promising results. In comparison, it can be noted that acoustic studies for disease classification have deeper focus on Parkinson’s disease than other modalities. This keener interest can be attributed to the more significant impact of Parkinson’s disease over speech production than other neurodegenerative diseases [9]. Hence, analysis of Parkinsonian speech is a very promising research avenue for early-diagnosis and monitoring applications.

1.2.3 Challenges in Acoustic Speech Analysis for Parkinson’s Disease Classification

Acoustic studies over the use of speech from people with Parkinson’s disease have been following perceptual studies in the type of speech tasks used for analysis and type of features analyzed. Acoustic methods aim to take advantage of the inherent objectivity due to algorithmic processing in handling every sample in the same manner. Acoustic studies also try to impart intelligence to the statistical methods by using advanced methods proven to take advantage of semantics within the data and perform very close to humans for various applications like speech translation [28]. The four primary challenges acoustic studies must face in this research area are as follows:

- *Disease prognosis*: The timeline of Parkinson’s disease is very heterogeneous and known to be different in different subjects. The order of appearance of various symptoms during disease progression is sporadic, and symptoms are known to worsen with time, having a massive impact on the quality of life. Thus, the analysis methods developed for the speech from people with Parkinson’s disease must be robust and pick up subtle variations for successful classification.

- *Speech task*: Different types of speech tasks are used to elicit phonatory, articulatory and prosodic irregularities due to Parkinson’s disease. Researchers working on specific speech tasks have often listed the merits of the tasks they used and occasionally discouraged the usage of some popular tasks under analysis in this area. It is challenging and critical to identifying the optimal tasks for a comprehensive analysis protocol in this research direction.
- *Clinical significance of features*: The spectro-temporal acoustic features used for classification studies are very diverse. For the most part, these features are devised to quantify various symptoms identified in different speech pathologies. Due to the availability of datasets containing the features and their labels, there is an abundance of studies that focus on developing advanced classification methods without knowing their clinical significance. It is imperative to develop features with the ability to capture the slightest abnormalities to yield superior classification.
- *Reliability and reproducibility*: Classification protocols developed using speech technologies must be stable and reproducible with a different dataset. It must be underlined that existing methodologies are often tested with the available dataset and fail to generalize with a new dataset. The novel analysis protocol presented in this work will be proved to be reproducible with a different dataset that increases the system's reliability and can work with voice and language disorders.

1.3 Research Motivation

The pathogenesis of many neurodegenerative disorders is complicated and heterogeneous, leading the research related to their treatment and rehabilitation into various directions [29]. Diagnosis of many of these disorders relies heavily upon the information provided by the patient or on the neuroimaging methods. The lack of biomarkers required for a reliable diagnosis makes symptomatic research a hopeful direction in understanding the underlying etiologies. Researchers are usually interested in studying very prominent symptoms that have a significant impact on the patient's quality of life. In Parkinson’s disease, prominent symptoms like rigidity and bradykinesia are treated as the top two troublesome disease effects to the patients [30]. The next prominent effect is the reduction in speech quality. Though most of the patients experience speech disorders due to neurodegenerative diseases, it is not fully explored. Its impact is perceived to be stronger only in the advanced stages of disease progression.

Speech-based studies for neurodegenerative disorders have been gaining popularity recently due to the low cost and easy analysis methods. Lopez-de-Ipena *et al.* have studied the effectiveness of Alzheimer’s disease classification using linear and non-linear features from spontaneous speech in a multi-lingual dataset containing 70 speakers [31, 32]. Similar research studies have focused on examining speech to classify Parkinson’s disease using sustained phonations [33, 34]. Though these studies have shown that classification using speech can result in over 90% accuracy, they do not include connected speech primarily examined during the diagnostic process. Sustained phonations are often used because of their simplicity during data acquisition as well as processing. As mentioned earlier, the classification protocols developed around speech are required to be reliable and reproducible. Features like the fundamental frequency, formant frequencies are used by themselves for training classification models, resulting in overfitting to the dataset and lack of generalization. Many of these standard features are prone to bear a resemblance to dysarthric speech features when the speakers are over 40 years old [35, 36]. Hence, using the features without any modulation can produce models that get trained on those aspects of speech typical for some healthy speakers and result in misclassifications. During the diagnosis process, speech evaluations in the rating scales focus on overall vocalic dynamics rather than the similarities in the voiceprint. Current analysis methods use very few features like jitter and shimmer that target quantification of vocalic perturbations during phonation. The established speech technology methods that assume stationarity over small and overlapping segments of speech cannot capture the perturbations occurring at much smaller intervals than the duration of these segments.

For pathologies where the speech deficits are closely associated with phonatory shortfalls, acoustic methods targeting the perturbations can significantly impact the reliability of automatic classification systems. These methods have the potential to accurately identify the mild deficits in phonation, which is hard, even for a trained clinician, to pick up just through hearing. For pathologies like Parkinson's disease, such methods can also help in preliminary evaluation of the speech conceived to be not affected in the early stages [37].

Overall, speech has immense potential in containing the effects of neurological disorders from early stages. Current acoustic methods cannot contain the vocalic dynamics effectively in developing an objective tool for assisting diagnostic evaluations and telemonitoring applications. These models can possess the ability to identify the changes that are hard to perceive through the naked ear. Acoustic methods currently adopted in the research can be improved by changing the segmentation and feature extraction procedures that work with connected speech. In the pursuit of an ideal system that can identify the effects of the pathology and not just identify the resemblances in voiceprint, it is essential to take advantage of similarities in vocalic dynamics through systemic quantification of the vocal perturbations.

In this work, a novel classification framework has been proposed to fill in the gaps in current methods. The proposed framework aims to exploit the simplicity in processing sustained phonations and use it with connected speech using complex segmentation and feature extraction protocols. This framework will be evaluated comprehensively using speech from people with Parkinson's disease to compare against the existing methodologies and provide evidence for reliability and generalization. The proposed framework will be tested using the features extracted through a state-of-the-art LSTM Autoencoder under unsupervised training methods in addition to the novel feature extraction methods. Classification performances are compared between supervised and unsupervised methods to measure the effectiveness of the proposed framework.

1.4 Contributions

A novel framework has been developed to process the connected speech recordings from individuals and classify whether the speaker has a neurological disorder that impacts speech production or not.

- Developed a robust and automated segmentation algorithm that can identify the voiced portions of speech and segment them pitch synchronously [38].
- Developed a novel feature set consisting of 15 different temporal and spectral features extracted from pitch synchronous speech segments. These features are aimed to quantify the subtle variations in the voiced segments between different pitch cycles [39].
- Developed and evaluated a covariance-based feature transformation protocol that can contain the vocalic dynamics effectively and improve the automatic classification [40].
- Evaluated the performance using sustained phonations for classification.
    - Evaluated the classification performance under established frameworks using sustained phonations and standard features.
    - Evaluated the performance using the proposed framework with pitch synchronous segmentation and a novel feature set.

- Validated the compatibility of the proposed framework with novel features extracted pitch synchronously from sustained phonations and connected speech to determine the optimal speech task.
- Validated the efficiency of novel feature set by comparing the performance against established Mel-frequency Cepstral Coefficients extracted from the pitch synchronous speech segments [41].
- Developed an unsupervised feature extraction model based on the state-of-the-art LSTM Autoencoders that can provide features spectrograms of voiced segments of variable lengths.
    - Evaluated the compatibility of proposed framework with unsupervised features by comparing the performance using novel feature-set.
- Trained an ensemble of classifiers and validated the compatibility of each classifier for the classification application using the proposed framework by identifying the likelihood of the classifiers to overfit [42].

Compared to the existing state-of-the-art, the novel framework exhibits innovation in terms of feature extraction and transformation. The comparative studies exploring the efficiency of various changes made to the traditional methods provide strong evidence of the proposed framework's superiority to the existing methodologies. This work compares the performance of the novel features against traditional methods; it also compares the performance using unsupervised features, that are hard to manipulate to obtain biased performance results. As opposed to many existing studies, the classification performances are evaluated using data from a different dataset to illustrate the generalization capacity of this framework. This dissertation will serve the research community and clinicians working with neurodegenerative diseases and other speech and language disorders by providing a novel and robust method for analysis and classification.

1.5 Organization

Rest of the chapters in this dissertation are organized to provide necessary background and other information that can establish the importance of the framework and the evaluation methods.

In Chapter 2, the background information related to Parkinson's disease, its origins, prevalence, symptoms, and diagnosis is provided. It also covers the impact of PD on speech and a brief literature overview of the symptomatic research and acoustic analysis methods. An overlook of various acoustic features used for analysis and different machine learning methods under research is presented. Finally, the importance of connected speech based studies and the limitations in using connected speech is discussed in this chapter.

In Chapter 3, the proposed framework and a detailed explanation of the related materials and methods are described. Results from different evaluations of the proposed methodology using different datasets under various experimental settings are presented in this chapter. These results help in determine the efficiency of the proposed framework and its robustness. This chapter also includes the study replicating the currently popular methods using sustained phonations and results that help identify the problems with acoustic studies using sustained phonations.

In Chapter 4, an LSTM Autoencoder-based feature extraction protocol along with its classification performance is provided. This study provides evidence as to the performance of the framework using unsupervised feature extraction methods in identifying PD. This chapter also contains detailed information on the LSTM Architecture and the implementation steps involved in the Autoencoder training, and the classification methodology. At the end of this chapter, performance metrics from the classification using novel feature set and Autoencoder-based features is compared and discussed.

In Chapter 5, findings of this work that contribute to the research community by providing evidence for the drawbacks in the current methods and impact and robustness of the proposed framework are summarized. Future directions for improving automated evaluation methods using connected speech are also provided.

Chapter 2: Background

2.1 Parkinson’s Disease

Parkinson’s disease (PD) is a progressive neurodegenerative condition affecting about 1 to 2 per 1000 of the population at any given time and it increases with age affecting 1% of the population over 60 years [43]. After Alzheimer’s disease, PD is the second most commonly occurring neurodegenerative disorder. Initially, PD was considered to be a movement disorder with tremor, rigidity, postural instability and bradykinesia as significant symptoms. These symptoms are manifestations of deterioration of dopaminergic neurons in basal ganglia (midbrain), which communicate motor signals from the brain to skeletal muscles [44].

2.1.1 Prevalence

PD has a prevalence of 572 per 100,000 in North America among people aged ≥ 45 Years. By 2010, 680,000 individuals aged ≥45 years were diagnosed with PD in the US. This number was projected to 930,000 and 1,238,000 by 2020 and 2030, respectively [45]. The average onset age of PD is in the range of 65 to 70 years, but less than 5% of patients have an onset age of <40 years [43]. The incidence of PD has gender-based differences as it has been reported that its incidence and prevalence are higher for men by 1.5 to 2 times than women [46-48]. Muangpaisan, *et al.* in [49], it is reported that the prevalence of PD in Asia is 51.3 to 176.9 per 100,000 in door-to-door surveys and prevalence in record-based studies ranged from 35.8 to 68.3 per 100,000. The surveys are conducted in different regions of Asia, and the total population in each region is normalized to 100,000. The average hospitalization cost per day is $28,400, focusing on the top 10 procedures that the patients might have to undergo [50]. The annual economic impact of the PD in the USA is estimated at $10.8 billion. 58% of this $10.8 billion is spent on direct costs, including expenses for hospitals, medicines, care and other services. The direct expenses for patients are $10,043 to $12,491 more in the case of PD compared to other modalities [51].

2.1.2 Symptoms

The standard and noticeable symptoms of PD are mainly caused due to the degeneration of the dopaminergic neurons, which diminishes the signal flow in the neurological pathways. The deficiency in dopamine causes many abnormalities in motor functions throughout the central nervous system [52]. Literature has the mentioned several motor and non-motor symptoms of PD. There are four cardinal features of PD that can be grouped under the acronym TRAP: Tremor at rest, Rigidity, Akinesia (or bradykinesia) and Postural instability [53]. The tremor at rest or resting tremor can be natural (due to old age) or drug-induced, or due to PD. The tremors observed in people with PD are mostly the “pill-rolling tremors” (the tremors that make it look like the patient is trying to roll a pill using his/her fingers). Essential Tremor (ET) is another form of tremor with the etiologies like heredity or behavioral causes (alcoholism, smoking). In older adults, it is hard to distinguish between ET and tremors due to PD. Rigidity or freezing refers to increased resistance in the muscular extensions or contractions throughout the movement. This rigidity can also result from many other conditions like tetanus, and when this exists along with the resting tremor, “cogwheel rigidity” can be felt during limb mobilization [54, 55]. Postural instability is a combination of many muscular dysfunctionalities in the body of the person with PD. The stooped posture is a primary reason for the falls of the patient. The gait of the patient with PD is slow and on a narrow base with short steps or shuffling.

Further, the non-motor symptoms include a broad array of disorders. Blochberger *et al.* in [56] broadly classified the non-motor symptoms into four groups: Neuropsychiatric symptoms (dementia, depression, anxiety etc.), autonomic symptoms (constipation, urinary incontinence, hyperhidrosis etc.), sleep disorders (rapid eye movement during sleep, narcolepsy, etc.) and sensory (pain, paraesthesia) symptoms. Bostantjopoulou *et al.* presented the Non-Motor Symptoms Questionnaire details used in the clinical examination of the non-motor symptoms. They have a 30-item self-completed questionnaire for a comprehensive indication of presence or absence of non-motor symptoms in PD [57]. PD also causes speech and swallowing difficulties, and linguistic discrepancies. Together, these symptoms can affect the social interaction, physical and mental well-being and overall quality of life of the patients with PD significantly [58-60].

2.1.3 Diagnosis and Treatment

There is no definitive way to identify the presence of PD using tests of body fluids (blood, urine) or neuroimaging like MRI or CT. Even with advanced neuroimaging techniques, research indicates the requirement of long-term follow-up studies to validate their use in pre-motor disease [61]. There are no known biomarkers for the diagnosis which can differentiate between a person with PD and healthy control. The lack of definitive biomarkers diagnostic process for PD depends mainly on patients' history and subjective decisions made by neurologists and clinicians. The diagnostic process by clinical examination is typically conducted in two stages: 1. Search for the evidence (symptoms) of Parkinsonism, 2. Search for the presence of Akinesia (loss or impairment of the power of voluntary movement) [62].

Current diagnosis methods are based upon a questionnaire developed by Movement Disorders Society (MDS) called Unified Parkinson’s Disease Rating Scale (UPDRS) [63]. Part II and part III of this scale are targeted towards the motor aspects. Part II collects the information from patients, and part III assesses the motor signs of PD. It is scaled from 0 to 137, with 0 representing no symptoms and 137 representing severe motor impairment. Section 3.1 of UPDRS part III examines the speech deficits by evaluating the volume, prosody and clarity. The diagnostic decision is dependent on the score on the rating scale given by the clinician. This method is inherently subjective, resulting in frequent misdiagnosis, which is confirmed after autopsy. Drug-induced Parkinsonism or aparkinsonian syndrome have a massive impact on the misdiagnosis in PD. Studies showed that 20% of PD diagnosed patients showing typical symptoms are confirmed to have manifested those symptoms from other etiologies [64].

In addition to the insufficient diagnostic protocol, there is no cure for PD. Treatments for PD are focused on maintaining the quality of life using physical/motor and cognitive training, medication and deep brain stimulation (DBS) [65-68]. Physical and cognitive training aims to improve the brain's functionality to work around the deficits caused due to PD. While post-training tests show significant improvements in reducing symptoms, it has little to no effect on neurodegeneration. Medication prescribed typically improves dopamine levels or helps in developing alternate ways for neurotransmission. Levodopa is a drug that has been showing significant improvement in this direction. However, PD patients have shown behavioral addiction after chronic treatment using levodopa [69]. Recent studies on DBS have targeted the development of an implantable device for PD patients. DBS is intended to block electrical signals from the regions of the brain that control the motor activity to reduce the magnitude of the symptoms. Research using DBS has been constantly exploring new areas over the last two decades [70]. Studies have shown the DBS has alternative side-effects inducing/improving symptoms like freezing of gait (FoG) and slurred speech [71, 72]. Recent advances in the use of focused ultrasound has been showing significant results in slowing the disease progression. FUS guided by magnetic resonance imaging (MRI) allows mechanical energy to be delivered precisely and without craniotomy to specific targets within the brain [73, 74].

2.1.4 Motor Impairments Assessment

The assessment of the motor symptoms is tricky, with many possibilities for misdiagnosis. Zach *et al.* have discussed the entrainment and pointing test used to check for the presence of the psychogenic tremor [75]. These tests require the patients to perform simple rhythmic movements (tapping with fingers or toes). The evaluation is made based on whether the tapping frequency is lower than the tremor frequency. The “pointing test” involves asking the patient to make a rapid ballistic movement (lifting and/or moving the arms swiftly, grabbing a static or moving pen, etc.) with the limb contralateral to the one where the tremor is examined. Clinicians around the world widely use many similar tests to evaluate tremors. It is also possible that these tests can result in a false negative. There are 21 conditions where rigidity can be observed as a symptom, and PD is among them [76]. It gives more reason for the clinical diagnosis to be more subjective and a good chance for a misdiagnosis. The gait of the patient with PD is slow and on a narrow base with short steps. Freezing of gait or the festination (faster steps) can occur under different circumstances [54].

Motor and non-motor symptoms of PD have nonlinear progression with disease duration and sporadic in their order of manifestations [77, 78]. Research has been focused on evaluating the symptoms by designing methodologies for severity quantification, which can help evaluate the disease presence and disease progression [79-91]. The creation of compensatory methods to improve patient’s quality of life is another direction for symptomatic research. Studies in [92-94] focus on improving the patient's gait by using pressure sensors, a camera (image processing) and acceleration sensors, respectively. They detect irregularities in gait automatically and trigger actuation signals in the form of auditory cues [92, 93] and vibratory feedback in [94]. Wearable sensors for analysis have gained popularity in quantifying and mitigating motor symptoms [92-101]. Wearable accelerometers are used in [97] and [98] to study and attempt to automatically stabilize the gait by syncing the gait with the tempo of the auditory cues. Advanced studies analyzing the EEG signals for disease classification and disease progression assessment have been showing significant progress by using novel features including higher order statistics [102, 103]. Research studies have also explored sensor fusion and compressive sensing on the gait data for extracting accurate and compact data that can significantly reduce the load of a biomedical health monitoring system [104]. The data acquisition methods in all the aforementioned symptomatic research studies are stressful on patients and their care-givers. Studies using smartphone-based systems where built-in sensors can be used for data acquisition are being developed to mitigate the stress involved in data acquisition. They can be implemented remotely without causing any discomfort to the patient [105-109]. The control over data acquisition in these studies is minimal. Due to the use of smartphones, the collected data is highly processed, and in most cases, it is impossible to tap into the smartphones and gain access to the raw data.

2.2 Parkinsonian Speech

Often, improving the quality of life for patients and care-givers requires healthy and regular social communication and interaction. Vocal communication has a huge role in the maintenance of healthy social relationships and interactions. The deficits in paralinguistics of human speech, such as rate, volume, pitch, and pause duration, correlate with the development of depression [110].

Speech production starts in the brain with a set of commands that directs various organs to execute an elaborate coordinated plan designed to produce an utterance. Hence, speech production requires a phonetic plan and a motor plan [111]. When a person is affected by a neurological disorder like PD, it can affect speech production due to the imprecise execution of the brain's motor plan. In advanced stages of PD, the effect can be heard and perceived, and in the early stages, it is hard for the naked ear to pick up the fluctuations often due to compensatory mechanisms. When a neurological disorder affects articulators of speech (vocal cords, jaw, lips, tongue) at different levels due to motor impairment, their collective impact can affect the produced speech [112]. Hence, speech can be used as a contributing feature in diagnosing and detecting the disorder's level of progression if known to affect the speech [113]. Compared to the research using motor symptoms, speech processing has a few pros and cons. The pros of speech analysis for PD are:

- *Non-Invasive*: When compared to the usual methods of diagnosis like the blood work or the recording of the electrical signals like the ECG, EEG, etc. or any other process, obtaining speech samples is entirely non-invasive and causes minimal discomfort to the subjects.

- *Cost-efficient*: Compared to the conventional methods, the cost for setting up the facility to record the speech is meager. Typically, it requires a noise-free environment, a decent microphone and a digital recording setup. On the whole, the equipment and microphone are massively less expensive.

- *Time-efficient*: Compared to the time required for the extraction of data using the neuroimaging techniques like the EEG, PET or SPECT, which can be 60 minutes to 3 hours, the time required to prep the subject and extract speech samples can be 10 min to 30 min. Depending on the level of progression and the efficiency of motor functioning of the patient, more time may be required to prepare the patient. However, the actual time required to obtain the recordings is significantly less.

- *Low stress*: PD patients do not need to wear sensors or perform a battery of motor functions or go through a lengthy procedure that takes hours to give the speech samples. It is a relatively straightforward procedure to follow for the subject and the person recording the samples. In case of any erroneous recordings, repeating the recording process does not take long, unlike most other sensor-based data collection.

- *Portability*: Even though it is ideal for collecting the speech samples in a sound-proof recording room, it is relatively easier to develop a portable speech recording system for data collection. Portable systems can help obtain speech samples from patients with difficulties going to the recording center. At a basic level, it requires a laptop, a microphone and portable isolation shield for creating noise-free environment. This equipment is easy to transport and setup anywhere for data collection.

In addition to the pros, speech analysis also comes with some cons that have to be taken care while designing the protocols. Some of these cons are:

- *Pre-existing health conditions*: This issue is not unique to speech as other motor symptoms could have etiologies other than PD. Suppose the subject being examined for language or voice disorders happens to suffer from other pre-existing conditions affecting speech production. In that case, it can impact the analysis and result in bias over classifiers. Inspection for the pre-existing conditions before sample collection can mitigate this issue.

- *Patient history*: Like pre-existing conditions, the habitual and medical history can derail the analysis. For example, a person with chronic addiction to smoking is expected to have dysfunctional vocal cords, impacting phonation. Hence the speech sample analysis in such cases cannot result in reliable decision making.

2.2.1 Speech Impairments Due to Parkinson’s Disease

Literature shows that one of the manifestations of motor and non-motor dysfunction due to PD is in the form of speech impairments like dysarthria (slurred and slow speech), dysphonia (poor voice quality), dysprosody (variations in melody, intonation, pauses, stresses, intensity, vocal quality), monopitch, reduced speech rate, reduced and monotonous loudness, articulatory imperfections, inappropriate silences, hoarseness and rushed speech [114, 115]. Approximately 90% of patients with PD are known to have speech disorders due to PD [116], which impacts the social well-being of the patients. Hence it is essential to study the effects of PD on speech both for quantification and quality improvement.

Compared to the research over other motor symptoms, the research over speech samples as a diagnostic aid is still at the initial stages and not enough for real-time use. From the pros and cons and a multitude of possible impairments, it can be deduced that disease classification applications can benefit from extensive research on speech analysis.

2.2.2 State-of-the-Art Research Using Acoustic Methods

Following [117], many published works in engineering and technology focus on using speech as a diagnostic aid for detecting PD. Besides the commands being generated in the brain, speech requires coordination of respiratory, phonatory, articulatory and resonatory subsystems. Motor effects of PD can cause impairments in these subsystems, which can be perceived

As mentioned in Chapter 1, studies using PD speech can be broadly divided into perceptual studies and analytical studies. Perceptual studies focus on ratings from trained and experienced listeners to quantify various symptoms observed in the speech samples. Analytical assessments based on signal processing and machine learning algorithms focus on using features extracted from the speech samples to train models for automatic classification.

While perceptual studies have the advantage of using human knowledge and intelligence, they suffer from the subjectivity of the listener and rely upon the usage of multiple raters to address inter-rater variances, which increases the experimental costs. Analytical assessments address the subjectivity issue using an automatic classifier. However, it is often limited by the computational complexity and lacks human intuition and intelligence to process speech samples with variability in the content being spoken. Hence, the development of objective methods for PD classification using speech needs to be brought closer to perceptual studies using intelligent control over speech content used for analysis.

2.2.2.1 Speech Tasks

A significant share of the data used in the research is sustained vowel phonation [22, 34, 118-120]. The subjects are asked to sustain a vowel phonation with the vowel choice and utterance duration variations. PD patients have restricted mobility in the articulators causing the formant centralization [121]. This effect makes sustained vowel phonations interesting for PD speech analysis, and many studies tried to take advantage of this effect.

Next to sustained vowel phonations, other types of voice recordings include diadochokinetic (DDK) task – repeating the sequence of syllables /pa/, /ta/ and /ka/ and utterances of a set of clinically selected words. DDK task combines bilabial, alveolar and velar places of articulation and hence affected more by PD [122, 123], clinically selected words proven to bring out the most functioning of the vocal tract [124, 125]. Some authors argue that the sustained vowel phonation or the DDK task makes the patient more conscious, and hence the recordings are not natural, which brings the usage of the sentences and monologues used in [125-128]. The justification for that argument is that even if the phonations are made under a controlled environment, the patient cannot completely shut the effect of the PD.

2.2.2.2 Analysis Methodology Using Signal Processing

Speech analysis methods for classification using acoustics methods can be broken down into three sequential steps as shown in Figure 2-1: Pre-processing, Feature extraction, and

Classifier training. Depending on the application, each step will be modified; for example, not all features work for every application involving speech. Recognition-based applications predominantly use MFCCs in the feature extraction step.

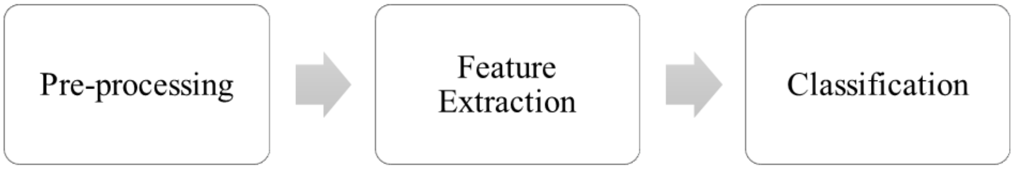


Figure 2-1 Flowchart showing classification procedure

Pre-processing the speech samples is carried out to make sure the samples fit the requirements of the subsequent processing steps. Signal processing methods have the essential stationarity requirement in the signal being analyzed. Signal stationarity means that the statistical properties of the signal do not vary. Hence, any patterns of behavior observed in the signal can be useful in analysis and drawing inferences. In speech processing, the stationarity assumption would only be satisfied over short segments. These segments are called blocks with durations typically ranging between 20 ms and 50 ms. This method of processing the speech signals segment-by-segment is called block processing. For PD speech analysis, the block size adopted by many research studies is 25 ms. Acoustic features are extracted from each block and used for analysis. These acoustic features represent the physiological shortcomings in various areas of speech production for each block. Classifiers trained over these features have better performance if they can effectively capture the perturbations due to PD within these segments.

This study used pitch synchronous segmentation methodology instead of bock processing fixed window (typically of 25 ms) based segmentation protocol for analysis. We address the spectral distortion and smoothing problems known in speech research for a very long time with this change. With pitch synchronous segmentation, each pitch cycle can be processed individually.

This helps in mitigating the spectral smoothing caused due to the inclusion of multiple pitch cycles in the analysis window of block processing, thereby providing better estimates of the spectrum and improving feature computations [129]. This segmentation method has a significant impact in improving the spectral estimates of speech with low fundamental frequency, often observed in PD speech from males.

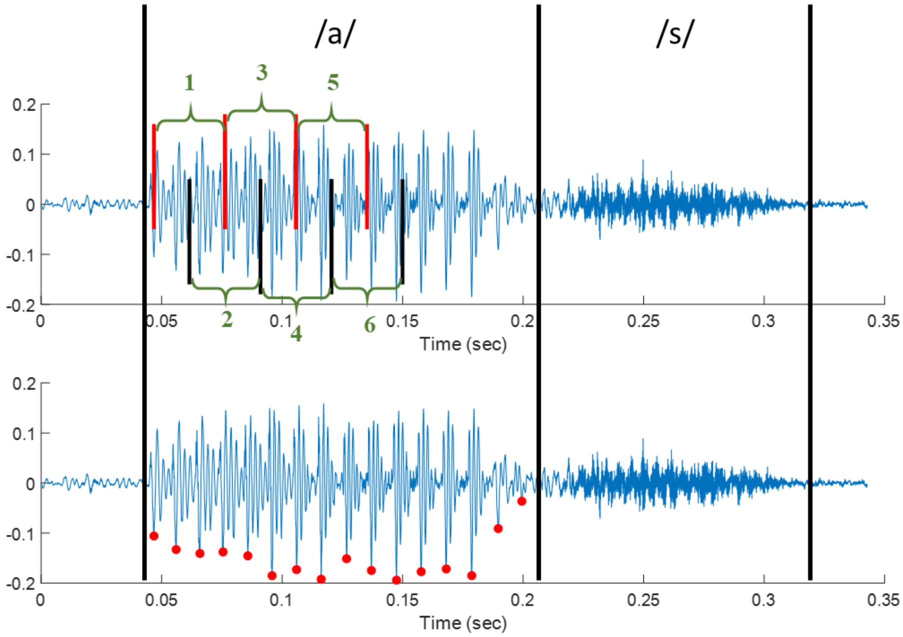


Figure 2-2 Block and pitch synchronous segments

Block processing (top) and pitch synchronous (bottom) segmentation methods are shown in Figure 2-2. In block processing, the blocks are overlapped by 50%, as shown by the block indices in green. Pitch synchronous segmentation can only be applied to voiced portions of speech where the vocal cords vibrate. The waveform shown in Figure 2-2 is the utterance of the word 'Gas' containing voiced /a/ sound and unvoiced /s/ sound.

Feature extraction is carried out over the short-time segments created after pre-processing. The features extracted from the speech samples and used for classification are selected by different researchers based on different criteria. However, some features are used extensively in almost all the studies. A list of such features from the time-domain, with the description and reason behind their consideration, is given below:

- *Pitch and Jitter*: Fundamental frequency (F0) is the closest acoustic feature to the vocal pitch of a person. Due to PD, pitch variations or lack thereof is significant compared to the F0 variations in healthy speech. Variations in pitch are extracted using when compared to healthy speech, are quantified using jitter.
- *Pitch period*: Pitch period is the time-domain version of the pitch, used as frequently as the pitch. Though intuitively, pitch and pitch period are supposed to have a similar effect on performance, many studies use pitch period and pitch in the list of acoustic features [130].
- *Amplitude perturbations (Shimmer)*: Shimmer measures the amplitude variations in the speech signal. Shimmer helps quantify the variations in pressure perceived by the microphone caused due to the irregularities in vocal fold vibrations.
- *Harmonics energy*: Asgari and Shafran mentioned in their harmonic Model of voiced speech in [122] that the glottal pulse sequence in the human speech production system is rich in harmonics. The harmonic structure observed in voiced portions has variations in energy distribution between PD speech and healthy speech due to pitch perturbations. As the harmonicity of each voice is distinct, normalized measures like harmonics to noise (HNR) or the noise to the harmonics (NHR) [122, 130-134] or harmonic coefficients [122] are used as features.
- *Formant frequencies*: Formants are the peaks observed when the frequency spectrum of the speech is observed. While many studies used the formants directly for analysis [112, 132, 135-140], various sub-features are calculated using formants. The frequently used features that are derived using the formants are:

- Formant Centralization Ratio (FCR): The dysarthria due to PD causes centralization of the formants, which can be quantified using equation (1), where

$$FCR = \frac{\mathrm{F1/i/ + F1/u/ + F2/u/ + F2/a/}}{\mathrm{F2/i/ + F1/a/}} \qquad (1)$$

  Where F1/i/ is the first formant for the vowel /i/ and F2/u/ is the second formant of the vowel /u/ and so on [141]. It is used in [112, 135, 136, 138, 140]

- Vowel Articulation Index (VAI): VAI is the reciprocal of the FCR, which also quantifies the formant centralization [112, 135, 136].
- Vowel Space Area (VSA): VSA is the area of the triangle in the formant subspace (first and second formants) using three vowels /a/, /i/ and /u/. Due to vocal impairments, the area of this triangle is expected to be reduced for PD patients [135, 136, 138, 140].
- F2 Ratio: It is the ratio between the second formant of the vowels /i/ and /u/ [112, 135, 136, 140].

In addition to these, some research studies also used the spectral centroid [126, 142], spectral entropy [122, 143], F1 vs F2-F1 plots [144], semitones [140].

- *Cepstral features*: MFCCs are the most widely used cepstral features for acoustic analysis. They are estimated by applying a cosine transform on discrete Fourier transform (DFT) bins grouped according to a Mel frequency scale (a logarithmic scale that is close approximation of human auditory system) [137]. It is one of the most frequently used features for PD classification using speech [34, 122, 124, 127, 132, 134, 137, 139, 143, 145-148].
- *Noise and Energy estimates*: In addition to HNR and NHR, other noise-based features are used for PD speech to quantify the hoarseness and breathy speech. Glottal to Noise

Excitation (GNE), Normalized Noise Energy (NNE) [131, 132] and Short-Time (ST) Energy [124, 137, 142].

Those are some of the important features used in the attempts to create a system that can aid clinical decision-making. Besides grouping the features based on their fundamental feature (eg: 'pitch' for the jitter, mean jitter and the maximum and minimum values) and linearity, one other type of grouping is based on muscular dynamics and functioning [136].

In addition to these most frequently used features, some features have been used less frequently, but they were reported to have a considerable impact on classification performance. They are:

- *Pulse features*: Pulse or *glottal excitation pulse* features include features computed over the glottal waveform estimated from the voiced portions [149, 150]. These features were used to detect any significant impact of PD over the voicing function isolated from the effects of the vocal tract.
- *Nonlinear features*: Alongside the features mentioned above, few studies have used various nonlinear features. Though they are not prevalent for analyzing PD speech, they are included in this list due to their classification impact reported in the respective studies [132, 151, 152]. The definitions of these features are not very clear in the literature and most of them have been used for different applications and different kinds of data.
    - Correlation dimension (D2) and largest Lyapunov exponent: The first two features are constructed based on the concepts from chaotic systems where speech is treated as a multi-dimensional nonlinear process. D2 is the measure of the level of multi-dimensionality.

- Hurst exponent: This measures long-term dependencies among various variables involved in time-series analysis of speech.
- Lempel-Ziv complexity: Measure of the data complexity.
- Recurrence Period Density Entropy (RPDE) and Detrended Fluctuation Analysis: These two features are entropy measures used to quantify the deterministic and stochastic components of the speech signal. Out of all non-linear features listed here, only RPDE has significant reuse in other studies using PD speech.

2.2.2.3 Machine Learning Methods

Kernel-based Support Vector Machines (SVM) and k-NN (K Nearest Neighbor) are the two most commonly used classifiers for acoustic studies on PD speech. The different kernels used for SVMs are Gaussian/Radial Basis kernel [34, 41, 112, 132, 137, 143, 151-155], sigmoid [139, 143, 145, 156], polynomial [122, 151, 155], linear [148, 155], and Multi-Layer Perceptron (MLP) [155-157]. The k-NN classifiers are used in [131, 150, 154, 156, 158, 159]. They depend on the closeness of samples in feature space and groups based on their Euclidean distance in feature space. The sample under examination is classified into the class to which most of its k nearest neighbors belong. Previous studies also compared the effectiveness of various classifier types for different applications [42].

Many other classifiers have been used and most of the studies using new classifiers appear to be experimental with minimal information over the suitability of the classifiers for PD classification. However, SVMs with Gaussian kernels have been reported to perform superior to other classifiers in many studies. The list of ML models and stuies used for this PD classification using speech is provided in Table 2-1. It is common in research to use datasets containing features and labels to explore the classification algorithms to obtain the best performance. Many studies using acoustic features for classification have followed a similar approach and used various datasets from the University of California, Irvine (UCI) machine learning repository [157, 159-164].

Table 2-1 Classifiers used for PD speech analysis

Classifier	**Studies**
Linear discriminant analysis (LDA)	[128, 130, 165]
Classification And Regression Trees (CART)	[136, 166]
Random Forests and Random Trees	[127, 136, 163]
Naïve Bayes	[127, 149]
Feed forward Back-propagation based Artificial Neural Network (FBANN)	[163]
Bayesian Network	[156]
Sequential Forward Feature Selections (SFFS)	[135]
Adaboost	[159]
Stacked generalization and complementary Neural Networks	[164]

In addition to classical ML models, advanced neural network architectures are also being studied for PD classification using speech. One of the earlier studies using artificial neural networks (ANN) for classification integrated fuzzy logic and linguistic hedges [160]. When the datasets became more extensive with the integration of data from multiple sources, multi-layer perceptron models were used to identify the optimal number of hidden layers [157, 162]. Different variants of ANN like Feedforward Back-propagation ANN, Adaptive Resonance Theory and Kohonen Neural Network (ART-KNN), CompleMenTary Neural Networks (CMTNN) have been used [159, 163, 164].

More recent studies have used several deep neural network (DNN) architectures for PD classification using speech. When large datasets containing raw samples are available, DNN architectures can be trained on them directly, which helps in learning relevant features on their own based on cost function description. Some studies used features extracted from speech data rather than raw speech samples to train DNNs expecting better classification [167-169]. Fewer studies have used raw speech samples to train DNN models mainly due to the data constraints. Unlike the feature sets, raw data is not available easily for the studies. Marek *et al.* and Juan *et al.* have trained convolutional neural networks (CNN), traditionally used for image classification applications, on magnitude spectra of speech samples to spectral the to [170, 171]. They have used segments of predefined length from sustained phonations and connected speech and used the spectrograms (time-frequency representation) of these segments as inputs to CNN to train the classifiers. Long-short term memory (LSTM) and recurrent neural networks (RNN) have been used for applications where data is sequential. PD speech has also been used with these advanced neural network models as it is sequential and holds much information in the form of dynamic perturbations. LSTMs also work with data of variable length, which is typical for speech-based applications. Jhansi *et al.* have combined the LSTMs and CNNs and devised CNN-LSTM models. They segmented the speech samples using fixed window length (block processing) and used the spectrogram of these individual segments as inputs to the CNN-LSTM model [26]. Like the spectrograms, Danish Raza *et al.* have used the features extracted from each segment of the speech sample to train an LSTM model [172].

A high-level comparison of these studies shows that acoustic studies using features are more prevalent than DNN based studies due to the lack of data. Recent studies have been using transfer learning approaches to train the DNN models that have already been trained originally with similar data for different applications [26, 170]. Jhansi *et al.* [26] showed the advantage of using connected speech over the established sustained phonations with sequential learning models with better classification performance.

2.3 Connected Speech in Parkinson's Disease

Speech task used for automatic acoustic analysis has predominantly been sustained phonations [22, 34, 119]. Fewer studies have been focused on the automatic analysis of connected speech. Rusz *et al.*, and Orozco *et al.*, have focused on connected speech for PD classification. In [173, 174] Rusz *et al*. used sustained phonations, DDK and connected speech tasks to conduct overall analysis targeting the classification problem from all three directions (phonatory, articulatory and prosodic studies). They used spectral measures like formants, vowel space area (VSA) and vowel articulation index (VAI) to classify PD and healthy controls. For prosodic analysis, they have examined syllabic level measures like fundamental frequency (F0), vowel onset time (VOT), intensity (relative loudness of speech), articulation rate, pause characteristics, and rhythm. Their analysis outcomes state that the prosodic analysis has the best chance of classification followed by articulatory analysis and connected speech is more suitable for classification than sustained phonations. In [175], Rusz *et al.* have studied DDKs using articulatory features like vowel similarity quotient, vowel variability quotient, VOT, formants and DDK rate. They discussed the importance of articulatory imprecision in PD and how it can classify PD from HC.

In [176] Orozco *et al.* used word utterances from three languages along with DDKs to conduct articulatory analysis using MFCCs and bark band energies (BBE) extracted from voiced and unvoiced portions individually. They reported a maximum of 99% classification accuracy using the unvoiced components of DDK tasks and 90% accuracy using voiced components. With isolated word utterances they achieved 84% to 96% classification accuracy. In [118], they have shown similar performance outcomes with isolated words, DDK, sentences, and reading text. MFCCs, noise measures, formant measures and BBEs are used to analyze both [118] and [176].

Skodda *et al.* have focused on prosodic analysis using F0, formants and speech rate computed with syllable and pause durations [177]. In [178], they used MFCCs for articulatory analysis; F0, energy, syllable durations, pause durations, jitter and shimmer with connected speech for prosodic analysis; glottis parameters for phonatory analysis. They reported that articulatory analysis using connected speech tasks has better performance in classification between PD and HC.

2.4 Summary

Research community has two schools of thought: 1. Sustained phonations are adequate for PD classification, 2. Connected speech and DDK have more vocalic information and hence more suitable for PD speech analysis than sustained phonations. Advanced classification methods applied to connected speech, DDK and sustained phonations show that connected speech has superior performance [26]. This adds to the existing argument which says connected speech has more information that helps in PD classification.

Articulatory and phonatory analyses are conducted predominantly using spectral and temporal measures typically extracted using toolkits like openSMILE [179] and VOICEBOX [180]. It is important to note that these features are typically extracted using block processing protocol where the speech is segmented using a 25 ms fixed window. Very early on in speech research, it has been established that block processing results in spectral smoothing [181]. Except for jitter, shimmer, whose extraction method is varied through different studies, and glottal parameters, the rest of the features are extracted using block processing over a fixed window. PD classification can benefit when features extracted can better capture the vocalic dynamics.

This dissertation examines pitch synchronous segmentation to address the smoothing effects due to block processing and a novel feature set to quantify the variations in speech between every two consecutive vocal fold closure incidents.

Chapter 3: Analysis Using Pitch Synchronous Segmentation and Novel Feature Set

3.1 Pitch Synchronous Analysis

Typical speech analysis follows a block processing approach where the block size can vary between 20 ms and 50 ms depending on the application. These blocks are known to retain the necessary statistical stationarity in data. In pitch synchronous (PS) segmentation, the pitch cycles in voiced speech segments are extracted and processed. Figure 2-2 in Chapter 1 shows a speech sample segmented using block processing (top) with 25 ms window size and 50% overlap. The fundamental frequency of speech is typically higher for females than males, which means that, within a fixed duration, the vocal activity is higher for females than males [182]. A 25 ms segment of the voiced portion sample from the male and female speech is given in Figure 3-1. It can be noticed that within the same time window, a male voice produces almost half as many pitch cycles (repeating patterns) as a female voice.

While features extracted from these blocks of fixed length represent the effect from multiple cycles, they must address varying amounts of vocalic activity for different speakers and contextual effects. Especially for phonatory analysis, where the cycle-to-cycle perturbations in vocal tract manifest as the symptoms, block processing falls short in bringing out the perturbations. PS segmentation preserves the information between every two vocal fold closures, thereby improving feature computation required for phonatory analysis.

The issue with PS segmentation of speech that can be used for different applications is the complexity of the voiced speech. The cycle length varies subtly between the cycles, and segmentation becomes very complex during vowel transitions. Previous works involving estimation pitch cycles adopt computationally expensive methods and/or get affected by varying noise content.

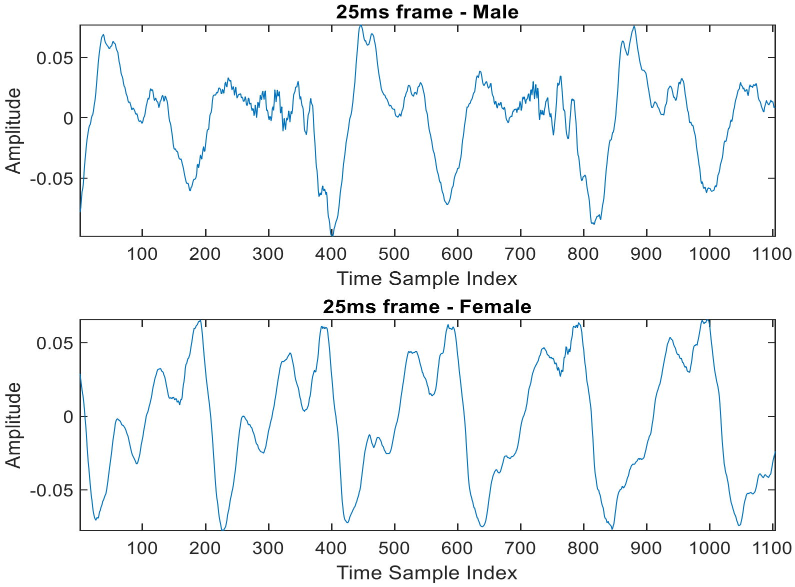


Figure 3-1 Speech sample of 25 ms duration in male (top) and female (bottom)

3.2 Automatic Segmentation Framework

An automatic segmentation framework has been developed for PS segmentation of the connected speech samples [38]. It is implemented in two stages: cycle endpoint estimation and error correction. The cycle endpoint estimation is performed coarsely and then fine-tuned to improve accuracy. The speech signal is initially segmented into 25 ms frames with a 50% overlap. From each segment, the fundamental frequency is estimated using a multi-stage Fourier transform. In this framework, fundamental frequency estimation is implemented using peak-picking from auto-correlation magnitude spectrum from second stage Fourier transform. The bin index and lag index of these peaks are very close in auto-correlation and multi-stage magnitude spectrum, as shown in Figure 3-2 (red peaks are at 408th lag and 392nd bin and black peaks are at 209th lag and 220th bin).

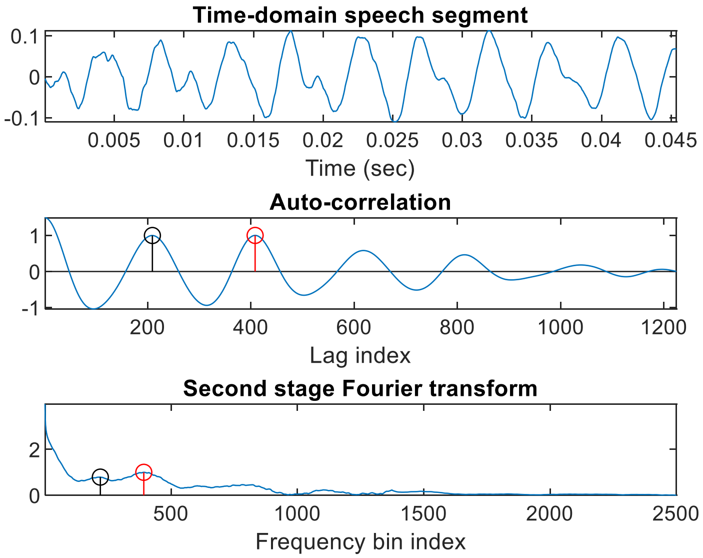


Figure 3-2 Periodicity in speech segment represented in speech waveform (top) estimated using auto-correlation (middle) and a two stage Fourier transform (bottom)

For speech segment s(n) the Fourier transform is computed with N1 bins in the first stage and N2 bins in the second stage, as mentioned in the below equations.

$$s(n) \xrightarrow{\text{FFT}} S_2(f), \quad N1\ bins \tag{2}$$

$$|S_2(f)| \xrightarrow{\text{FFT}} S_2(f), \quad N2\ bins \tag{3}$$

For a fundamental frequency F0, the second stage Fourier transform's peak location can be computed using the below equation.

$$F_0 = fs \times \left(\frac{N2}{N1}\right) \times \left(\frac{1}{P}\right) \tag{4}$$

$$fs = Sampling\ frequency, P_{loc} = Peak\ location$$

Among many peaks detected in the second-stage Fourier transform, the location of peak with maximum harmonic energy will be used for the ‘P’ in equation (4). The fundamental frequency for all 25 ms segments is estimated using this method. Their mean value is used as the reference in cycle endpoint detection, where pitch cycles are finely segmented by precisely identifying endpoints. The algorithm used for cycle endpoint estimation is given below:

Cycle endpoint estimation algorithm

```
Input: speech signal, $F_0$ estimates along with start and end locations of each 25ms
        segment, average $F_0$ value and minima location in first segment.
while speech segment of length $1/F_0$ is available
        extract a frame of size 3 times $(1/F_0)$
        if difference between previous cycle length and reference is beyond the
                limits, estimate auto-correlation over a sequence of 2*reference
        else  estimate auto-correlation over a sequence of length
                2*previous cycle length
        extract Peak value and peak location
        eliminate peaks far reference
        if all peaks are eliminated, select the peak with most prominence
        else Select that peak whose location is closest to reference
        use the peak location to identify next cycle end point
        save the minima close to the sample identified in previous step
end
```

After the coarse estimation of cycle endpoints, offset and cycle length (doubling and halving errors) errors are identified and corrected using Hilbert transforms. Figure 3-3 shows a pitch cycle with partial sections of its adjacent cycles on both ends. The peaks of Hilbert transform of a segment of the speech from the waveform shown in Figure 3-3 will have a maximum value only when the segment ends align with pitch cycle ends marked by the red vertical lines. Hence, a small offset in the cycle ends can be identified and corrected automatically by examining the Hilbert transforms.

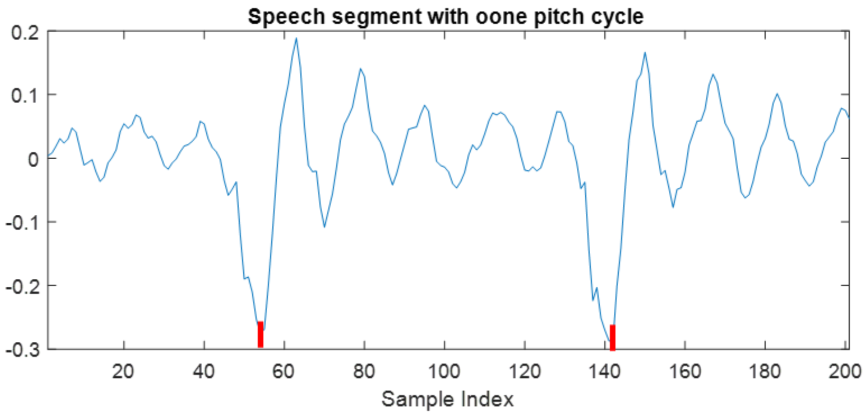


Figure 3-3 Speech segment with one pitch cycle including partial sections of adjacent cycles

The pitch cycle length identified across the speech sample is passed through a fifth-order median filter to identify doubling and halving errors. Cycles resulting in a positive error are checked for doubling by comparing against the local mean. If the ratio between the cycle length and local mean is ≈2, it is treated as a doubling error and appropriately breaks it down into two parts. The estimated pitch cycles with cycle length to local mean F_0 ratio close to 0.5 are identified to be halving errors. The offset and cycle length corrections are performed recurrently. The performance variation is quantified using the number of doubling and halving errors identified from median filtering after each iteration. The iterative process is stopped when no significant improvement is observed.

3.3 Analysis Framework

The novel analysis framework proposed in this dissertation uses covariances within the features extracted from the pitch synchronous segments of the voiced sections to train classifiers for PD detection. This framework differs from existing methods in three main domains. Firstly, the pre-processing is based on PS segmentation instead of block processing and uses voiced segments bordered by unvoiced/silent portions on both ends. Second, novel pitch synchronous features (PSFs) were extracted from the PS segments instead of popular features mentioned in the previous sections. Third, a covariance-based feature transformation is developed to group the 1st order differences in PSFs from each voiced segment of the connected speech identified in pre-processing. With variations at multiple stages of analysis compared to established methods, the proposed framework's evaluation is planned to address the variations at each stage as explained in the later sections.

3.3.1 Speech Databases

The framework proposed in this dissertation is evaluated using two different datasets. Usage of multiple databases helps verify the reproducibility of results and comprehend the effect of dataset size over classification performance. Connected speech is evaluated using passages in Italian (Database 1) and English (Database 2) languages read by people with PD and healthy controls. Both Databases also contained sustained vowel phonations which were used for evaluating the existing methods. The database descriptions are as follows:

Database 1 was collected by Giovanni et al. and analyzed using an automatic Speech-to-Text system. It was acquired from IEEE DataPort [183]. It contains two phonetically balanced Italian passages read by 50 subjects with 28 PD (19 male and 9 female) and 22 HC (10 male and 12 female). According to the authors of [184], none of the patients reported speech or language disorders unrelated to their PD symptoms before their study and were receiving antiparkinsonian treatment. The HY scale ratings were < 4 for all the patients except for 2 patients with stage 4 and 1 patient with stage 5. The database consisted of different speech tasks recorded from the participants, but we included only the passage readings in this analysis. The recordings were made at a sampling rate of 44.1 kHz with 16 bits/sample. The duration of each passage recording varied between 1 to 4 minutes, with a mean of 1.3 minutes. The recordings were performed in an echo-free room with the distance between speaker and microphone varying between 15 and 25 cm.

Database 2 was selected from a larger database collected at the Movement Disorders Center, University of Florida [185]. It consists of speech tasks such as passage reading ("The Rainbow Passage", Fairbanks, 1940) and sustained vowel phonations of which only the former was used in this study. This dataset is more balanced than Database 1, with ten age and gender-matched data in both PD and HC classes. Recordings were made using a Marantz portable recorder (Marantz America, LLC, Mahwah, NJ) and stored digitally with a sampling rate of either 44.1 kHz or 22.05 kHz and 16 bits/sample. The institutional review board approved all test procedures at the University of Florida, and testing was completed following an informed consent process. The duration of recordings from this database varied between 25 and 90 sec with a mean of 41 sec.

3.3.2 Feature Set Description

The features developed for PS analysis have been analyzed in previous studies [39, 40]. As the proposed framework includes normalization step, the features extracted from normalized pitch synchronous segments were not included to avoid redundancies. Details of the novel PSFs along with their descriptions are as follows:

- *Pitch Period*: The duration span of each PS segment computed in seconds.
- *Length of Curve*: The sum of the Euclidean distances between every two consecutive time samples in the data segment.
- *Total Energy*: Total energy of the PS segment.
- *Quarter Segment Energy:* Each PS segment is divided into four parts of equal duration span. Total energy in each of those parts pooled together forms the four features in quarter segment energy.
- *Correlation Canceller Efficiency:* Starting from the second pitch cycle in the super segment, the estimation error (e_n) between a data segment (y_n) and its correlation (with its

previous segment) correlation canceller estimate ($\hat{y}_n$) is computed. Then, the expectation of estimation error and data segment is used as the feature output.

$$e_n = \; y_n - \hat{y}_n \;\;,\;\; n \geq 2 \tag{5}$$

$$\hat{y}_n = a\, y_{n-1} \;\;,\;\; a = \frac{R_{yy}(1)}{R_{yy}(0)} \tag{6}$$

$R_{yy}(k)$ is the autocorrelation sequence of y_n and k is the lag.

$$\text{Correlation Canceller Efficiency (CCE)} = \mathrm{E}[e_n \, y_n] \tag{7}$$

- *Correlation Canceller Mean Square Error*: This is the mean squared estimation error (e_n) computed in the feature correlation canceller efficiency (CCE).
- *Peak Frequency*: This is the frequency in hertz where the maximum magnitude spectrum output can be observed. (f_N is the Nyquist frequency)

$$y(n) \xrightarrow{\text{FFT}} Y(f) \tag{8}$$

$$\text{Peak Frequency} \;=\; \{f_1 :\; |\mathrm{Y}(f_1)| \geq\; |\mathrm{Y(f)}|, \mathrm{f} \in [0, f_N]\} \tag{9}$$

- *Quarter Band Magnitude*: The magnitude spectrum of the segment is divided into four equal bands between 0 Hz and f_n Hz, and the total magnitude in each portion is extracted.

$$M(n) = \sum_{f=\left(\frac{n-1}{4}\right)f_N}^{\left(\frac{n}{4}\right)f_N} |Y(f)| \quad n = 1 \; to \; 4, \tag{10}$$

- *Spectral Factor*: This is the ratio between the energy of high frequency components to the total energy in the segment.

$$F = \frac{mean(|Y(f_1)|^2)}{mean(|Y(f)|^2)} \quad , \tag{11}$$

$$f_1 = \left(\frac{fs}{5}\right) Hz \; to \; Nyquist \; frequency$$

$$fs = Sampling\; frequency, f = 0 \; Hz \; to \; Nyquist \; frequency$$

On the whole, 15 PSFs were extracted from each PS segment of the voiced speech. Only quarter segment energy and quarter band magnitude return four elements, and the rest of the features are singular values.

3.3.3 Classifiers

An ensemble of 17 classifiers, available in MATLAB, were used for classification experiments in this study. It included variants of Trees, k-Nearest Neighbors (KNNs), Support Vector Machines (SVMs), Ensemble learners and logistic regression. Using an ensemble of classifiers, the suitability of different classifiers for this application was evaluated with unbiased decision-making. The organization of all the members in these groups is shown in Figure 3-4. All the classifiers were used with their default hyperparameter settings provided by MATLAB's Classification Learner Application.

3.4 Evaluation Methodology

In this section, the methodology evaluation steps designed for the framework discussed. The alternatives for various steps from the established protocols are mentioned along with the evaluation procedures.

- *Pre-processing*: This includes the voiced/unvoiced classification followed by PS segmentation of voiced segments. To emphasize the importance of PS segmentation, block processing based analysis with 25 ms window will be carried out and classification performance comparisons will be made between both segmentation methods. Mel-Frequency Cepstral Coefficients (MFCCs) were used for comparing the segmentation protocols.
- *Feature Extraction*: PSFs and MFCCs extracted from PS segments and will be evaluated for efficient classification in this step.

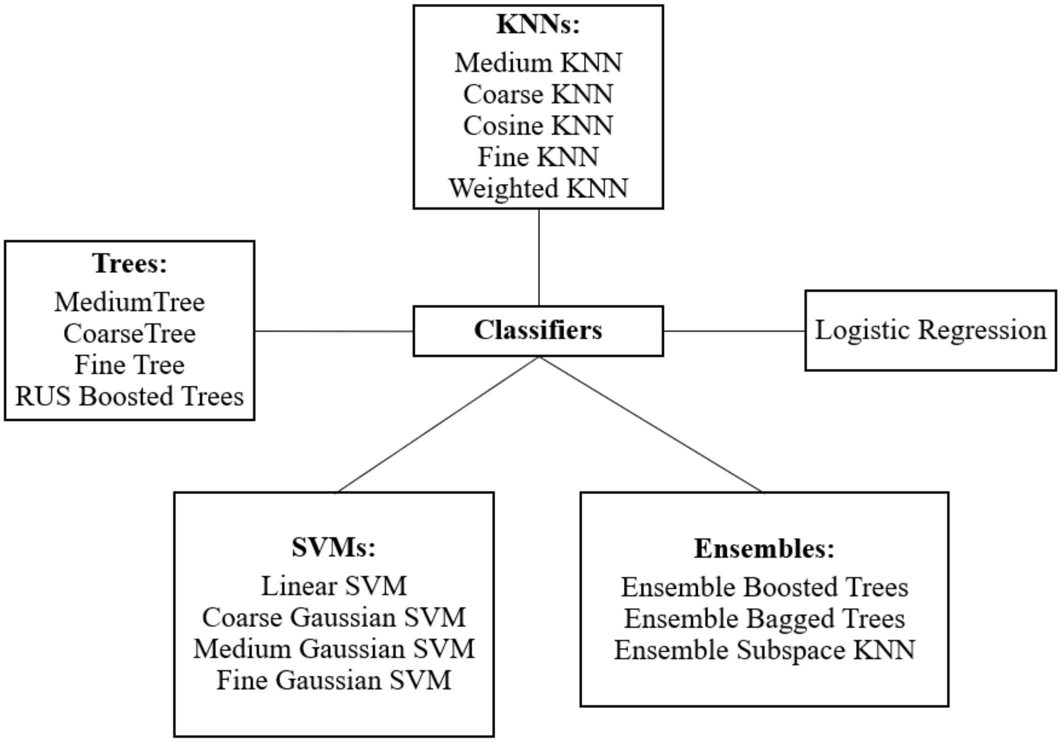


Figure 3-4 Classifiers used for analysis

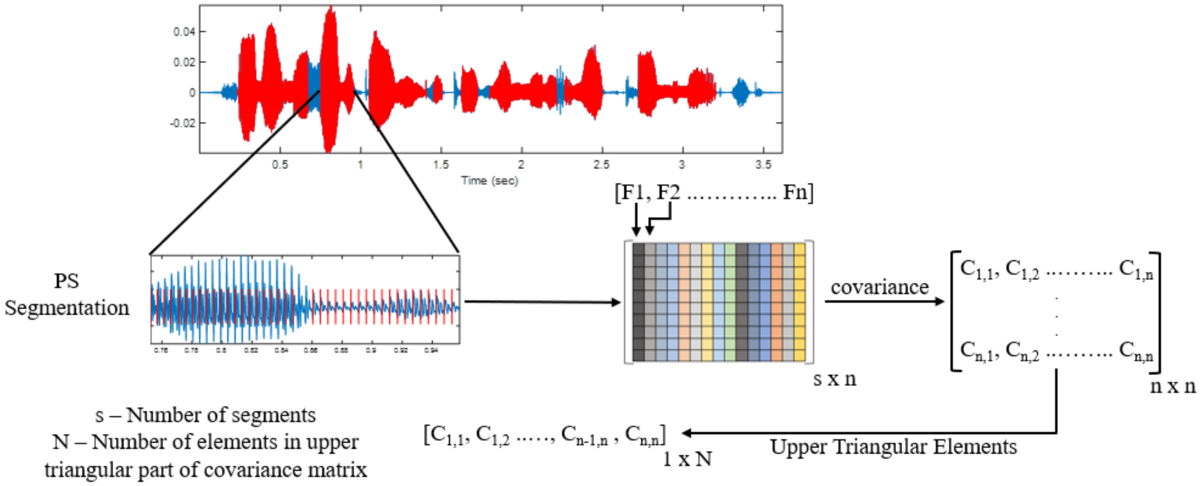

0.04
0.02
0
-0.02
0.5
1
1.5
2
2.5
3
3.5
Time (sec)
[F1, F2 ………… Fn]
PS
Segmentation
covariance
$C_{1,1}$, $C_{1,2}$ ……… $C_{1,n}$
$C_{n,1}$, $C_{n,2}$ ……… $C_{n,n}$
s x n
n x n
[$C_{1,1}$, $C_{1,2}$ ….., $C_{n-1,n}$, $C_{n,n}$]
1 x N
Upper Triangular Elements
s – Number of segments
N – Number of elements in upper triangular part of covariance matrix

Figure 3-5 Syllabic analysis protocol

- *Analysis methods*: Covariance-based analysis method proposed in this framework will be implemented as shown in Figure 3-5. After voiced and unvoiced classification, super segments were identified as the voiced components bordered by unvoiced/silent portions on both ends (segments in red in Figure 3-5). These super segments will be segmented further into blocks or PS segments. Features extracted from all segments within a super segment will be grouped, and covariances of these feature groups will be used for training classifiers.

In Database 1, a total of 12600 super segments were identified. From each paragraph reading an average of 134±24 super segments have been identified. Each super segment contained an average of 30 cycles with a standard deviation of 27 cycles. In Database 2, a total of 1357 super segments were identified. An average of 68±20 super segments has been identified from each paragraph. Each super segment contained an average of 64 cycles with a standard deviation of 58 cycles. Most of the existing methods use segmental analysis methods. The features extracted from each speech segment (block processing or PS) were treated as an individual data point while training the classifiers.

- *Raw features and 1st Order differences*: The framework also uses the 1st order differences of the features within each super segment. Comparisons from using the different variants of PSFs and MFCCs will be carried out to prove the importance of using 1st order differences over raw features before covariance computation.
- *HoldOut validation*: Training and testing datasets were created by pooling random samples in the available data in two steps due to data imbalances. First, the dataset is divided into train and test sets with 80% and 20% of the available data. Then, ‘N’ random samples from each class in the original train set were pooled to create a balanced final training data. ‘N’

was chosen to be 50% of the size of minority class in the original train set. The 50% factor was selected to preserve distributional similarity while providing enough samples for training the models. The train and test sets created using this protocol did not have any common samples and were similar in their class-wise distributions.

3.4.1 Evaluation Criteria

For evaluating each variation introduced into the established methods through the proposed framework, systematic steps will be followed as mentioned in this section.

3.4.1.1 Segmentation Method and Analysis Method Combinations

Both segmentation and analysis methods have two choices (PS and block processing; segmental and syllabic analyses), resulting in four possible cases (combinations) to be evaluated. A novel comparison technique designed to measure the classifier's efficiency in learning the effects of PD rather than identify the speakers in each class will be used. First, the classification performance was identified under standard conditions, then speakers were assigned with random PD/HC labels, and classification experiments were repeated without any other changes to the protocol. The segmentation and analysis method combination(s) good at identifying speakers in each class will deliver comparable results with original and random PD/HC labels. Hence, performance reduction magnitude will be used to compare these combination(s).

The performance reduction will be computed as per equation (12) and used as an evaluation metric in this step. The experiments were conducted using MFCCs from Database 1 for both feature variants and both genders individually using all classifiers. OA represents the accuracies with original labels and RA represents the same with random labels.

$$Relative\ Reduction = \frac{OA-RA}{OA} \times 100 \qquad (12)$$

3.4.1.2 Feature Type and Feature Variant Combinations

The performance reduction magnitude will be used to compare the feature (MFCCs or PSFs) and feature variant (raw features or 1st order differences) combinations. The evaluations will include gender dependent and gender independent cases individually.

3.4.1.3 Evaluation of Classifiers

The 17 classifiers used for experiments will be evaluated for overfitting by comparing overfit factor and test accuracies. The overfit factor between test accuracy and train accuracy will be calculated as per equation (13). It measures the relative difference between training and testing accuracies. A higher value for overfit factor indicates better performance of the classifier only on the data it has seen during training which signals overfitting to training data.

$$Overfit\ factor = \frac{Train\ Acc - Test\ Acc}{Train\ Acc} \tag{13}$$

A higher value for test accuracy with random label assignment indicates the classifier's ability to remember speakers more efficiently than learning effects due to PD. These two metrics will be calculated from the results of experiments with optimal choices identified in previous steps. They were used to identify and eliminate classifiers with a high tendency to overfit.

3.4.1.4 Testing Using Different Database

The optimal combination of segmentation, analysis method, feature type and feature variant from above mentioned criteria will be tested as the final framework. The efficacy of the final framework will be tested by training the classifiers on a larger dataset, Database 1 and testing on a different dataset, Database 2 where the speakers, language and acquisition environment were all different. Due to the differences in data acquisition conditions like the difference in equipment and variations in speaker to microphone distances, features extracted from different databases suffered differences in their numerical ranges. Normalization at the syllabic level will address these differences and maintain uniformity between the features from both databases. z-scores with zero mean and unit standard deviation will be extracted from feature data of every super segment using equation (14) and used for training and testing.

$$z_{ij} = \frac{(x_{ij} - \mu_i)}{\sigma_i} \tag{14}$$

$$x - Feature\ value, \qquad i - Feature\ Index, \qquad j - Segment\ index,$$

$$\mu - Mean, \qquad \sigma - Standard\ deviation$$

3.4.2 Instituted Method Evaluation

In addition to the proposed framework evaluations for various criteria, evaluation of the popular methods in the research is carried out using the sustained phonations available in both databases. First, steps used for classification in [34] will be retraced to evaluate the performance. Then, the proposed framework will be evaluated using the sustained phonations. In both cases, evaluations are carried out with different datasets used for training (Database 1) and testing (Database 2).

Performance comparisons will be targeted in the following directions:

i. Instituted method with sustained phonations versus proposed framework with connected speech

ii. Instituted method versus proposed framework with sustained phonations

iii. Proposed framework with connected speech versus proposed framework with sustained phonations

A total of 495 sustained phonations of five vowel sounds (/a/, /e/, /i/, /o/ and /u/) are available from Database 1 and Database 2 contained one sustained phonation of /a/ vowel from each speaker resulting in 20 phonations. From each phonation, 333 features grouped under 12 different families will be extracted using the same protocols developed by Tsanas [34] using

VOICEBOX [180]. These features will be used to train the optimal classifiers identified from the 17 classifiers evaluated using overfit factor for the proposed framework. The proposed framework will be implemented using the sustained phonations without any variations to compare connected speech and sustained phonations.

3.4.3 Performance Metrics

The classification performances will be evaluated primarily using accuracy for the first three criteria. The final framework evaluations will be evaluated using various additional metrics. Descriptions of all the metrics used for evaluations are as follows:

- *True Positives (TP)*: Number of PD samples predicted as PD
- *True Negatives (TN)*: Number of HC samples predicted as HC
- *False Positives (FP)*: Number of HC samples predicted as PD
- *False Negatives (FN)*: Number of PD samples predicted as HC
- *Accuracy*: Proportion of test samples predicted to their correct labels (PD or HC).

$$Accuracy = \frac{(\mathrm{TP+TN})}{(\mathrm{TP+TN+FP+FN})} \tag{15}$$

- *Precision (P)*: Proportion of PD predictions that were correct.

$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP+FP})} \tag{16}$$

- *Recall (R)*: Proportion of all PD samples correctly predicted.

$$Recall = \frac{TP}{(\mathrm{TP+FN})} \tag{17}$$

- *MCC*: Matthews Correlation Coefficient (MCC) is an improvement over F1-Score as it includes the TN into its computation.

$$MCC = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{18}$$

- *F1-Score*: Harmonic mean of precision and recall.

$$Accuracy = \frac{2\times\mathrm{P}\times\mathrm{R}}{(\mathrm{P}+\mathrm{R})} \tag{19}$$

- *ROC-AUC*: Area under Receiver Operating Characteristic (ROC) curve.

All the metrics except MCC have values between 0 and 1, with 1 being the best possible value. MCC can have values between -1 and 1, with 1 being the best possible value.

3.5 Results and Discussion

In this section, results for each one of the evaluation criteria mentioned earlier will be presented. Accuracies from multiple classifiers are reported using boxplots to show their distribution. Results for gender-dependent and gender-independent versions are reported individually.

3.5.1 Results From Proposed Framework

Four combinations of segmentation and analysis methods are compared to identify the optimal choice. Results from PS and block processing (25 ms block size, 50% overlap) with original labels and randomly assigned labels along with their relative differences were used. Results from MFCCs using segmental and syllabic analyses for both genders and both feature variants using holdout validation are provided in Figure 3-6. The labels on xlabels are formatted to include gender (M-Male, F-Female) and analysis (Seg-Segmental, Syll-Syllabic).

Both segmentation methods have a comparable performance for all the gender, analysis methods and feature variant combinations from these results. Between segmental and syllabic analyses, segmental analysis had better performance than syllabic analysis when raw features were used. As raw MFCCs can contain the speaker's voiceprint better than 1st order differences, the results follow the intuition and provide better accuracies with segmental analysis. When 1st order differences were used, syllabic analysis showed better performance than segmental analysis. This observation was also aligned with the intuition that vocalic perturbations have more impact due to PD than the voice and syllabic analysis protocol can better extract these perturbations.

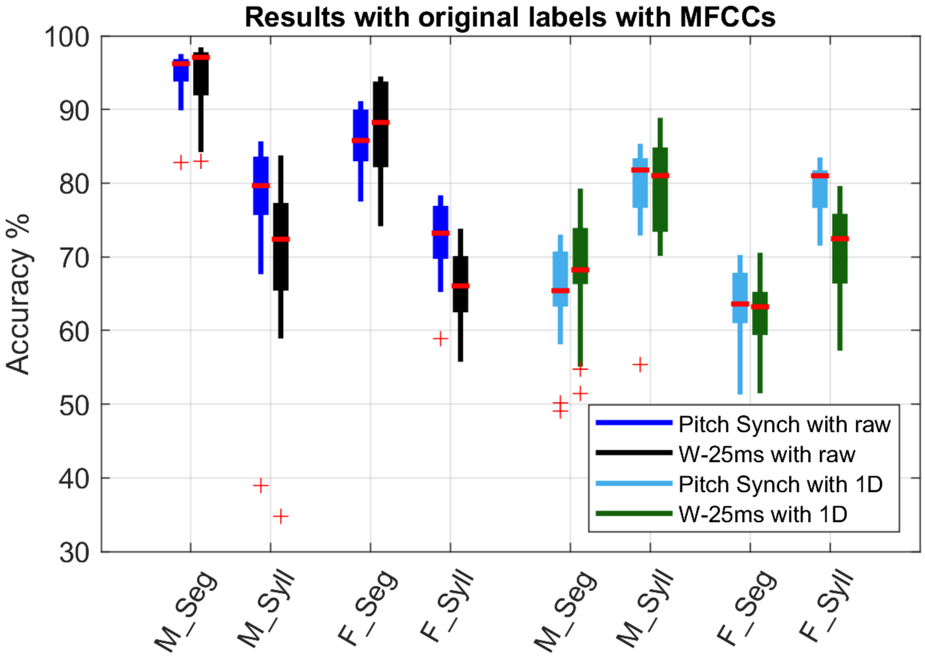


Figure 3-6 Classification results with different segmentations using MFCCs with original labels from Database 1

When classifiers were trained with random labels, performance decreased in all cases. The relative accuracy reduction due to random label assignment is shown in Figure 3-7. These results show that except for females under segmental analysis and raw MFCCs, PS segmentation resulted in comparable or higher degradation in all other cases. These results also had some negative values for percentage decrease, suggesting better classification with randomized labels. Inspection revealed that classifiers with fine kernels were responsible for such effects. These classifiers were also prone to overfitting problems, as discussed in the subsequent sections.

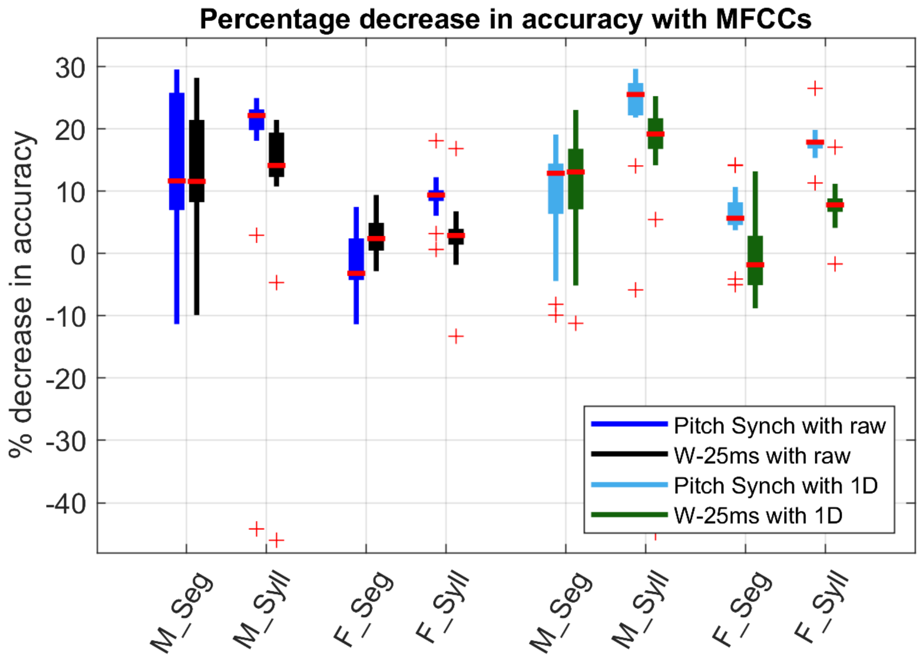


Figure 3-7 Percentage reduction in classification performance using MFCCs with original and random labels from Database 1

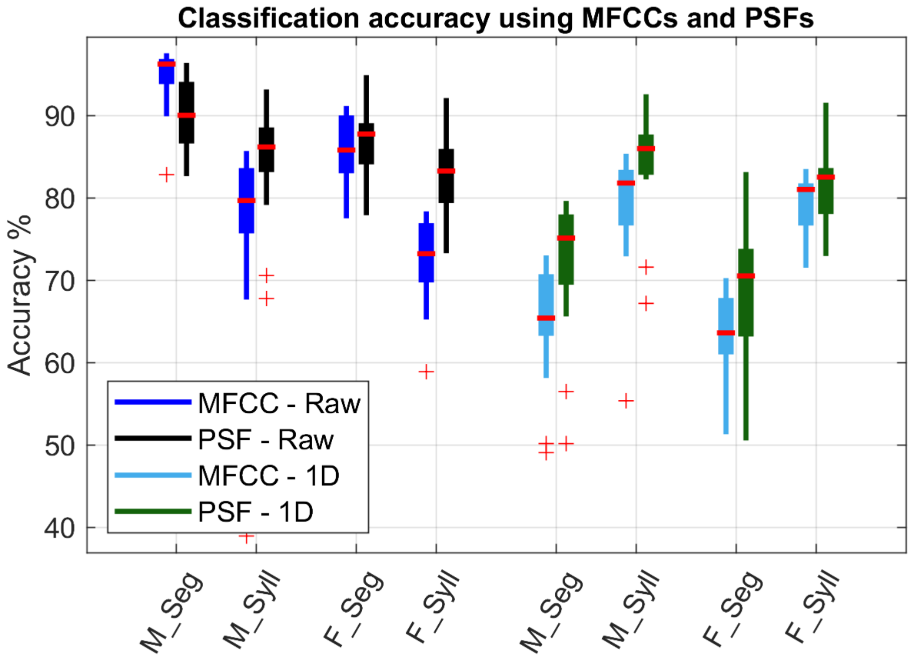


Figure 3-8 Classification performance comparison between MFCCs and PSFs extracted from Database 1

Overall, for MFCCs, in all gender and feature variant combinations, PS segmentation with syllabic analysis had a better reliable performance with 1st order differences in MFCCs.

The MFCCs and PSFs are evaluated using their original/raw values and their 1st order differences. Comparison between feature type (MFCC or PSF) and feature variants (raw or 1st order differences) was made using classification accuracies with original labels and the relative decline in performance due to random label assignment. The performance comparison between MFCCs and PSFs using Database 1 for both genders, both feature variants and both analysis methods is shown in Figure 3-8. When raw features were used, except for males under segmental analysis, PSFs had better median accuracies than MFCCs in all other combinations.

The performance difference between raw MFCCs and raw PSFs was more significant for syllabic analysis than segmental analysis. It shows that the ability to contain speaker information is better for raw MFCCs than raw PSFs (segmental analysis), as shown in Figure 3-8. It can also be noticed that for 1st order differences, performance differences were greater for segmental analysis than syllabic analysis. It shows a higher ability of 1st order differences in PSFs to capture the impact of PD on vocalic dynamics than 1st order differences in MFCCs (syllabic analysis).

Percentage reduction in classification performance with original labels and randomly assigned labels was also higher for PSFs than MFCCs in all cases, as shown in Figure 3-9 (black and green bars), suggesting the effectiveness of PSFs over MFCCs in containing the effects of PD. When results from segmental and syllabic analyses were compared in Figure 3-9, syllabic analysis had a higher performance decrease than segmental analysis for both feature types. Additionally, the gap between segmental and syllabic analyses' percentage decrease was higher for MFCCs than PSFs for both genders in both feature variants. It was due to the ability of MFCCs to identify the speakers and the power of PSFs to capture the effects of PD better. As shown in Figure 3-7, Figure

3-9 also shows negative performance differences due to the overfitting observed in few classifiers using fine kernels. Overall, PSFs with syllabic analysis were observed to be optimal where the performances were comparable to segmental analysis in both feature variants (raw and 1st order differences) for both genders. Between feature variants, raw features worked better under the segmental analysis method and 1st order differences were better suited for syllabic analysis. However, from segmentation method and analysis method combinations, syllabic analysis using features extracted from PS segments was better suited for PD classification. Hence, from the syllabic analysis results, the optimal combination of feature type and variant for PD classification can be identified as 1st order differences in PSFs.

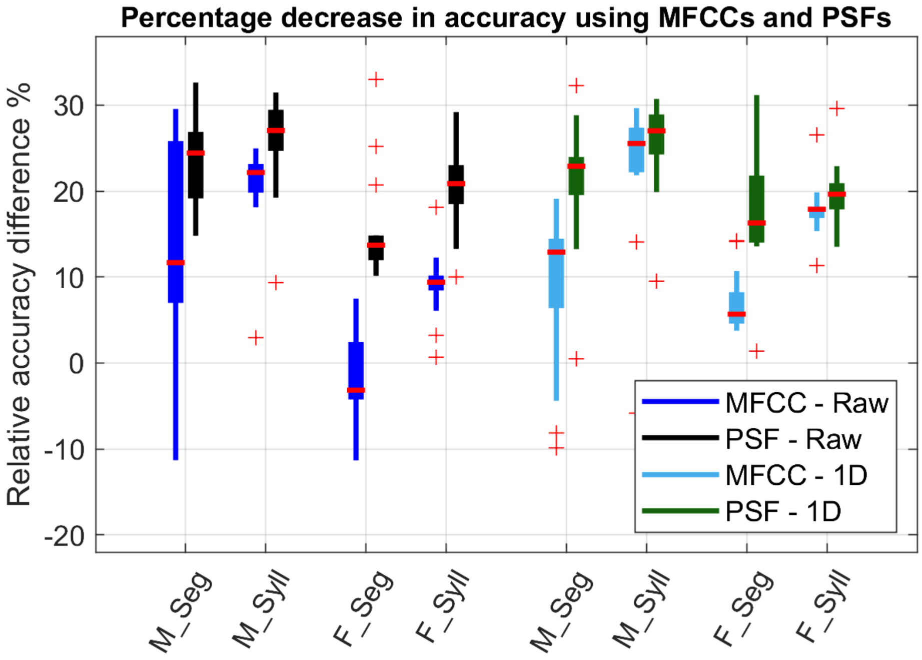


Figure 3-9 Percentage reduction in classification performance with original and random labels - Comparison between MFCCs and PSFs (Database 1)

Each of the 17 classifiers was evaluated using two metrics: Overfit factor using syllabic analysis shown in Table 3-1 and test accuracy with segmental analysis using random label assignment shown in Table 3-2. The classifiers in the ensemble were arranged in increasing order of overfit factor observed in males using raw MFCC with syllabic analysis in both Table 3-1 and Table 3-2. Index numbers were given to each classifier in the left-most columns of both tables used to identify them later in discussions.

Table 3-1 Overfit factor using PSFs with original labels under syllabic analysis

No.	**Classifier**	**Male**		**Female**	
		Raw	**1st order Diff**	**Raw**	**1st order Diff**
1	MediumKNN	0.05	0.04	0.01	0.04
2	CoarseKNN	0.04	0.02	-0.00	-0.01
3	CosineKNN	0.04	0.04	0.01	0.04
4	**LinearSVM**	0.03	0.03	0.00	0.01
5	**CoarseTree**	0.01	0.02	0.00	0.03
6	**CoarseGaussianSVM**	0.07	0.06	-0.02	-0.00
7	**MediumTree**	0.04	0.06	0.06	0.06
8	**EnsembleBoostedTrees**	0.06	0.07	0.06	0.07
9	**RUSBoostedTrees**	0.04	0.06	0.06	0.05
10	**LogisticRegression**	0.00	0.01	0.02	0.03
11	FineTree	0.10	0.12	0.13	0.14
12	MediumGaussianSVM	0.10	0.12	0.03	0.03
13	FineKNN	0.15	0.15	0.19	0.20
14	WeightedKNN	0.14	0.13	0.16	0.17
15	EnsembleBaggedTrees	0.10	0.10	0.11	0.11
16	EnsembleSubspaceKNN	0.29	0.29	0.27	0.27
17	FineGaussianSVM	0.31	0.32	0.16	0.10

All the classifier experiments (training, testing and obtaining results) were repeated ten times, and median values of overfit factors and test accuracies are provided in Table 3-1 and Table

3-2, respectively. From Table 3-1, for MFCCs and PSFs, in both genders, overfit factors for classifiers 1 to 10 are significantly lower than the rest in most cases. They were highlighted using a thicker outside border in Table 3-1. Starting from classifier 11 till classifier 17, overfit factors increase, implying increased overfitting to training data.

Table 3-2 Median test accuracy using MFCCs with random labels under segmental analysis

No.	**Classifier**	**Male**		**Female**	
		Raw	**1st order Diff**	**Raw**	**1st order Diff**
1	MediumKNN	90.18	60.63	93.73	64.34
2	CoarseKNN	85.33	61.79	89.11	64.13
3	CosineKNN	89.46	59.77	93.14	62.51
4	**LinearSVM**	67.69	54.19	80.44	47.96
5	**CoarseTree**	66.52	56.23	77.01	60.7
6	**CoarseGaussianSVM**	75.12	58.22	83.92	47.05
7	**MediumTree**	68.83	56.39	76.81	59.83
8	**EnsembleBoostedTrees**	74.04	56.73	82.76	57.2
9	**RUSBoostedTrees**	68.83	56.39	76.81	59.83
10	**LogisticRegression**	67.62	54.01	79.92	54.14
11	FineTree	72.18	57.1	80.62	58.81
12	MediumGaussianSVM	86.63	62.27	90.78	65.39
13	FineKNN	91.27	59.4	93.91	59.64
14	WeightedKNN	91.97	61.24	94.29	62.37
15	EnsembleBaggedTrees	88.67	60.09	92.39	62.81
16	EnsembleSubspaceKNN	90.62	59.5	94.08	62.16
17	FineGaussianSVM	92.14	60.6	95.86	66.02

Table 3-2 contains classifier-wise median test accuracies under segmental analysis when labels were randomized. The classifiers with lower accuracy for 1st order differences and higher accuracies for raw features were highlighted using a thicker outside border. Though the first three classifiers showed low overfit values in Table 3-1 (classifiers 1 to 10), they had significantly higher accuracy values even with random labels. It shows that k-NNs are better equipped for speaker verification applications than PD classification. They learn minimal information about the effects of PD and rely more on the closeness of samples in feature space. Following these observations, only the classifiers 4 to 10 were recognized as optimal classifiers. Their names were also bolded in the second column in both tables. The values were color-coded in both tables, with green and red colors denoting desirable and undesirable values.

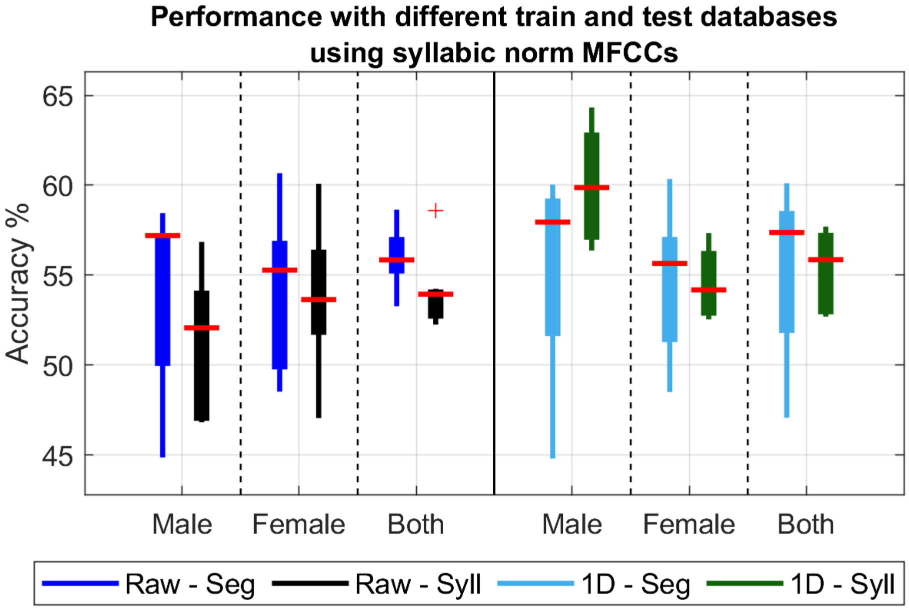


Figure 3-10 Classification accuracies with different training and testing datasets using MFCCs

For the final step, training and testing were carried out using different databases. Between speakers in Database 1 and Database 2, nationality, spoken language, research group involved in data collection, length of paragraphs and sampling frequency (16 kHz for Database1 and 44.1 kHz for Database 2). Figure 3-10 and Figure 3-11 show the results with syllabic normalizations applied before training the classifiers.

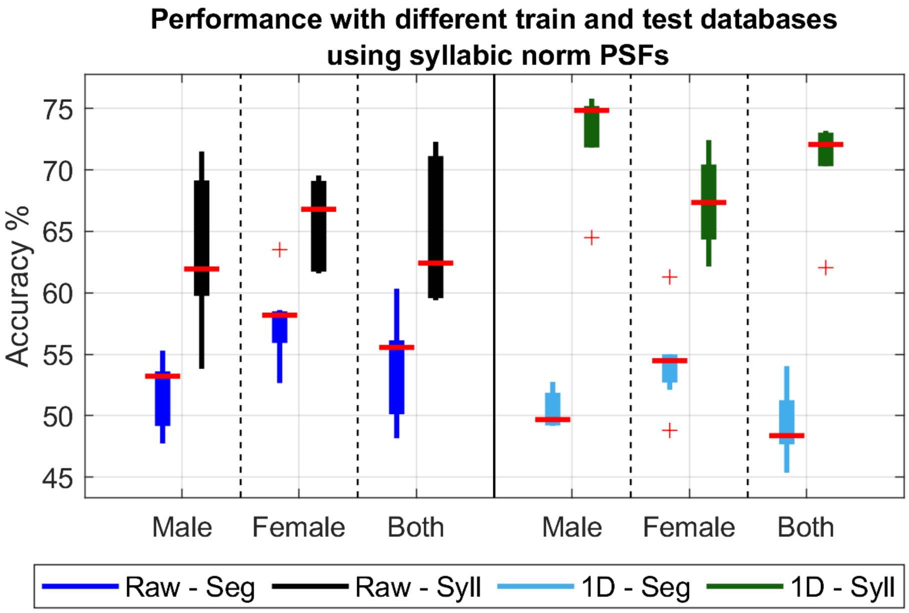


Figure 3-11 Classification accuracies with different training and testing datasets using PSFs

For both variants of MFCCs (raw and 1st order differences) with gender independent grouping (black), coarse Gaussian SVM and logistic regression resulted in relatively higher performance across all metrics, as shown in Figure 3-12. When PSFs were considered, Medium Trees also have comparable performance to coarse Gaussian SVM and logistic regression, as shown in Figure 3-13. Hence, coarse Gaussian SVM and logistic regression stand out as better classifiers for PD classification using MFCCs or PSFs under syllabic analysis protocol. When different databases were used for training and testing with PS segmentation, PSFs and syllabic analysis, Coarse Gaussian SVM and logistic regression have classification accuracy close to 75% without any indication of bias in F1-scores MCC. A closer look into the results in Figure 3-8 shows that these two classifiers produced accuracies of 85% under the same conditions when only Database 1 was used for training and testing.

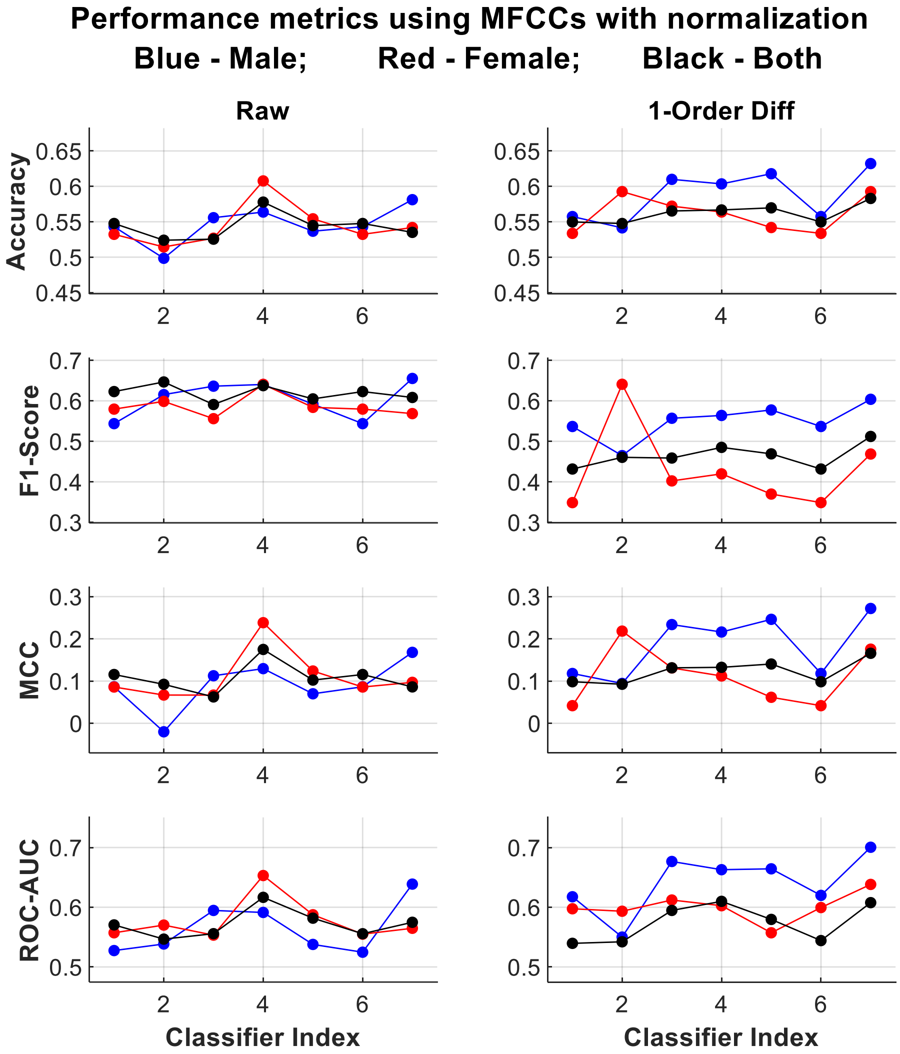


Figure 3-12 Performance metrics with different training and testing datasets using MFCCs in three gender groups

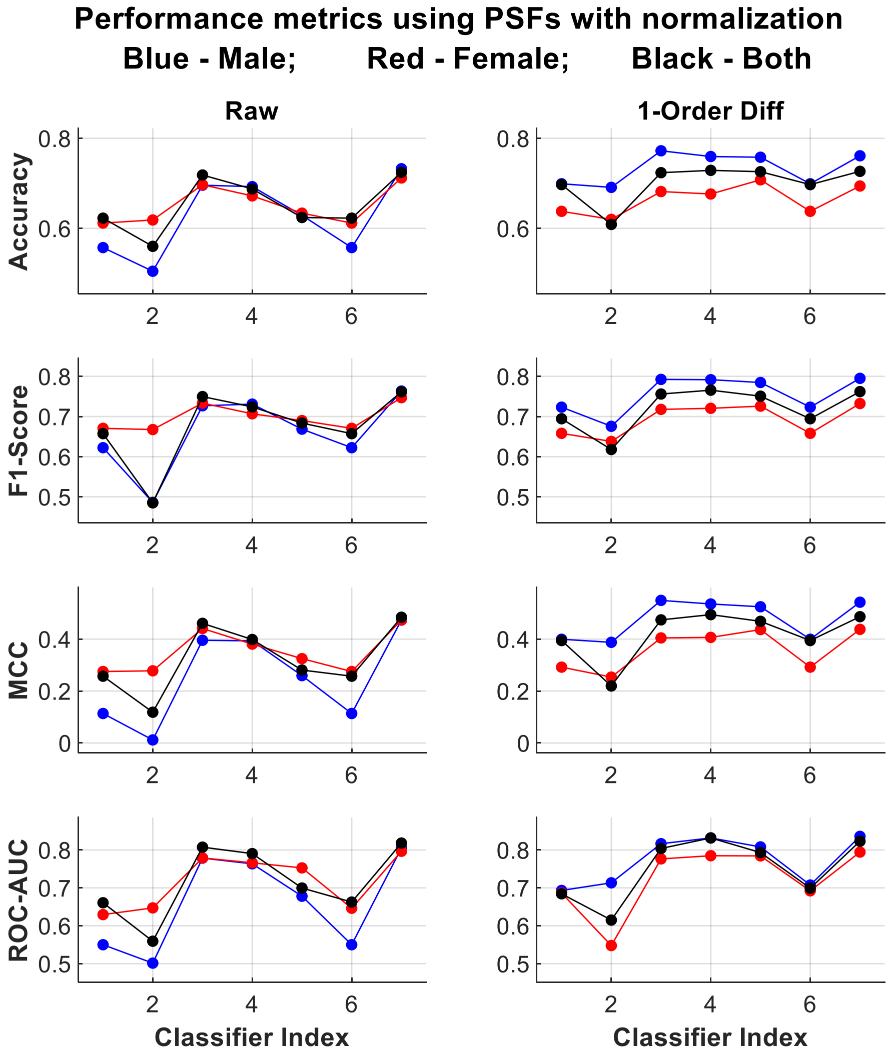


Figure 3-13 Performance metrics with different training and testing datasets using PSFs in three gender groups

The results presented so far were computed when each segment (in segmental analysis) or super segment (in syllabic analysis) was treated as an individual sample. The percentage of correctly predicted super segments from each participant using logistic regression with 1st order differences in PSFs is shown in Figure 3-14. Results from coarse Gaussian SVM followed the results from logistic regression very closely. More than 80% of the super segments from all PD speakers (red) and more than 50% of super segments from 6 out of 10 HC speakers were correctly predicted with their respective labels. Hence, the proposed framework had a high recall and good precision for PD detection.

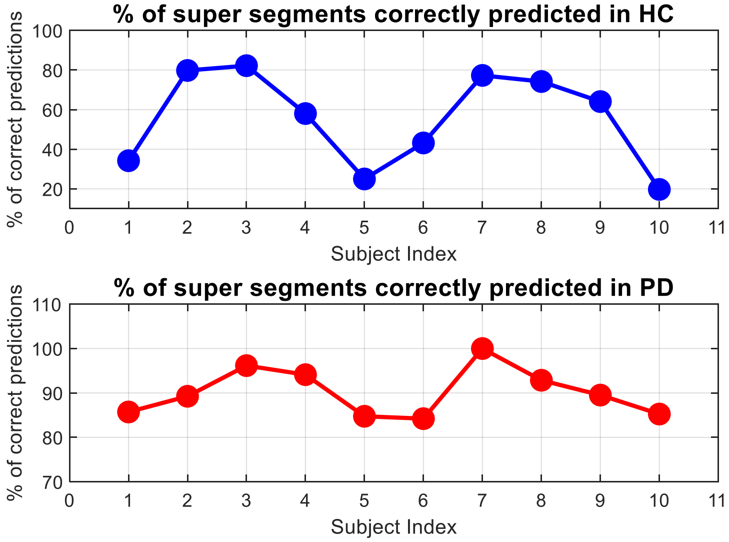


Figure 3-14 Percentage of super segments correctly predicted in each subject

3.5.2 Results From Instituted Method

When the sustained phonations were used for PD classification, the classifiers trained included the seven identified as optimal from the results presented earlier. These classifiers were trained using the features from a random pool containing 80% of the samples from Database 1. Trained classifiers were evaluated over the remaining 20% of the samples from Database 1 and all the samples from Database 2. Results from the evaluation of leftover samples of Database 1 are labeled ‘Train’, and results from evaluations over Database 2 are labeled ‘Test’ in Figure 3-15 and Figure 3-16.

When PSFs are used for classification, they are processed using the proposed framework, and the covariance values were used for training and testing. VOICEBOX-based features and PSFs have performed well while evaluated over the leftover samples from Database 1. With completely unseen data from Database 2, results worsened and had no value while using VOICEBOX features. With PSFs, they had a median accuracy of 70%.

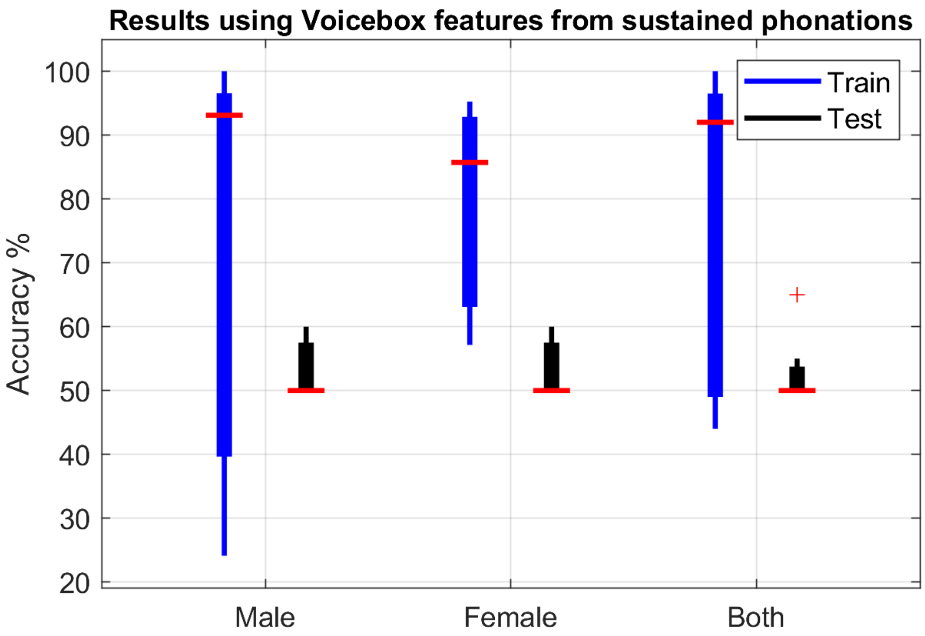


Figure 3-15 Results using the VOICEBOX features from sustained phonations

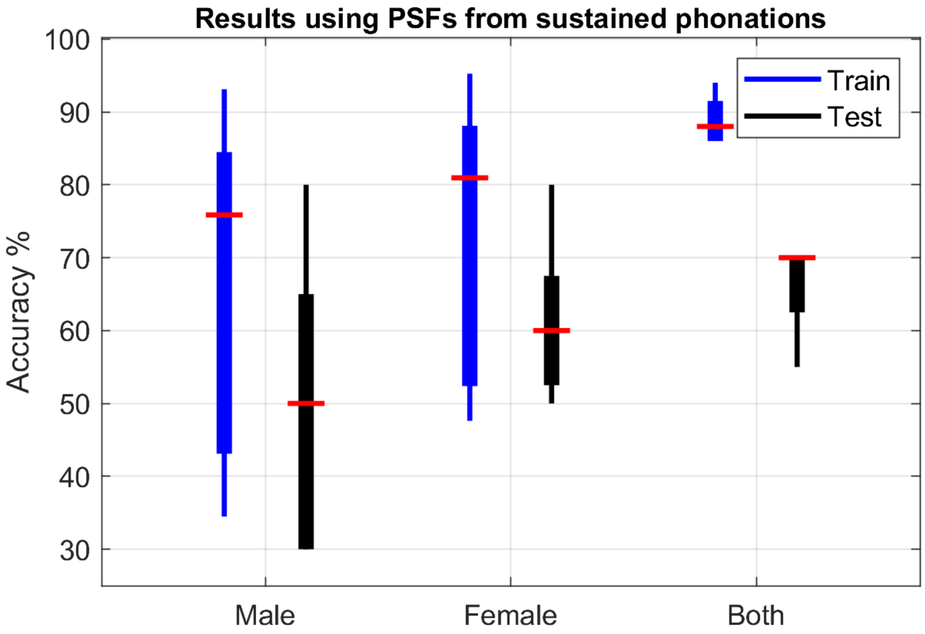


Figure 3-16 Results using the PSFs from sustained phonations

3.5.3 Comparisons

From the results for the proposed framework, it becomes clear that covariances of the normalized PSFs from voiced segments of the connected speech provide reliable performance. On the other hand, VOICEBOX features used in many studies did not deliver any performance when the trained models are tested on data from a different dataset.

When the instituted method with sustained phonations is compared against the proposed framework with connected speech, the proposed framework had better and reliable results. Instituted methods performed very well when evaluated over the leftover data from the dataset used for training (Database 1). When the proposed framework was used with sustained phonations, the test results were no different from the test results from instituted method. However, the train set evaluations show that PSFs perform less than the VOICEBOX features. This performance difference shows that VOICEBOX features are superior for identifying the speakers and have less impact on PD classification.

Comparing the proposed framework results with connected speech and sustained phonations shows better performance with connected speech. The vocalic dynamics from connected speech contain the effects of PD better than sustained phonations. While sustained phonations are praised for the simplicity in acquisition and processing, the lack of intonational variations makes them less suitable for PD classification.

Chapter 4: Analysis Using Unsupervised Feature Extraction with Pitch Synchronous Segmentation

4.1 Background

4.1.1 Long Short-Term Memory (LSTM)

Artificial neural networks have been under research study for classification for over several decades. Modern neural network architectures are developed specifically for various applications. While convolutional neural networks are extensively used for image-based applications, they cannot handle sequential data like speech. Neural network architectures like recurrent neural networks and long short-term memory were developed to maintain a memory component that can help analyze the features of the signal that vary with time.

RNNs are similar to multi-layer perceptron networks with minor variations. In addition to the inputs from preceding layers, each neuron in RNNs gets feedback from the previous time point. This feedback mechanism helps the RNNs take the information learned from the previous time steps into consideration while adjusting the weights. RNNs can deliver incredible performance when the inputs are shorter in length without much variation. When the input sequences become longer, it becomes more challenging for them to consider all the information learned from the initial time steps. This issue is called the vanishing gradient problem. The longer input sequences can also result in the opposite effect causing exploding gradient problem. Due to these issues, RNNs are modified and replaced with LSTM models.

LSTM is a modified version of RNN that handles the vanishing and exploding gradients problem using internal memory units to contain information over long periods. These memory units have a linear relation to the memory units from the preceding time-steps. LSTM models use three gates: input, forget and output, to gauge how much information from current, previous and output model representations are appropriate for modeling the current timestep. Equations (20) to (23) can be used to obtain output for input, forget, cell memory and output gates, respectively.

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{R}_i \boldsymbol{y}_{t-1} + \boldsymbol{b}_i) \tag{20}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{R}_f \boldsymbol{y}_{t-1} + \boldsymbol{b}_f\right) \tag{21}$$

$$\boldsymbol{c}_t = \boldsymbol{c}_{t-1} * \boldsymbol{f}_t + \boldsymbol{i}_t * \tanh\,(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{R}_c \boldsymbol{y}_{t-1} + \boldsymbol{b}_c) \tag{22}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{R}_o \boldsymbol{y}_{t-1} + \boldsymbol{b}_o) \tag{23}$$

where **W** represents the input weights, **R** represents the recurrent weights, **b** is the bias component, **i** is the input, and **y** is the output. The * operator represents the element-wise multiplication operation [26].

4.1.2 Autoencoders

Autoencoders are powerful tools that have been used for applications like dimensionality reduction and anomaly detection. They map a high dimensional input (x) to a low dimensional latent space (z) in the bottleneck. These latent values in z are then used to reconstruct the original input. The first half of the Autoencoder, where x is mapped to z is called the encoder (g : x → z) and the second part, where input is reconstructed, is the decoder (f : z → x). The encoder and decoder are constructed such that they are symmetrical about the bottleneck, as shown in Figure 4-1. The neural networks in the encoder and decoder halves can be created using any architecture like fully connected networks, CNNs, RNNs or LSTMs. The training is done based on a loss computed from the difference between the input and its reconstructed counterpart. The z is used as a set of features representing each input for PD classification.

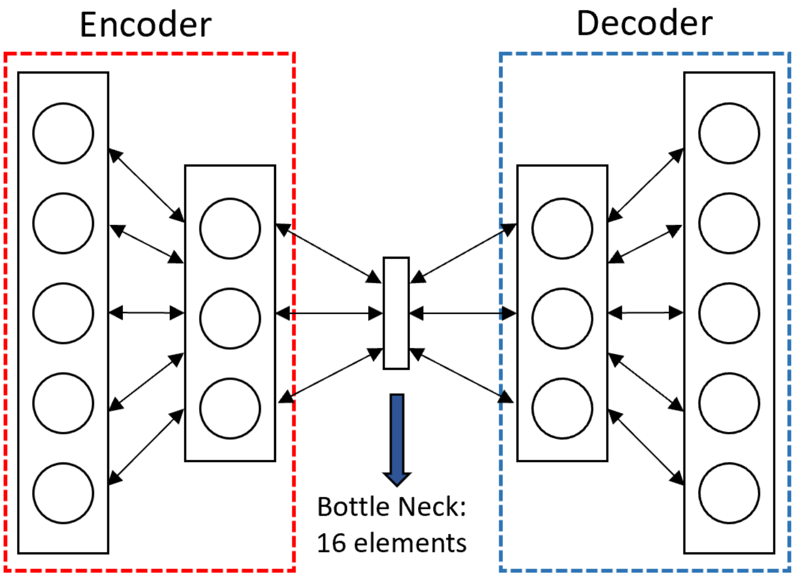


Figure 4-1 Autoencoder model

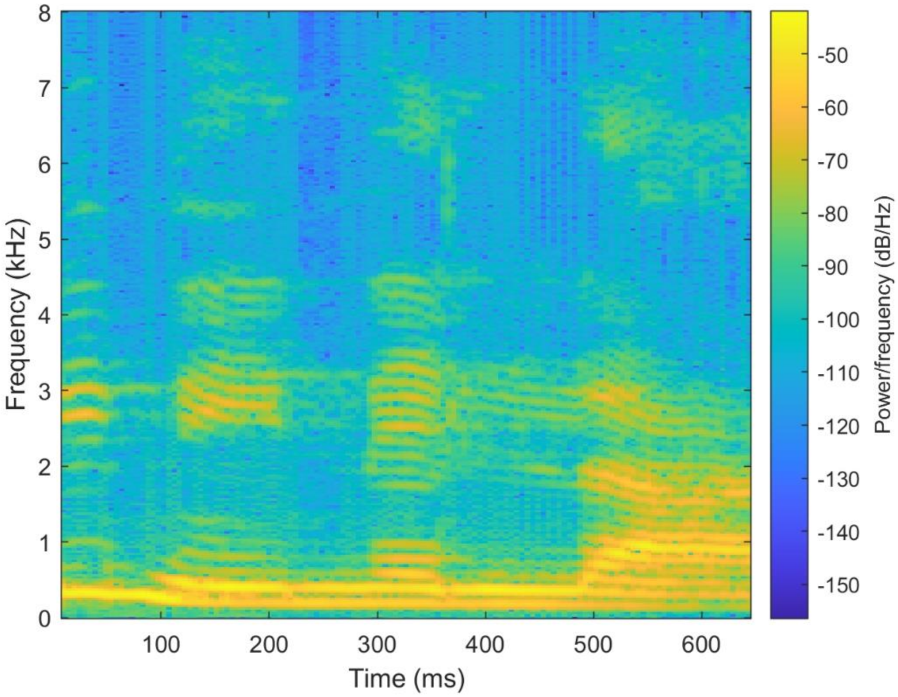


Figure 4-2 Spectrogram of voiced segment in a speech sample

The Autoencoders are used to take the spectrograms of varying lengths and reconstruct them precisely. Studies using spectrograms (images) typically employ CNN Autoencoders for targeted dimensionality reduction in feature space [171, 186-188]. Spectrograms are the time-frequency representation of speech with time on one axis and frequency on the other, as shown in Figure 4-2. Due to the higher variability in the data segment lengths, the spectrograms have variability in the time axis. However, CNNs typically accept inputs of fixed dimensions, and due to this limitation, spectrograms must be resized using image resizing methods, or the audio segments have to be clipped. Either modification can negatively impact the data, and when used with Autoencoders, the bottleneck elements will represent modified data rather than raw data.

4.1.3 LSTM-Autoencoder

LSTM-Autoencoders are developed for dimensionality reduction in this study. The covariance-based dimensionality reduction using PSFs adopted in the previous chapter is replaced with LSTM-Autoencoder to evaluate the classification performance using the unsupervised feature extraction method. With the availability of cycle-to-cycle variations, LSTM-Autoencoder can automatically learn the features that can quantify the variations.

Studies on LSTM architectures have suggested using simple (vanilla) LSTM to perform reasonably well on various datasets for different applications [189]. The hyper-parameters used for LSTM architecture are all widely used in the research. The network weights are updated using adaptive moment (ADAM) estimation method. The training criterion was focused on reconstruction efficiency only.

4.1.4 Transfer Learning

Transfer learning is a popular technique used with deep neural network training. Transfer learning helps in training efficient models even with relatively smaller datasets by harnessing the knowledge of a deep neural network trained over a much larger dataset. Transfer learning methods are typically employed when the task to be accomplished with limited data availability is close to the task addressed by the pre-trained model over a larger dataset.

4.2 Autoencoder Training

The Autoencoder is trained using speech samples from the TIMIT database initially, and then it was retrained using the speech samples from Database 1. This transfer learning based method is used to avoid any potential discrepancies due to the limited number of speakers and super segments available from Database 1.

4.2.1 TIMIT Dataset and Pre-processing

TIMIT is a widely used dataset for speech recognition containing recordings of short sentences read by 630 speakers [190]. Each speaker read ten sentences which after preprocessing yielded over 56000 super segments for training the Autoencoder. All the recordings were made in a sound-proof environment at 44.1 kHz. Each sentence recording was phonetically transcribed, and the phonemic boundaries are also made available. The 6300 samples were pitch synchronously segmented and super segments were identified using the automatic segmentation algorithm explained in previous chapter.

During pre-processing, each super segment identified in the speech recordings from TIMIT, Database 1 and Database 2 are turned into spectrograms containing frequencies between 0 Hz and 22.1 kHz, which is the Nyquist frequency with 256 bins. The spectrograms were created to have a fixed dimension of 100 rows by 256 columns representing 100 pitch cycles and 256 frequency bins. 98% of the super segments from all three databases had less than 100 pitch cycles in them. Spectrograms were created using the Fourier transform of each pitch cycle repeated four times and stacked vertically. When the number of pitch cycles in a super segment is less than 100, the missing rows of the spectrogram contains zeros. These spectrograms were normalized using min-max normalization with a maximum value of 1 and minimum of zero. Super segments with more than 100 pitch cycles are clipped to fit into the input dimensions. The preprocessing methodology adopted for each super segment is shown in Figure 4-3.

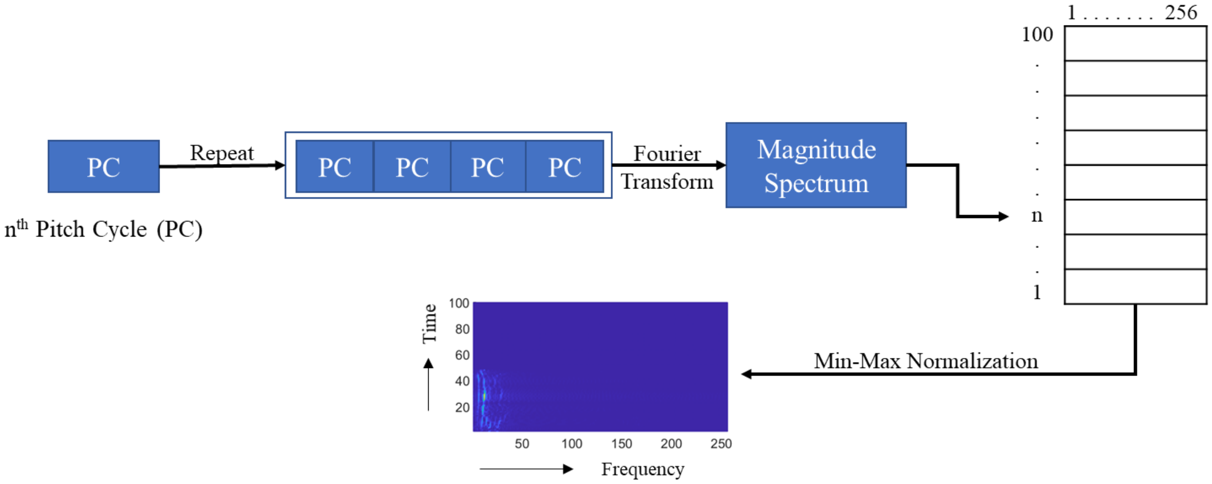
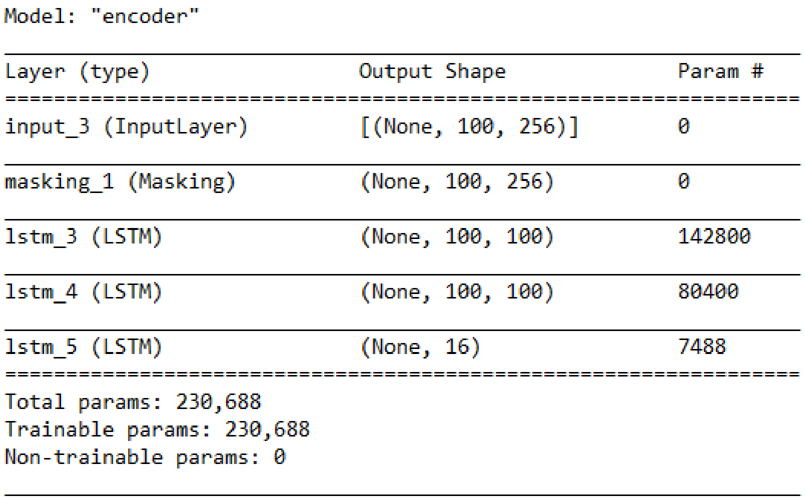


Figure 4-3 Super segment pre-processing for Autoencoder training

4.2.2 Autoencoder Architecture

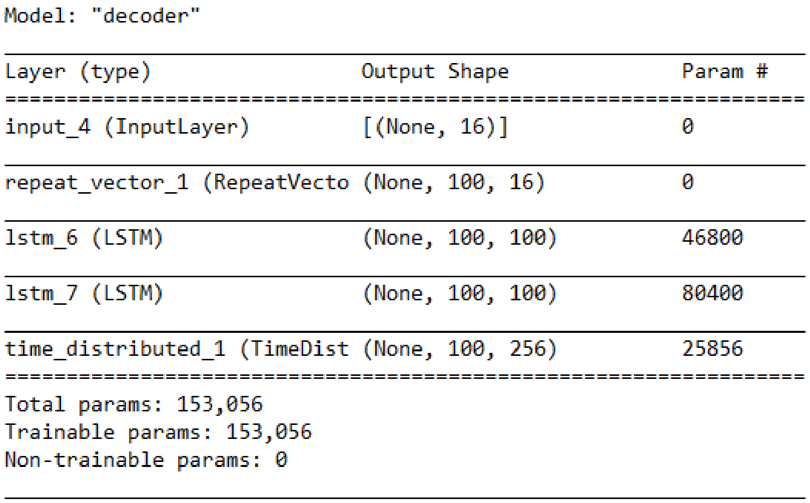
The LSTM-Autoencoder architecture used in this study is simple, containing five layers on the encoder and decoder as shown in Figure 4-4 and Figure 4-5. The inputs to the Autoencoder are structured to have a dimension of 100 by 256. It means that the super segments containing up to 100 pitch cycles can be used at a time. Experimentation for the optimal Autoencoder architecture included different variants with 1 to 2 hidden layers and 50 or 100 LSTM units in those layers. The architectures shown in Figure 4-4 and Figure 4-5 are employed as the optimal version due to the performance improvement over the less deep variants.

Deeper versions of the Autoencoder required much longer to converge with incremental improvements over the optimal architecture presented here. Masking technique is used in this LSTM architecture to account for the variable lengths in the spectrograms. The layer ‘masking_1’ in Figure 4-4 provides the ‘lstm_3’ layer with binary True/False values specifying the rows with Fourier transforms and empty rows. Portions of the inputs containing zeros for the entire rows are omitted from computations during training. Hence, the models are trained only on the relevant data without modifying the raw data. For activation, rectified linear unit (relu) is used for every unit in all layers of encoder and decoder.



```
Model: "encoder"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_3 (InputLayer)         [(None, 100, 256)]        0
_________________________________________________________________
masking_1 (Masking)          (None, 100, 256)          0
_________________________________________________________________
lstm_3 (LSTM)                (None, 100, 100)          142800
_________________________________________________________________
lstm_4 (LSTM)                (None, 100, 100)          80400
_________________________________________________________________
lstm_5 (LSTM)                (None, 16)                7488
=================================================================
Total params: 230,688
Trainable params: 230,688
Non-trainable params: 0
_________________________________________________________________
```

Figure 4-4 Encoder architecture



```
Model: "decoder"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_4 (InputLayer)         [(None, 16)]              0
_________________________________________________________________
repeat_vector_1 (RepeatVecto (None, 100, 16)           0
_________________________________________________________________
lstm_6 (LSTM)                (None, 100, 100)          46800
_________________________________________________________________
lstm_7 (LSTM)                (None, 100, 100)          80400
_________________________________________________________________
time_distributed_1 (TimeDist (None, 100, 256)          25856
=================================================================
Total params: 153,056
Trainable params: 153,056
Non-trainable params: 0
_________________________________________________________________
```

Figure 4-5 Decoder architecture

4.2.3 Train Protocol

The Autoencoder was developed using TensorFlow 2.0 [191] on the Python platform. With over 56000 samples from the TIMIT dataset, training, cross-validation and testing sets were created with 80%, 10% and 10% split. The model was trained for 25 epochs with a batch size of 16 samples. With a training time of 790 seconds per epoch, the model was trained for 5.5 hours. The mean square error between the original input and reconstructed inputs was used for monitoring the training and determining the model convergence.

Once the model has been trained on TIMIT data, it was retrained over the samples from Database 1 for ten epochs. Database 1 contained 12601 samples, and similar to the training protocol using TIMIT, these samples were also split into training, cross-validation and testing sets with the same ratio. The average training time per epoch while retraining was 197 seconds with a batch size of 16 samples. The model was created such that it has the portability to extract the encoder half and use it for transforming the spectrograms from Database 1 and Database 2 into 16-dimensional vectors.

Database 2 resulted in 1357 samples, which were not part of Autoencoder training. They were only transformed using the autoencoders after the second phase training using Database 1. Similar to the classification protocols adopted in Chapter 3, classifiers were trained on 80% of the 12601 samples from Database 1 for training. The rest are used for testing along with all the samples from Database 2.

4.3 Results and Discussion

The results are presented here in two parts, first the results from the autoencoder training and the efficiency of transfer learning are presented. Then the actual PD classification results are presented. Autoencoder training is carried out

4.3.1 Training Results

Even though the Autoencoders were trained for 25 epochs over 80% of the TIMIT data, convergence could be seen after 9th epoch as shown in Figure 4-6. When the models had fewer hidden layers or fewer units in the hidden layers, the convergence could only be seen after 15th epoch. For the optimal model containing 2 hidden layers and 100 hidden units in each layer, the mean squared error (MSE) after the first phase of learning over the TIMIT data was identified to be 2.68×10^{-4}. After retraining the model in the second phase over Database 1, MSE was identified to be 2.43×10^{-4} for Database 1 and 2.51×10^{-4} for Database 2.

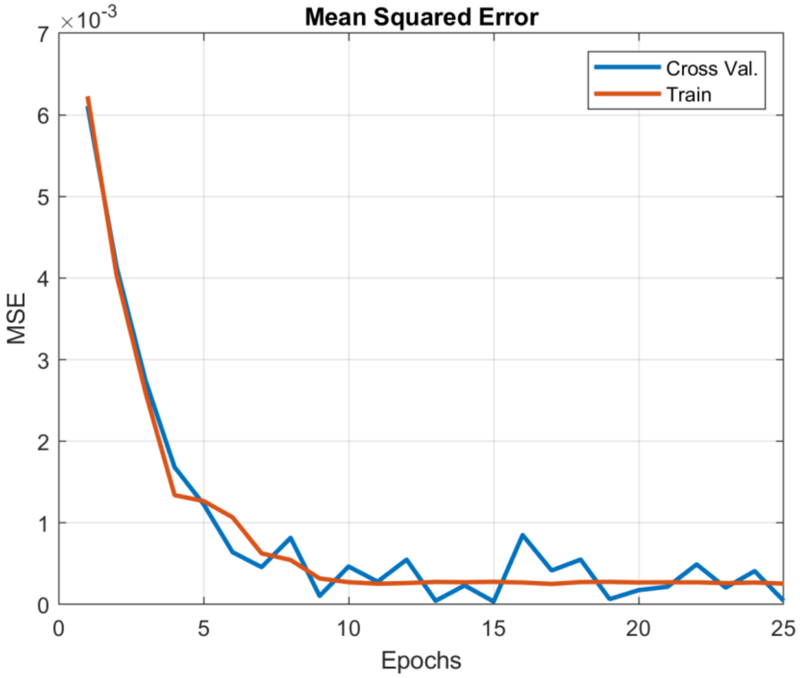


Figure 4-6 Autoencoder MSE during training

4.3.2 Transfer Learning Efficiency

The transfer learning approach followed in training the Autoencoder has shown great performance in terms of having a MSE from second stage comparable to first stage where the input dataset was much larger. The reconstructions of a random sample from each of the three datasets

(TIMIT, Database 1 and Database 2) are shown in Figure 4-7. It can be observed from the figure that reconstruction preserves the harmonic structure as well as the energy in each of the harmonics. The impressive reconstruction reflects on the efficiency of latent features at the bottleneck.

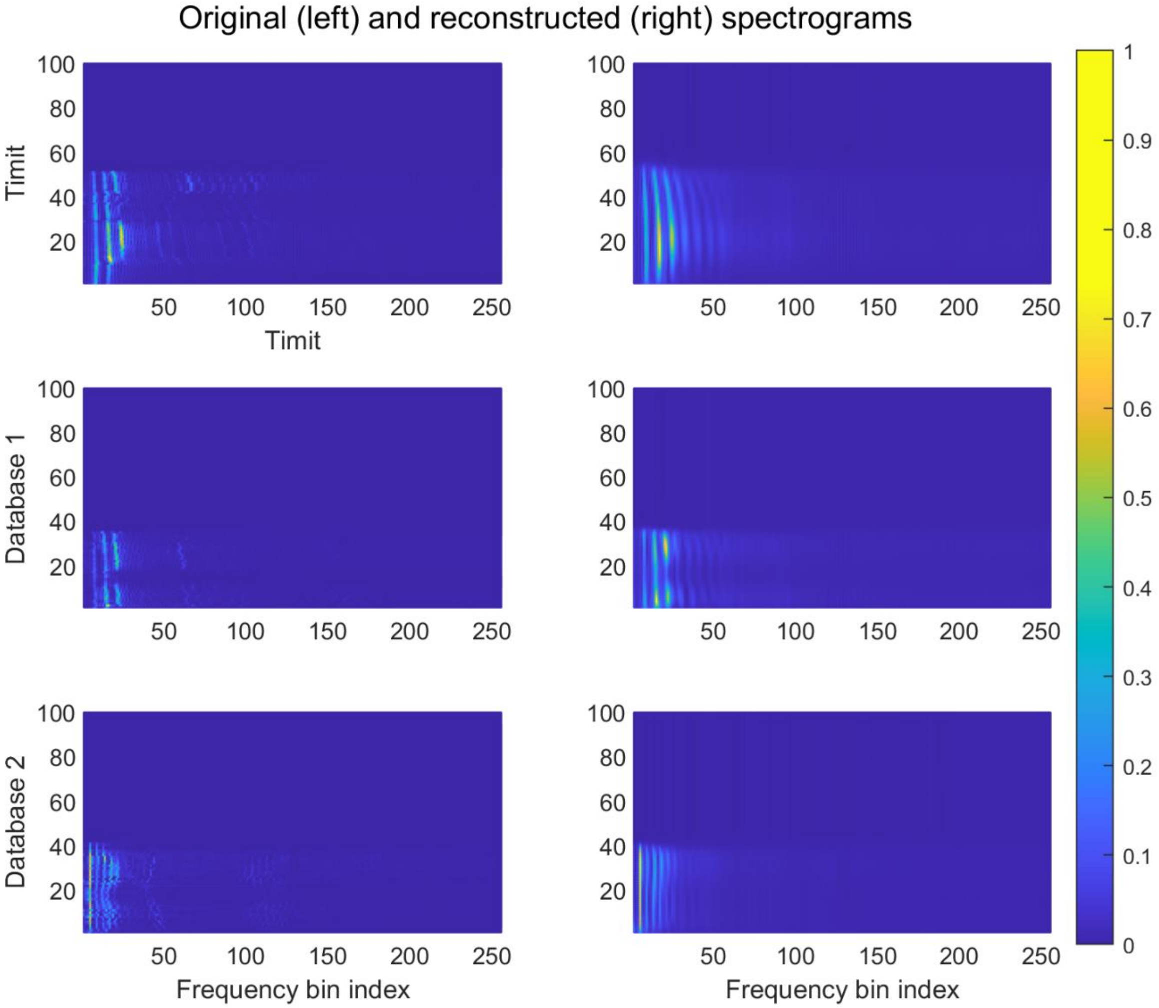


Figure 4-7 Original and reconstructed spectrogram samples

4.3.3 Classification Results

The seven classifiers identified optimal for PD classification in Chapter 3 are used for analyzing the LSTM Autoencoder-based features. These classifiers are listed in Table 4-1. Classifiers are trained on 80% of samples from Database 1 and tested on the remaining 20% samples and all the samples from Database 2.

Table 4-1 List of classifiers evaluated

No.	Classifiers
1	LinearSVM
2	CoarseTree
3	CoarseGaussianSVM
4	MediumTree
5	EnsembleBoostedTrees
6	RUSBoostedTrees
7	LogisticRegression

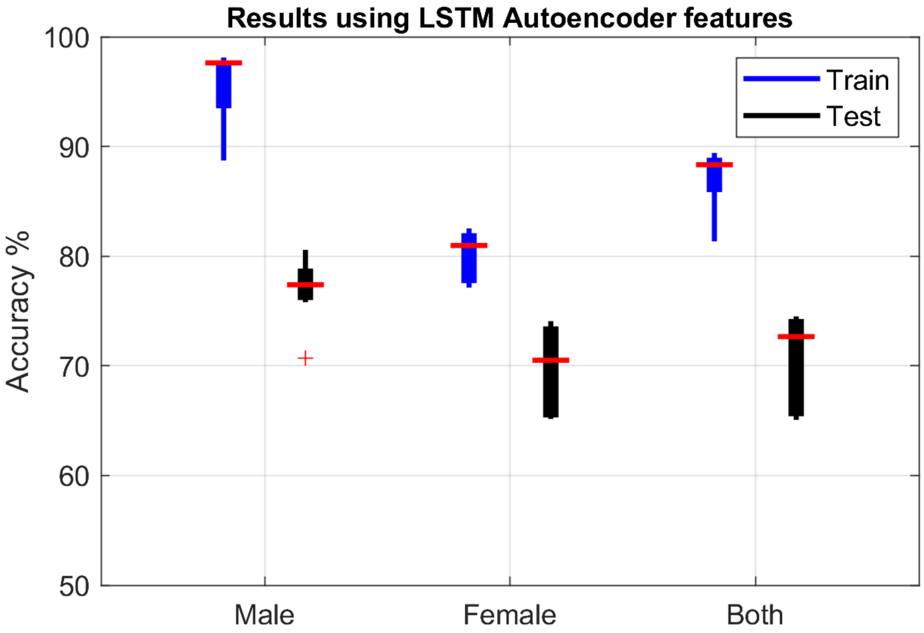


Figure 4-8 Classification accuracies using latent space features from LSTM Autoencoders

Classification accuracies across the classifiers for both databases are shown in Figure 4-8. These results show that unsupervised LSTM Autoencoder-based features have a performance comparable to the performance from PSFs presented in the previous chapter. Across the different gender groups, the male population has better classification results than female and gender independent groups. The difference between the train (20% of Database 1) and test (Database 2) accuracies is much lesser compared to the results from sustained phonations and VOICEBOX based features. Though the features are learned by the Autoencoder automatically, the preprocessing had a positive impact on the classification performance. The percentage of the super segments correctly predicted for each subject using Autoencoder features is shown in Figure 4-9. With a 50% threshold, seven out of ten HC subjects and eight out of ten PD subjects will be correctly assigned to the respective class.

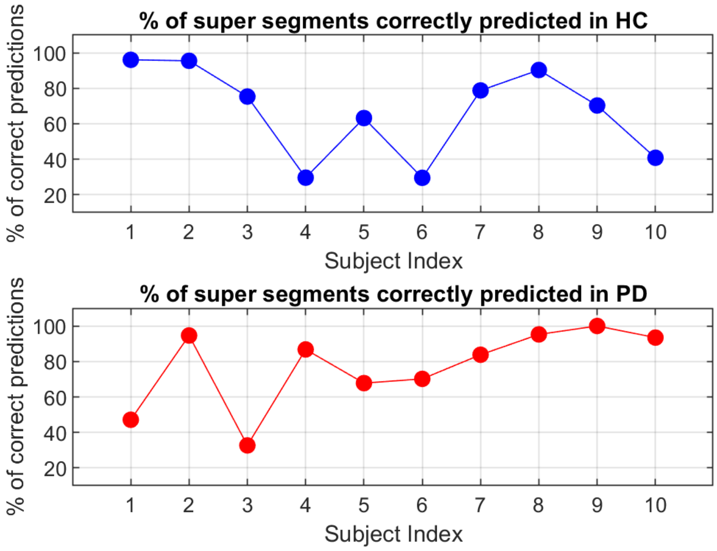


Figure 4-9 Percentage of super segments correctly predicted in each subject using Autoencoder features

In addition to the classification accuracy, other metrics like F1-Score, MCC and ROC-AUC are shown in Figure 4-10. Coarse Gaussian SVM, medium tree and logistic regression have shown the superior performance among the seven classifiers based on the results for gender independent (black) group. Between the two genders, the better performance with males can be attributed to the better intonational variability maintained by female population compared to males

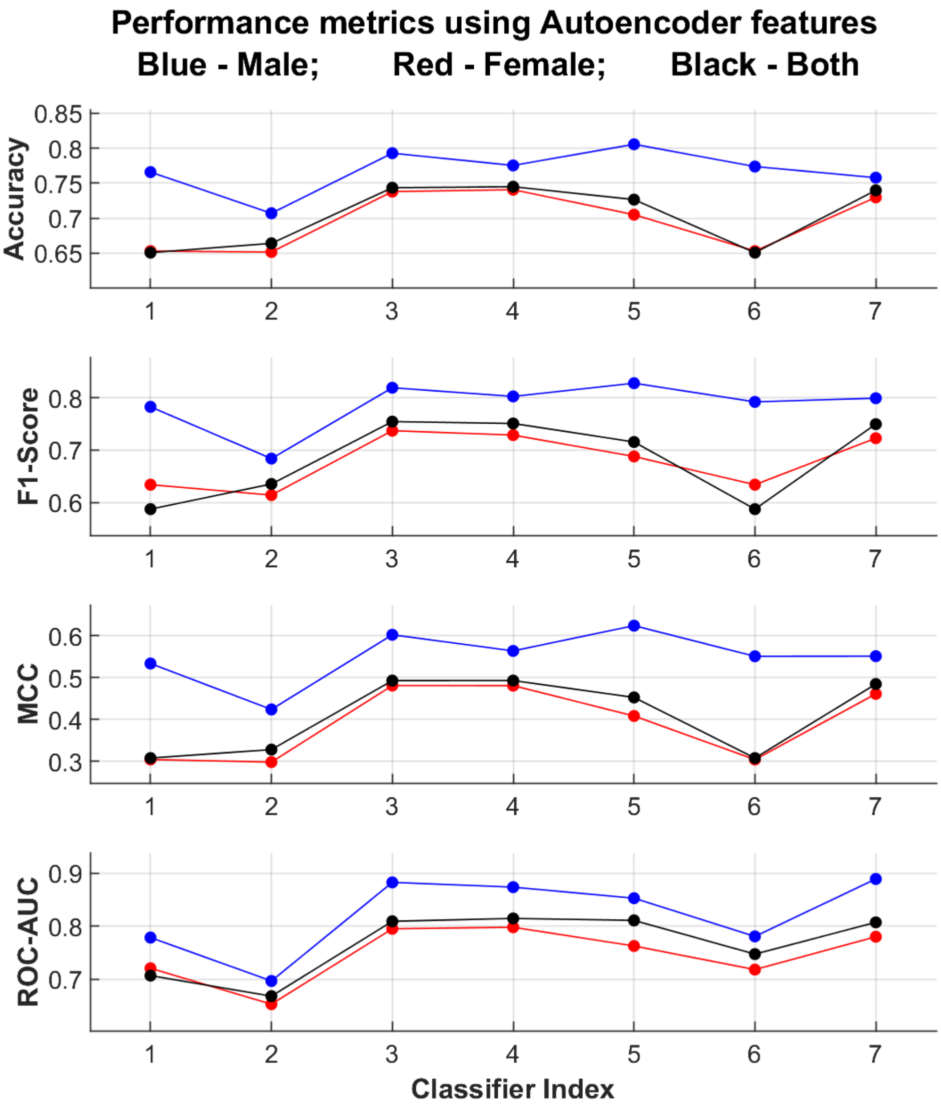


Figure 4-10 Classification performance metrics using features from LSTM autoencoder

in both datasets. The three classifiers with better performance also have relatively less delineation between male (blue) and female (red) groups, showing their higher reliability than others.

4.3.4 Comparisons

Eight different classification performance metrics are compared between classification using pitch synchronous features and latent space features from LSTM Autoencoders as shown in Figure 4-11. The comparison shows the closeness in performance results between using both feature types. Except for sensitivity/recall and specificity, all the other metrics do not vary much. For all the three classifiers, PSFs have higher recall than Autoencoder-based features. Higher recall makes PSFs a better choice for PD detection, as false positives can be processed and eliminated with additional tests. With higher specificity, Autoencoder-based features perform better in identifying the negatives. Comparing the results in Figure 3-14 and Figure 4-9, it is observed that more than 80% of the super segments from every speaker with PD have been classified correctly using PSFs, and 4 out of 20 speakers in the HC group have a chance for misclassification. On the other hand, with Autoencoder-based features, for the PD group, the percentage of super segments

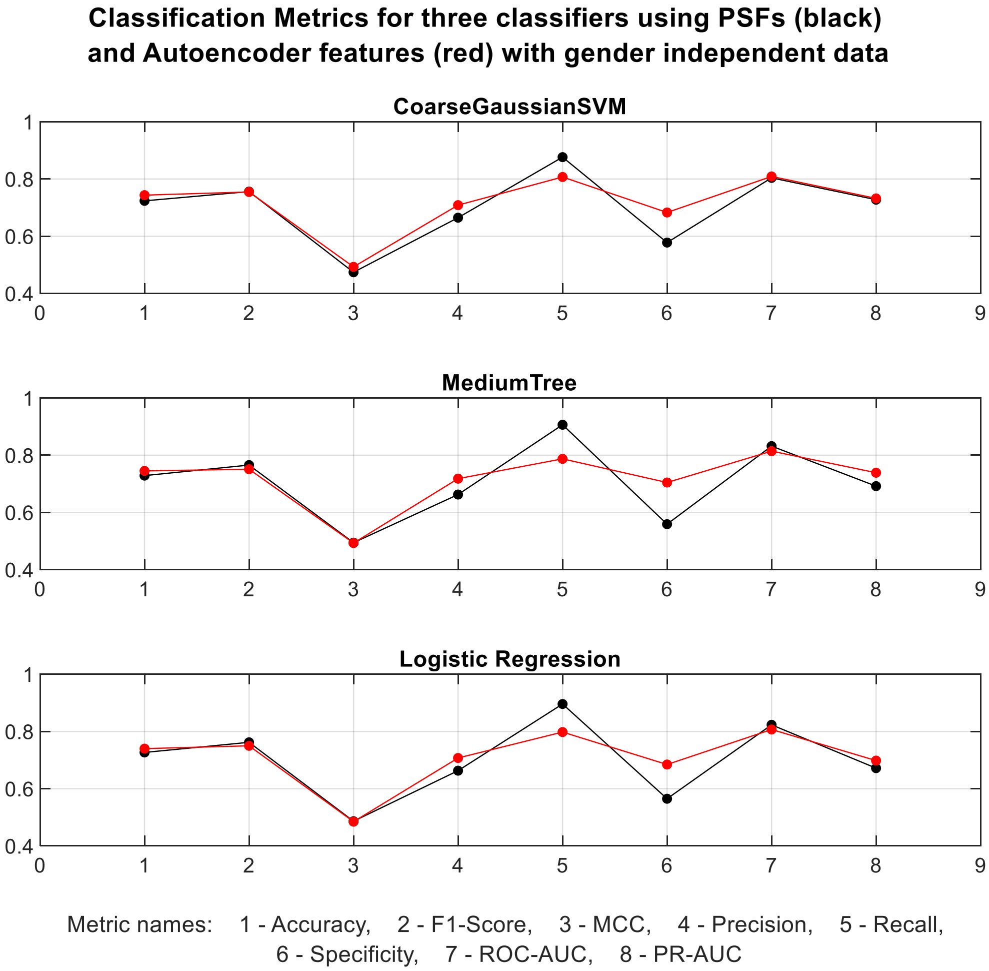


Figure 4-11 Classification metrics with gender independent data using PSFs and Autoencoder features

that are correctly classified went down for many subjects, and two subjects have this percentage below 50% which can lead to misdiagnosis. The percentages are higher for HC subjects than PSFs, but still have three subjects with the misdiagnosis possibility. The Autoencoder-based features were transformed using principal component analysis (PCA) to reduce the dimensionality while retaining 95% variance. The performance variation before and after using PCA was miniscule across all the metrics.

Overall, classification performance is very close for both Autoencoder-based features and PSFs across all the metrics as shown in Figure 4-11. However, PSFs provide more reliable performance in terms of classifying PD subjects correctly. The PS segmentation for the analysis of super segments for PD classification has proven to impact classification even with data from a different research study proving the generalization capacity of this method.

Chapter 5: Conclusion

5.1 Conclusion

Analysis of connected speech has been part of evaluations for neurodegenerative disease diagnosis for a long time. While the effect on speech production differs between the diseases, it is well known that speech evaluations play a vital role in decision making. Current evaluation methods largely depend on the perceptual rating systems, which suffer from subjectivity issues. While research studies utilizing perceptual analysis focus on improving the reliability by employing multiple listeners and conducting statistical tests to identify inter-rater reliabilities, the inherent subjectivity issues are hard to mitigate.

Advances in signal processing and AI significantly impact the efficiency of various tasks employing speech and other time-varying signals. The cutting-edge algorithms developed for different applications using speech are applied to neurodegenerative speech analysis to improve objective decision-making. While the algorithms prove to work in some studies using traditional methods, their reliability is often suspected in the research community. The capacity to generalize over different datasets is paramount for these applications and testing the established and traditional methods seldom provides evidence of their generalization ability. Hence, research on the administration methodology of the advanced methods for speech evaluations for neurodegenerative diseases is of utmost importance.

For neurodegenerative diseases like Parkinson's disease (PD), the impact on speech production in the phonatory and articulatory domains is well known in the literature. Existing acoustic analysis methods predominantly focus on using sustained phonation of /a/ vowel to extract various acoustic features that can be used to train automatic classification systems. These studies try to take advantage of the simplicity in processing the sustained phonations from a signal processing standpoint. It is observed from the literature that though sustained phonations dominate this research area, it cannot contain the simple vocalic dynamics that are present in the connected speech (conversational speech, reading or monologues). Due to this drawback of sustained phonations, some research studies have focused on connected speech for automatic analysis.

Due to its highly dynamic nature, analysis of connected speech becomes tricky and often requires manual intervention compared to sustained phonations. Hence, in this dissertation, a novel automatic analysis framework with various variations to the traditional methods has been developed. The framework differs from traditional methods in many ways and targets the vocalic dynamics for identifying PD. The proposed framework uses pitch synchronous segmentation instead of block processing to avoid the spectral distortion problem and maintain consistency in analysis. An automatic segmentation algorithm developed for this framework has proven to be efficient in extracting the pitch cycles from voiced portions of the connected speech.

The framework has been tested systematically to analyze the impact of every change introduced to the established methods. The novel pitch synchronous feature set was developed to capture the cycle-to-cycle variations within the voiced portions of the connected speech. A covariance-based feature transformation method has been developed for capturing the perturbations quantified using 1st order differences in the features. Compared to the established methods, the novel feature set had superior classification performance for sustained phonations and connected speech. The efficacy of the proposed framework's two basic components, PS segmentation and representation of the voiced segments using feature vectors, has been evaluated using an Autoencoder-based unsupervised feature extraction system. In this system, the latent variables at the bottleneck of LSTM Autoencoders trained to reconstruct the spectrogram representation of the voiced segments were used to train the classifiers.

Similar to most studies in literature, evaluation of sustained phonations using established methods provided 92% cross-validation accuracy where the data used for testing is a subset of Database 1 withheld before training the models. When those models are tested using data from a different dataset, Database 2, accuracy falls to 50%. This shows that existing methods fail at creating ubiquitous models. When covariances of PSFs from sustained phonations are used for classification, the cross-validation accuracy was 89% and test accuracy on data from Database 2 was 70%. However, the gender specific performance shows that the test accuracy using male data was still at 50%. This shows the potential of PSF in capturing the effects of PD, but the speech task used for analysis hinders the performance.

The systematic evaluation of proposed framework shows the efficacy of PS segmentation, PSFs and feature covariances over block processing, MFCCs and segmental analysis. Comparison between raw and 1st order differences shows the superiority of 1st order differences which proves that phonatory perturbations hold vital information for PD classification. With the proposed framework using PSF covariances, the cross-validation accuracy was 86%, and test accuracy using a different dataset was 72%. The gender specific results show 73% accuracy for male group and 68% accuracy for female group. These results portray the strength of proposed framework in generating ubiquitous models. With LSTM Autoenoders, the cross-validation and test accuracies were 89% and 73%. Gender specific results show 78% accuracy for males and 76% accuracy for females. This proves that PS segmentation and features representing the voiced segments of connected speech contain vital information helpful for PD detection.

The results provide strong evidence to the robustness of the proposed framework. Classification studies also showed the dependency of some classifiers over the closeness of the samples in the feature space. The overfit factor for each classifier shows how this dependency can deliver superior performance with the data from speakers with whom the models are already familiar. Overall, this research shows the importance of using the connected speech over sustained phonations and the efficiency of a novel automatic evaluation framework for connected speech in neurodegenerative diseases with the data from PD and healthy controls.

5.2 Future Research

There are various directions for this research that can help in advancing the automatic analysis and evaluations.

- Development of advanced data acquisition protocols to target various voiced sounds. In the current study, all voiced segments from connected speech have been used for analysis. Further research can target the specific sounds to identify the patterns that can help improve the speech task.
- Speech duration and the impact of vocal fatigue is another direction where the current studies have not found any significant impact. The speech task can be made longer to include the same phrases at the beginning and end to identify the impact of vocal fatigue and how that can vary between pathological and healthy voices.
- The manifestations of neuromotor effects on speech production differ from person to person. Future studies can focus on developing generative adversarial networks (GAN) to develop subject-specific models for longitudinal studies. The GANs can help synthesize reliable data that can be used to conduct comparative studies for telemonitoring applications.

- Real-time testing of the proposed framework becomes challenging due to the computational requirements for pitch synchronous segmentation. Hence cloud-based solutions can be developed to make use of wearable devices and IoT applications for real-time implementations.
- Proposed framework can be developed further to create an evaluation system for the drug effects and impact of procedures like DBS. It can also be used to evaluate the impact of different speech therapy methods. The objective evaluation methods developed through this framework can be used to evaluate and compare the diagnostic performances of neurologists or clinicians who conduct clinical tests as specified by the rating scales.

References

[1] H. Von Helmholtz, *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Vieweg, 1863.

[2] G. S. Ohm, "Über die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen," *Annalen der Physik,* vol. 135, no. 8, pp. 513-565, 1843.

[3] G. Proakis John and G. Manolakis Dimitris, "Digital signal processing: principles, algorithms, and applications," *Pentice Hall,* 1996.

[4] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara,* vol. 1, p. 67, 2005.

[5] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences,* vol. 9, no. 19, p. 4050, 2019.

[6] T. Bäckström, *Speech coding: with code-excited linear prediction*. Springer, 2017.

[7] A. I. Al-Shoshan, "Speech and music classification and separation: a review," *Journal of King Saud University-Engineering Sciences,* vol. 19, no. 1, pp. 95-132, 2006.

[8] S. Greenberg and W. A. Ainsworth, "Speech processing in the auditory system: an overview," *Speech processing in the auditory system,* pp. 1-62, 2004.

[9] ClevelandClinic, "Dysarthria," 11/16/2020 Art. no. 17653.

[10] P. Gómez Vilda, "Biomedical applications of voice and speech processing," *Loquens,* vol. 4, no. 1, p. e035, 06/30 2017.

[11] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: a review," *Frontiers in psychology,* vol. 8, p. 269, 2017.

[12] N. Miller, U. Nath, E. Noble, and D. Burn, "Utility and accuracy of perceptual voice and speech distinctions in the diagnosis of Parkinson’s disease, PSP and MSA-P," *Neurodegenerative disease management,* vol. 7, no. 3, pp. 191-203, 2017.

[13] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain and cognition,* vol. 56, no. 1, pp. 24-29, 2004.

[14] J. Lam and K. Tjaden, "Clear speech variants: An acoustic study in Parkinson's disease," *Journal of Speech, Language, and Hearing Research,* vol. 59, no. 4, pp. 631-646, 2016.

[15] V. Martel Sauvageau, J.-P. Roy, M. Langlois, and J. Macoir, "Impact of the LSVT on vowel articulation and coarticulation in Parkinson’s disease," *Clinical Linguistics & Phonetics,* vol. 29, no. 6, pp. 424-440, 2015.

[16] K. L. Lansford and J. M. Liss, "Vowel acoustics in dysarthria: Speech disorder diagnosis and classification," 2014.

[17] M. Perez *et al.*, "Classification of huntington disease using acoustic and lexical features," in *Interspeech*, 2018, vol. 2018, p. 1898: NIH Public Access.

[18] E. A. Strand, J. R. Duffy, H. M. Clark, and K. Josephs, "The Apraxia of Speech Rating Scale: A tool for diagnosis and description of apraxia of speech," *Journal of communication disorders,* vol. 51, pp. 43-50, 2014.

[19] A. Illa *et al.*, "Comparison of speech tasks for automatic classification of patients with amyotrophic lateral sclerosis and healthy subjects," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6014-6018: IEEE.

[20] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," *arXiv preprint arXiv:2004.06833,* 2020.

[21] J. Wang, P. V. Kothalkar, B. Cao, and D. Heitzman, "Towards Automatic Detection of Amyotrophic Lateral Sclerosis from Speech Acoustic and Articulatory Samples," in *Interspeech,* 2016, pp. 1195-1199.

[22] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease," *Nature Precedings,* pp. 1-1, 2008.

[23] Y. B. Zikria, M. K. Afzal, and S. W. Kim, "Internet of Multimedia Things (IoMT): Opportunities, Challenges and Solutions," ed: Multidisciplinary Digital Publishing Institute, 2020.

[24] K. A. Al Mamun, M. Alhussein, K. Sailunaz, and M. S. Islam, "Cloud based framework for Parkinson’s disease diagnosis and monitoring system for remote healthcare applications," *Future Generation Computer Systems,* vol. 66, pp. 36-47, 2017.

[25] S.-S. Shin, G. Y. Kim, B. M. Koo, and H.-G. Kim, "Parkinson's disease diagnosis using speech signal and deep residual gated recurrent neural network," *The Journal of the Acoustical Society of Korea,* vol. 38, no. 3, pp. 308-313, 2019.

[26] J. Mallela *et al.*, "Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson’s Disease and healthy controls with CNN-LSTM using transfer learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2020, pp. 6784-6788: IEEE.

[27] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease," in *INTERSPEECH,* 2017, pp. 314-318.

[28] Y. Jia *et al.*, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037,* 2019.

[29] G. Devitt, K. Howard, A. Mudher, and S. Mahajan, "Raman spectroscopy: an emerging tool in neurodegenerative disease research and diagnosis," *ACS chemical neuroscience,* vol. 9, no. 3, pp. 404-420, 2018.

[30] M. Schenkman, T. M. Cutson, C. W. Zhu, and K. Whetten-Goldstein, "A longitudinal evaluation of patients' perceptions of Parkinson's disease," *The Gerontologist,* vol. 42, no. 6, pp. 790-798, 2002.

[31] K. López-de-Ipiña *et al.*, "On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis," *Sensors,* vol. 13, no. 5, pp. 6730-6745, 2013.

[32] K. López-de-Ipina *et al.*, "Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach," *Computer Speech & Language,* vol. 30, no. 1, pp. 43-60, 2015.

[33] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of the royal society interface,* vol. 8, no. 59, pp. 842-855, 2011.

[34] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE transactions on biomedical engineering,* vol. 59, no. 5, pp. 1264-1271, 2012.

[35] J. R. Duffy, *Motor Speech disorders-E-Book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[36] G. L. Wallace, "Assessment of oral peripheral structure and function in normal aging individuals with the Frenchay," *Journal of communication disorders,* vol. 24, no. 2, pp. 101-109, 1991.

[37] W. Maetzler, I. Liepelt, and D. Berg, "Progression of Parkinson's disease in the clinical phase: potential markers," *The Lancet Neurology,* vol. 8, no. 12, pp. 1158-1171, 2009.

[38] S. Appakaya, S. A. Khoshnevis, E. Sheybani, and R. Sankar, "A novel pitch cycle detection algorithm for tele monitoring applications," in *2020 Wireless Telecommunications Symposium (WTS)*, 2020, pp. 1-4: IEEE.

[39] S. Appakaya and R. Sankar, "Effectiveness of Speech Analysis in Classification of Neurodegenerative Diseases: A Study on Parkinson's Disease," in *SoutheastCon 2018*, 2018, pp. 1-5: IEEE.

[40] S. Appakaya and R. Sankar, "Classification of Parkinson's disease Using Pitch Synchronous Speech Analysis," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 1420-1423: IEEE.

[41] S. Appakaya and R. Sankar, "Parkinson's Disease Classification using Pitch Synchronous Speech Segments and Fine Gaussian Kernels based SVM," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 236-239: IEEE.

[42] S. Appakaya, R. Sankar, and I.-H. Ra, "Classifier Comparison for Two Distinct Applications Using Same Data," in *9th International Conference on Smart Media and Applications (SMA 2020)*, 2020, pp. 1-4: ACM.

[43] O.-B. Tysnes and A. Storstein, "Epidemiology of Parkinson's disease," *Journal of Neural Transmission,* vol. 124, no. 8, pp. 901-905, 2017.

[44] S. Fahn, "Description of Parkinson's disease as a clinical syndrome," *ANNALS-NEW YORK ACADEMY OF SCIENCES,* vol. 991, pp. 1-14, 2003.

[45] C. Marras *et al.*, "Prevalence of Parkinson’s disease across North America," *NPJ Parkinson's disease,* vol. 4, no. 1, p. 21, 2018.

[46] M. Baldereschi *et al.*, "Parkinson’s disease and parkinsonism in a longitudinal study: two-fold higher incidence in men," *Neurology,* vol. 55, no. 9, pp. 1358-1363, 2000.

[47] C. A. Haaxma *et al.*, "Gender differences in Parkinson’s disease," *Journal of Neurology, Neurosurgery & Psychiatry,* vol. 78, no. 8, pp. 819-824, 2007.

[48] S. K. Van Den Eeden *et al.*, "Incidence of Parkinson’s disease: variation by age, gender, and race/ethnicity," *American journal of epidemiology,* vol. 157, no. 11, pp. 1015-1022, 2003.

[49] W. Muangpaisan, H. Hori, and C. Brayne, "Systematic Review of the Prevalence and Incidence of Parkinson's Disease in Asia," (in English), *Journal of Epidemiology,* vol. 19, no. 6, pp. 281-293, 2009.

[50] S. Mukherjee, H. Wu, and J. Jones, "Healthcare data analytics for Parkinson’s disease patients: a study of hospital cost and utilization in the United States," in *AMIA annual symposium proceedings*, 2016, vol. 2016, p. 1950: American Medical Informatics Association.

[51] J. J. Chen, "Parkinson's disease: health-related quality of life, economic cost, and implications of early treatment," *Am J Manag Care,* vol. 16 Suppl Implications, pp. S87-93, Mar 2010.

[52] D. Purves, A. Fitzpatrick, L. Katz, A. La Mantia, and J. McNamara, "Neuroscience. Sunderland Mass: Sinauer Assoc," *Inc. Publ,* pp. 121-44, 1997.

[53] J. Jankovic, "Parkinson’s disease: Clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry,* vol. 79, pp. 368-376, 2012.

[54] J. Massano and K. P. Bhatia, "Clinical Approach to Parkinson's Disease: Features, Diagnosis, and Principles of Management," *Cold Spring Harbor Perspectives in Medicine,* vol. 2, no. 6, p. a008870, 2012.

[55] F. Verde *et al.*, "An old woman with pressure ulcer, rigidity, and opisthotonus: never forget tetanus!," *Lancet,* vol. 384, no. 9961, p. 2266, Dec 20 2014.

[56] D. Annett Blochberger, MRPharmS, and Shelley Jones, DipClinPharm, MRPharmS, "Parkinson's disease: clinical features and diagnosis," *Clinical Pharmacist,* vol. 3, pp. 361-366, 2011.

[57] S. Bostantjopoulou, Z. Katsarou, C. Karakasis, E. Peitsidou, D. Milioni, and N. Rossopoulos, "Evaluation of non-motor symptoms in Parkinson's Disease: An underestimated necessity," *Hippokratia,* vol. 17, no. 3, pp. 214-9, Jul 2013.

[58] L. P. Leow, M.-L. Huckabee, T. Anderson, and L. Beckert, "The impact of dysphagia on quality of life in ageing and Parkinson's disease as measured by the swallowing quality of life (SWAL-QOL) questionnaire," *Dysphagia,* vol. 25, no. 3, pp. 216-220, 2010.

[59] P. Martinez-Martin, C. Rodriguez-Blazquez, M. M. Kurtis, K. R. Chaudhuri, and N. V. Group, "The impact of non-motor symptoms on health-related quality of life of patients with Parkinson's disease," *Movement Disorders,* vol. 26, no. 3, pp. 399-406, 2011.

[60] E. K. Plowman-Prine *et al.*, "The relationship between quality of life and swallowing in Parkinson's disease," *Movement disorders: official journal of the Movement Disorder Society,* vol. 24, no. 9, pp. 1352-1358, 2009.

[61] T. R. Barber, J. C. Klein, C. E. Mackay, and M. T. Hu, "Neuroimaging in pre-motor Parkinson's disease," *NeuroImage: Clinical,* vol. 15, pp. 215-227, 2017.

[62] P. W. C. K. Dr. Jolyon Meara, *Parkinson's disease and parkinsonism in the elderly.* (Cambridge University Press). 2000.

[63] C. G. Goetz *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results," *Movement disorders: official journal of the Movement Disorder Society,* vol. 23, no. 15, pp. 2129-2170, 2008.

[64] S. G. Reich, "Diagnosing Parkinson's Disease," in *Therapy of Movement Disorders*: Springer, 2019, pp. 3-6.

[65] P. Zemankova, O. Lungu, and M. Bares, "Psychosocial Modulators of Motor Learning in Parkinson's Disease," (in eng), *Frontiers in human neuroscience,* vol. 10, pp. 74-74, 2016.

[66] I. H. K. Leung, C. C. Walton, H. Hallock, S. J. G. Lewis, M. Valenzuela, and A. Lampit, "Cognitive training in Parkinson disease," *Neurology,* vol. 85, no. 21, p. 1843, 2015.

[67] S. H. Fox *et al.*, "International Parkinson and movement disorder society evidence-based medicine review: Update on treatments for the motor symptoms of Parkinson's disease," *Movement Disorders,* vol. 33, no. 8, pp. 1248-1266, 2018.

[68] S. Szlufik, M. Szumilas, J. Dutkiewicz, D. Koziorowski, T. Mandat, and E. Slubowska, "The impact of STN DBS on kinetic tremor in parkinson's disease patients," *Parkinsonism & Related Disorders*, vol. 22, pp. e109-e110, 2016.

[69] V. Voon *et al.*, "Impulse control disorders and levodopa-induced dyskinesias in Parkinson's disease: an update," *The Lancet Neurology*, vol. 16, no. 3, pp. 238-250, 2017.

[70] A. Wagle Shukla and M. Okun, "State of the art for deep brain stimulation therapy in movement disorders: A clinical and technological perspective," *IEEE Rev Biomed Eng*, Jul 7 2016.

[71] C. Schrader *et al.*, "GPi-DBS may induce a hypokinetic gait disorder with freezing of gait in patients with dystonia," *Neurology*, vol. 77, no. 5, pp. 483-488, 2011.

[72] J. Wertheimer *et al.*, "The impact of STN deep brain stimulation on speech in individuals with Parkinson's disease: the patient's perspective," *Parkinsonism & related disorders*, vol. 20, no. 10, pp. 1065-1070, 2014.

[73] G. Foffani *et al.*, "Focused ultrasound in Parkinson's disease: A twofold path toward disease modification," *Movement Disorders*, vol. 34, no. 9, pp. 1262-1273, 2019.

[74] S. Moosa, R. Martínez-Fernández, W. J. Elias, M. Del Alamo, H. M. Eisenberg, and P. S. Fishman, "The role of high-intensity focused ultrasound as a symptomatic treatment for Parkinson's disease," *Movement Disorders*, vol. 34, no. 9, pp. 1243-1251, 2019.

[75] H. Zach, M. Dirkx, B. R. Bloem, and R. C. Helmich, "The Clinical Evaluation of Parkinson's Tremor," *Journal of Parkinson's Disease*, vol. 5, pp. 471-474, 2015.

[76] M. George Krucik, MBA, "What causes muscle rigidity? 21 possible conditions," *http://www.healthline.com/symptom/muscle-rigidity*.

[77] K. A. Jellinger, "Neuropathology of sporadic Parkinson's disease: evaluation and changes of concepts," *Movement disorders*, vol. 27, no. 1, pp. 8-30, 2012.

[78] R. Xia and Z.-H. Mao, "Progression of motor symptoms in Parkinson's disease," *Neuroscience bulletin*, vol. 28, no. 1, pp. 39-48, 2012.

[79] R. C. Eberhart, "Tremor quantification using digital actigraphy," in *[Engineering in Medicine and Biology, 1999. 21st Annual Conference and the 1999 Annual Fall Meetring of the Biomedical Engineering Society] BMES/EMBS Conference, 1999. Proceedings of the First Joint*, 1999, vol. 1, p. 521 vol.1.

[80] A. Salarian, H. Russmann, C. Wider, P. R. Burkhard, F. J. G. Vingerhoets, and K. Aminian, "Quantification of Tremor and Bradykinesia in Parkinson's Disease Using a Novel Ambulatory Monitoring System," *IEEE Transactions on Biomedical Engineering,* vol. 54, no. 2, pp. 313-322, 2007.

[81] K. Harish, M. V. Rao, R. Borgohain, A. Sairam, and P. Abhilash, "Tremor quantification and its measurements on parkinsonian patients," in *2009 International Conference on Biomedical and Pharmaceutical Engineering*, 2009, pp. 1-3.

[82] G. Rigas, A. T. Tzallas, D. G. Tsalikakis, S. Konitsiotis, and D. I. Fotiadis, "Real-time quantification of resting tremor in the Parkinson's disease," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009, pp. 1306-1309.

[83] M. N. Alam, B. Johnson, J. Gendreau, K. Tavakolian, C. Combs, and R. Fazel-Rezai, "Tremor quantification of Parkinson's disease - a pilot study," in *2016 IEEE International Conference on Electro Information Technology (EIT)*, 2016, pp. 0755-0759.

[84] F. Corona, M. Pau, M. Guicciardi, M. Murgia, R. Pili, and C. Casula, "Quantitative assessment of gait in elderly people affected by Parkinson's Disease," in *2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2016, pp. 1-6.

[85] S. D. Din, A. Godfrey, and L. Rochester, "Validation of an Accelerometer to Quantify a Comprehensive Battery of Gait Characteristics in Healthy Older Adults and Parkinson's Disease: Toward Clinical and at Home Use," *IEEE Journal of Biomedical and Health Informatics,* vol. 20, no. 3, pp. 838-847, 2016.

[86] Y. Zhou, M. E. Jenkins, M. D. Naish, and A. L. Trejos, "The measurement and analysis of Parkinsonian hand tremor," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2016, pp. 414-417.

[87] V. Agostini, G. Balestra, and M. Knaflitz, "Segmentation and Classification of Gait Cycles," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 22, no. 5, pp. 946-952, 2014.

[88] F. Parisi *et al.*, "Inertial BSN-Based Characterization and Automatic UPDRS Evaluation of the Gait Task of Parkinsonians," *IEEE Transactions on Affective Computing,* vol. 7, no. 3, pp. 258-271, 2016.

[89] E. Bakstein, K. Warwick, J. Burgess, x00D, Stavdahl, and T. Aziz, "Features for detection of Parkinson's disease tremor from local field potentials of the subthalamic nucleus," in *Cybernetic Intelligent Systems (CIS), 2010 IEEE 9th International Conference on*, 2010, pp. 1-6.

[90] M. D. Djuri *et al.*, "Automatic Identification and Classification of Freezing of Gait Episodes in Parkinson's Disease Patients," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 22, no. 3, pp. 685-694, 2014.

[91] J. R. Duann, C. H. Lin, C. M. Chen, M. K. Lu, C. H. Tsai, and J. C. Chiou, "Anatomic differences between Parkinson's disease and essential tremor using ICA-based brain morphometry," in *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, 2013, pp. 573-576.

[92] C. K. Liao, C. D. Lim, C. Y. Cheng, C. M. Huang, and L. C. Fu, "Vision based gait analysis on robotic walking stabilization system for patients with Parkinson's Disease," in *2014 IEEE International Conference on Automation Science and Engineering (CASE)*, 2014, pp. 818-823.

[93] M. Bachlin *et al.*, "Wearable Assistant for Parkinson's Disease Patients With the Freezing of Gait Symptom," *IEEE Transactions on Information Technology in Biomedicine,* vol. 14, no. 2, pp. 436-446, 2010.

[94] K. N. Winfree, I. Pretzer-Aboff, D. Hilgart, R. Aggarwal, M. Behari, and S. Agrawal, "An untethered shoe with vibratory feedback for improving gait of Parkinson's Patients: The PDShoe," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 1202-1205.

[95] S. V. Perumal and R. Sankar, "Gait and tremor assessment for patients with Parkinson’s disease using wearable sensors," *Ict Express,* vol. 2, no. 4, pp. 168-174, 2016.

[96] S. V. Perumal and R. Sankar, "Gait monitoring system for patients with Parkinson's disease using wearable sensors," in *2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT)*, 2016, pp. 21-24: IEEE.

[97] J. I. Pan and Y. C. Huang, "Intelligent fall prevention for Parkinson's disease patients based on detecting posture instabilily and freezing of gait," in *Informatics in Control, Automation and Robotics (ICINCO), 2015 12th International Conference on*, 2015, vol. 01, pp. 608-613.

[98] H. Uchitomi *et al.*, "Interpersonal synchrony-based dynamic stabilization of the gait rhythm between human and virtual robot - Clinical application to festinating gait of Parkinson's disease patient," in *Micro-NanoMechatronics and Human Science (MHS), 2012 International Symposium on,* 2012, pp. 460-465.

[99] Y. Zhao *et al.*, "A novel wearable laser device to regulate stride length in Parkinson's disease," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 5895-5898.

[100] R. J. L. M. Verstappen, C. T. Freeman, E. Rogers, T. Sampson, and J. H. Burridge, "Robust higher order repetitive control applied to human tremor suppression," in *2012 IEEE International Symposium on Intelligent Control*, 2012, pp. 1214-1219.

[101] S. Boksuwan, T. Benjanarasuth, C. Kanamori, and H. Aoyama, "Tremor suppression for handheld micromanipulator using robust hybrid control," in *Control, Automation and Systems (ICCAS), 2014 14th International Conference on*, 2014, pp. 267-271.

[102] S. A. Khoshnevis, I.-H. Ra, and R. Sankar, "Early Stage Diagnosis of Parkinson's Disease Using HOS Features of EEG Signals," 2020.

[103] S. A. Khoshnevis and R. Sankar, "Classification of the stages of Parkinson's disease using novel higher-order statistical features of EEG signals," *Neural Computing and Applications,* pp. 1-13, 2020.

[104] S. A. Khoshnevis, S. B. Appakaya, E. Sheybani, and R. Sankar, "Compression of Gait IMU signals Using Sensor Fusion and Compressive Sensing," in *2020 Wireless Telecommunications Symposium (WTS)*, 2020, pp. 1-5: IEEE.

[105] N. Kostikis, D. Hristu-Varsakelis, M. Arnaoutoglou, and C. Kotsavasiloglou, "A Smartphone-Based Tool for Assessing Parkinsonian Hand Tremor," *IEEE Journal of Biomedical and Health Informatics,* vol. 19, no. 6, pp. 1835-1842, 2015.

[106] S. Mazilu, U. Blanke, and G. Troster, "Gait, wrist, and sensors: Detecting freezing of gait in Parkinson's disease from wrist movement," in *Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on*, 2015, pp. 579-584.

[107] L. Pepa, L. Ciabattoni, F. Verdini, M. Capecci, and M. G. Ceravolo, "Smartphone based Fuzzy Logic freezing of gait detection in Parkinson's Disease," in *Mechatronic and Embedded Systems and Applications (MESA), 2014 IEEE/ASME 10th International Conference on*, 2014, pp. 1-6.

[108] S. Mazilu *et al.*, "GaitAssist: A wearable assistant for gait training and rehabilitation in Parkinson's disease," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, 2014, pp. 135-137.

[109] R. LeMoyne, T. Mastroianni, M. Cozza, C. Coroian, and W. Grundfest, "Implementation of an iPhone for characterizing Parkinson's disease tremor through a wireless accelerometer application," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 4954-4958.

[110] C. Segrin, "Social skills deficits associated with depression," *Clinical psychology review*, vol. 20, no. 3, pp. 379-403, 2000.

[111] M. Belinchón Carmona, J. M. Igoa González, and A. Rivière Gómez, *Psicología del lenguaje: investigación y teoría* (no. 401.9 B4). 1994.

[112] H. Hazan, D. Hilu, L. Manevitz, L. O. Ramig, and S. Sapir, "Early diagnosis of Parkinson's disease via machine learning on speech data," in *Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, 2012, pp. 1-4.

[113] K. Tjaden, "Speech and Swallowing in Parkinson’s Disease," *Topics in geriatric rehabilitation*, vol. 24, no. 2, pp. 115-126, 2008.

[114] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of speech and hearing research*, vol. 12, no. 2, pp. 246-269, 1969.

[115] S. Pinto, A. Ghio, B. Teston, and F. Viallet, "Dysarthria across Parkinson's disease progression. Natural history of its components: Dysphonia, dysprosody and dysarthria," *Revue neurologique*, vol. 166, no. 10, pp. 800-810, 2010.

[116] L. O. Ramig, C. Fox, and S. Sapir, "Speech treatment for Parkinson’s disease," *Expert Review of Neurotherapeutics*, vol. 8, no. 2, pp. 297-309, 2008/02/01 2008.

[117] F. Quek, R. Bryll, M. Harper, C. Lei, and L. Ramig, "Audio and vision-based evaluation of parkinson's disease from discourse video," in *Bioinformatics and Bioengineering Conference, 2001. Proceedings of the IEEE 2nd International Symposium on*, 2001, pp. 245-252.

[118] J. Orozco-Arroyave *et al.*, "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 481-500, 2016.

[119] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 181-190, 2013.

[120] C. Blog, "Effects of Intensive Voice Treatment (LSVT) on Vowel Articulation in Dysarthric Individuals With Idiopathic Parkinson Disease: Acoustic and Perceptual Findings Shimon Sapir, Jennifer L. Spielman, Lorraine O. Ramig, Brad H. Story, and Cynthia Fox," *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 899-912, 2018.

[121] K. Tjaden, J. Lam, and G. Wilding, "Vowel acoustics in Parkinson's disease and multiple sclerosis: comparison of clear, loud, and slow speaking conditions," (in eng), *Journal of speech, language, and hearing research : JSLHR*, vol. 56, no. 5, pp. 1485-1502, 2013.

[122] M. Asgari and I. Shafran, "Extracting cues from speech for predicting severity of Parkinson'S disease," in *2010 IEEE International Workshop on Machine Learning for Signal Processing*, 2010, pp. 462-467.

[123] M. Novotny, J. Pospisil, C. R, and J. Rusz, "Automatic detection of voice onset time in dysarthric speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4340-4344.

[124] T. Bocklet, N. E, G. Stemmer, H. Ruzickova, and J. Rusz, "Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, 2011, pp. 478-483.

[125] H. Dubey, J. C. Goldberg, K. Mankodiya, and L. Mahler, "A multi-smartwatch system for assessing speech characteristics of people with dysarthria in group settings," in *2015 17th International Conference on E-health Networking, Application & Services (HealthCom)*, 2015, pp. 528-533.

[126] T. Villa-Canas, J. D. Arias-Londono, J. F. Vargas-Bonilla, and J. R. Orozco-Arroyave, "Time-frequency approach in continuous speech for detection of Parkinson's disease," in *Signal Processing, Images and Computer Vision (STSIVA), 2015 20th Symposium on*, 2015, pp. 1-6.

[127] S. Zhao, F. Rudzicz, L. G. Carvalho, C. Marquez-Chin, and S. Livingstone, "Automatic detection of expressed emotion in Parkinson's Disease," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4813-4817.

[128] E. O, A. Karatutlu, and U. C, "Detection of Parkinson's disease from vocal features using random subspace classifier ensemble," in *2015 Twelve International Conference on Electronics Computer and Computation (ICECCO)*, 2015, pp. 1-4.

[129] Y. Medan and E. Yair, "Pitch synchronous spectral analysis scheme for voiced speech," *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, no. 9, pp. 1321-1328, 1989.

[130] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015-1022, 2009.

[131] E. A. Belalcazar-Bolanos, J. R. Orozco-Arroyave, J. D. Arias-Londono, J. F. Vargas-Bonilla, and N. E, "Automatic detection of Parkinson's disease using noise measures of speech," in *Symposium of Signals, Images and Artificial Vision - 2013: STSIVA - 2013*, 2013, pp. 1-5.

[132] J. R. Orozco-Arroyave *et al.*, "Characterization Methods for the Detection of Multiple Voice Disorders: Neurological, Functional, and Laryngeal Diseases," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1820-1828, 2015.

[133] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 594-597.

[134] G. P, Vilda *et al.*, "Monitoring Parkinson's Disease from phonation improvement by Log Likelihood Ratios," in *Bioinspired Intelligence (IWOBI), 2015 4th International Work Conference on*, 2015, pp. 105-110.

[135] Z. Smekal, J. Mekyska, Z. Galaz, Z. Mzourek, I. Rektorova, and M. Faundez-Zanuy, "Analysis of phonation in patients with Parkinson's disease using empirical mode decomposition," in *Signals, Circuits and Systems (ISSCS), 2015 International Symposium on*, 2015, pp. 1-4.

[136] J. Mekyska *et al.*, "Assessing progress of Parkinson's disease using acoustic analysis of phonation," in *Bioinspired Intelligence (IWOBI), 2015 4th International Work Conference on*, 2015, pp. 111-118.

[137] A. Frid, H. Hazan, D. Hilu, L. Manevitz, L. O. Ramig, and S. Sapir, "Computational Diagnosis of Parkinson's Disease Directly from Natural Speech Using Machine Learning Techniques," in *Software Science, Technology and Engineering (SWSTE), 2014 IEEE International Conference on*, 2014, pp. 50-53.

[138] V. J. C, Correa, J. R. Orozco-Arroyave, J. D. Arias-Londono, J. F. Vargas-Bonilla, and N. E, "Design and implementation of an embedded system for real time analysis of speech from people with Parkinson's disease," in *Symposium of Signals, Images and Artificial Vision - 2013: STSIVA - 2013*, 2013, pp. 1-5.

[139] Vikas and R. K. Sharma, "Early detection of Parkinson's disease through Voice," in *Advances in Engineering and Technology (ICAET), 2014 International Conference on*, 2014, pp. 1-5.

[140] J. Mekyska, I. Rektorova, and Z. Smekal, "Selection of optimal parameters for automatic analysis of speech disorders in Parkinson's disease," in *Telecommunications and Signal Processing (TSP), 2011 34th International Conference on*, 2011, pp. 408-412.

[141] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech," (in eng), *Journal of speech, language, and hearing research : JSLHR*, vol. 53, no. 1, pp. 114-125, 2010.

[142] A. Monteiro, H. Dubey, L. Mahler, Q. Yang, and K. Mankodiya, "Fit: A Fog Computing Device for Speech Tele-Treatments," in *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2016, pp. 1-3.

[143] M. Asgari and I. Shafran, "Predicting severity of Parkinson's disease from speech," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 5201-5204.

[144] V. Narang, D. Misra, and G. Dalal, "Acoustic Space in Motor Disorders of Speech: Two Case Studies," in *Asian Language Processing (IALP), 2011 International Conference on*, 2011, pp. 211-215.

[145] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective Automatic Assessment of Rehabilitative Speech Treatment in Parkinson's Disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 181-190, 2014.

[146] W. Ji and Y. Li, "Stable dysphonia measures selection for Parkinson speech rehabilitation via diversity regularized ensemble," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2264-2268.

[147] J. R. Orozco-Arroyave *et al.*, "Towards an automatic monitoring of the neurological state of Parkinson's patients from speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6490-6494.

[148] A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad, "Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease," in *Electrical and Information Technologies (ICEIT), 2015 International Conference on*, 2015, pp. 300-304.

[149] A. Bourouhou, A. Jilbab, C. Nacir, and A. Hammouch, "Comparison of classification methods to detect the Parkinson disease," in *2016 International Conference on Electrical and Information Technologies (ICEIT)*, 2016, pp. 421-424.

[150] A. Benba, A. Jilbab, and A. Hammouch, "Hybridization of best acoustic cues for detecting persons with Parkinson's disease," in *Complex Systems (WCCS), 2014 Second World Conference on*, 2014, pp. 622-625.

[151] M. Shahbakhti, D. Taherifar, and A. Sorouri, "Linear and non-linear speech features for detection of Parkinson's disease," in *Biomedical Engineering International Conference (BMEiCON), 2013 6th*, 2013, pp. 1-3.

[152] E. A. Belalcazar-Bolanos, J. D. Arias-Londono, J. F. Vargas-Bonilla, and J. R. Orozco-Arroyave, "Nonlinear glottal flow features in Parkinson's disease detection," in *Signal Processing, Images and Computer Vision (STSIVA), 2015 20th Symposium on*, 2015, pp. 1-6.

[153] M. Novotny, J. Rusz, C. R, and R. E, "Automatic Evaluation of Articulatory Disorders in Parkinson's Disease," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1366-1378, 2014.

[154] B. E. Sakar, C. O. Sakar, G. Serbes, and O. Kursun, "Determination of the optimal threshold value that can be discriminated by dysphonia measurements for unified Parkinson's Disease rating scale," in *Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on*, 2015, pp. 1-4.

[155] A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad, "Using RASTA-PLP for discriminating between different Neurological diseases," in *2016 International Conference on Electrical and Information Technologies (ICEIT)*, 2016, pp. 406-409.

[156] M. S. Wibawa, H. A. Nugroho, and N. A. Setiawan, "Performance evaluation of combined feature selection and classification methods in diagnosing parkinson disease based on voice feature," in *2015 International Conference on Science in Information Technology (ICSITech)*, 2015, pp. 126-131.

[157] S. Jain and S. Shetty, "Improving accuracy in noninvasive telemonitoring of progression of Parkinson's Disease using two-step predictive model," in *2016 Third International Conference on Electrical, Electronics, Computer Engineering and their Applications (EECEA)*, 2016, pp. 104-109.

[158] B. E. Sakar *et al.*, "Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828-834, 2013.

[159] W. Caesarendra, F. T. Putri, M. Ariyanto, and J. D. Setiawan, "Pattern recognition methods for multi stage classification of parkinson's disease utilizing voice features," in *2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, 2015, pp. 802-807.

[160] M. F. CAGLAR, B. CETISLI, and I. B. TOPRAK, "Automatic Recognition of Parkinson's Disease from Sustained Phonation Tests Using ANN and Adaptive Neuro-Fuzzy Classifier," *Journal of Engineering Science and Design*, vol. 1, no. 2, pp. 59-64, 2010.

[161] M. Shahbakhti, D. Taherifar, and Z. Zareei, "Combination of PCA and SVM for diagnosis of Parkinson's disease," in *2013 2nd International Conference on Advances in Biomedical Engineering*, 2013, pp. 137-140.

[162] Q. W. Oung *et al.*, "Objective assessment of Parkinson's disease symptoms severity: A review," in *Biomedical Engineering (ICoBE), 2015 2nd International Conference on*, 2015, pp. 1-6.

[163] M. S. Islam, I. Parvez, D. Hai, and P. Goswami, "Performance comparison of heterogeneous classifiers for detection of Parkinson's disease using voice disorder (dysphonia)," in *Informatics, Electronics & Vision (ICIEV), 2014 International Conference on*, 2014, pp. 1-7.

[164] P. Kraipeerapun and S. Amornsamankul, "Using stacked generalization and complementary neural networks to predict Parkinson's disease," in *Natural Computation (ICNC), 2015 11th International Conference on*, 2015, pp. 1290-1294.

[165] M. Su and K. S. Chuang, "Dynamic feature selection for detecting Parkinson's disease through voice signal," in *RF and Wireless Technologies for Biomedical and Healthcare Applications (IMWS-BIO), 2015 IEEE MTT-S 2015 International Microwave Workshop Series on*, 2015, pp. 148-149.

[166] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests," *IEEE Transactions on Biomedical Engineering,* vol. 57, no. 4, pp. 884-893, 2010.

[167] H. Gunduz, "Deep learning-based Parkinson’s disease classification using vocal feature sets," *IEEE Access,* vol. 7, pp. 115540-115551, 2019.

[168] W. Rahman *et al.*, "Detecting Parkinson's Disease from Speech-task in an accessible and interpretable manner," *arXiv preprint arXiv:2009.01231,* 2020.

[169] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H. Ghomi, "Parkinson’s disease diagnosis using machine learning and voice," in *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2018, pp. 1-7: IEEE.

[170] J. C. Vásquez-Correa *et al.*, "Convolutional neural networks and a transfer learning strategy to classify parkinson’s disease from speech in three different languages," in *Iberoamerican Congress on Pattern Recognition*, 2019, pp. 697-706: Springer.

[171] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and E. Nöth, "Deep learning approach to Parkinson’s disease detection using voice recordings and convolutional neural network dedicated to image classification," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 717-720: IEEE.

[172] D. R. Rizvi, I. Nissar, S. Masood, M. Ahmed, and F. Ahmad, "An LSTM based Deep learning model for voice-based detection of Parkinson’s disease," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, p. 8, 2020.

[173] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson’s disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350-367, 2011.

[174] J. Rusz *et al.*, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171-2181, 2013.

[175] M. Novotný, J. Rusz, R. Čmejla, and E. Růžička, "Automatic evaluation of articulatory disorders in Parkinson’s disease," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1366-1378, 2014.

[176] J. R. Orozco-Arroyave *et al.*, "Automatic detection of Parkinson's disease from words uttered in three different languages," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[177] S. Skodda, W. Grönheit, and U. Schlegel, "Intonation and speech rate in Parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission," *Journal of Voice*, vol. 25, no. 4, pp. e199-e205, 2011.

[178] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues," in *Interspeech*, 2013, pp. 1149-1153.

[179] 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: ACM.

[180] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB. 2006," *URL http://www. ee. ic. ac. uk/... hp/staff/dmb/voicebox/voicebox. html. Available online*, 2003.

[181] S. Kim, T. Eriksson, H.-G. Kang, and D. H. Youn, "A pitch synchronous feature extraction method for speaker recognition," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 1, pp. I-405: IEEE.

[182] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," *Unpublished manuscript*, 1995.

[183] D. Giovanni and G. Francesco. "Italian Parkinson's Voice and Speech" [Online]. Available: http://dx.doi.org/10.21227/aw6b-tg17

[184] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system," *IEEE Access,* vol. 5, pp. 22199-22208, 2017.

[185] M. D. Skowronski, R. Shrivastav, J. Harnsberger, S. Anand, and J. Rosenbek, "Acoustic discrimination of Parkinsonian speech using cepstral measures of articulation," *The Journal of the Acoustical Society of America,* vol. 132, no. 3, pp. 2089-2089, 2012.

[186] B. Karan, S. S. Sahu, and K. Mahto, "Stacked auto-encoder based Time-frequency features of Speech signal for Parkinson disease prediction," in *2020 International Conference on Artificial Intelligence and Signal Processing (AISP),* 2020, pp. 1-4: IEEE.

[187] A. Lauraitis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Detection of Speech Impairments Using Cepstrum, Auditory Spectrogram and Wavelet Time Scattering Domain Features," *IEEE Access*, 2020.

[188] L. Zahid *et al.*, "A Spectrogram-Based Deep Feature Assisted Computer-Aided Diagnostic System for Parkinson's Disease," *IEEE Access,* vol. 8, pp. 35482-35495, 2020.

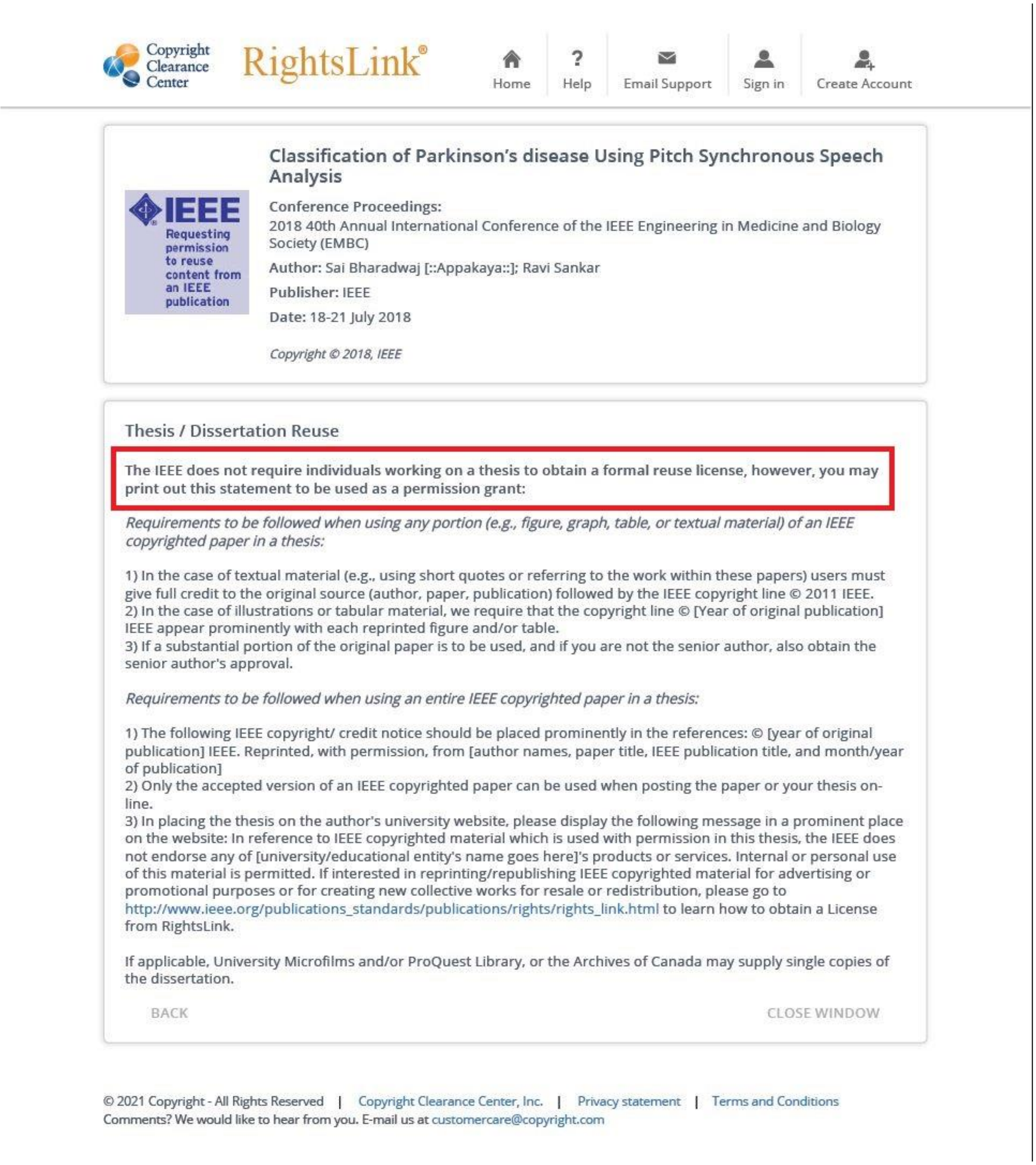
[189] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE transactions on neural networks and learning systems,* vol. 28, no. 10, pp. 2222-2232, 2016.

[190] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n,* vol. 93, p. 27403, 1993.

[191] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16),* 2016, pp. 265-283.

Appendix A: Copyright Permissions

The permission below is for the use of published content in Chapter 3, Section 3.2


Copyright Clearance Center
RightsLink®
Home
Help
Email Support
Sign in
Create Account
IEEE
Requesting permission to reuse content from an IEEE publication

Classification of Parkinson's disease Using Pitch Synchronous Speech Analysis

Conference Proceedings:
2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

Author: Sai Bharadwaj [::Appakaya::]; Ravi Sankar

Publisher: IEEE

Date: 18-21 July 2018

Copyright © 2018, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

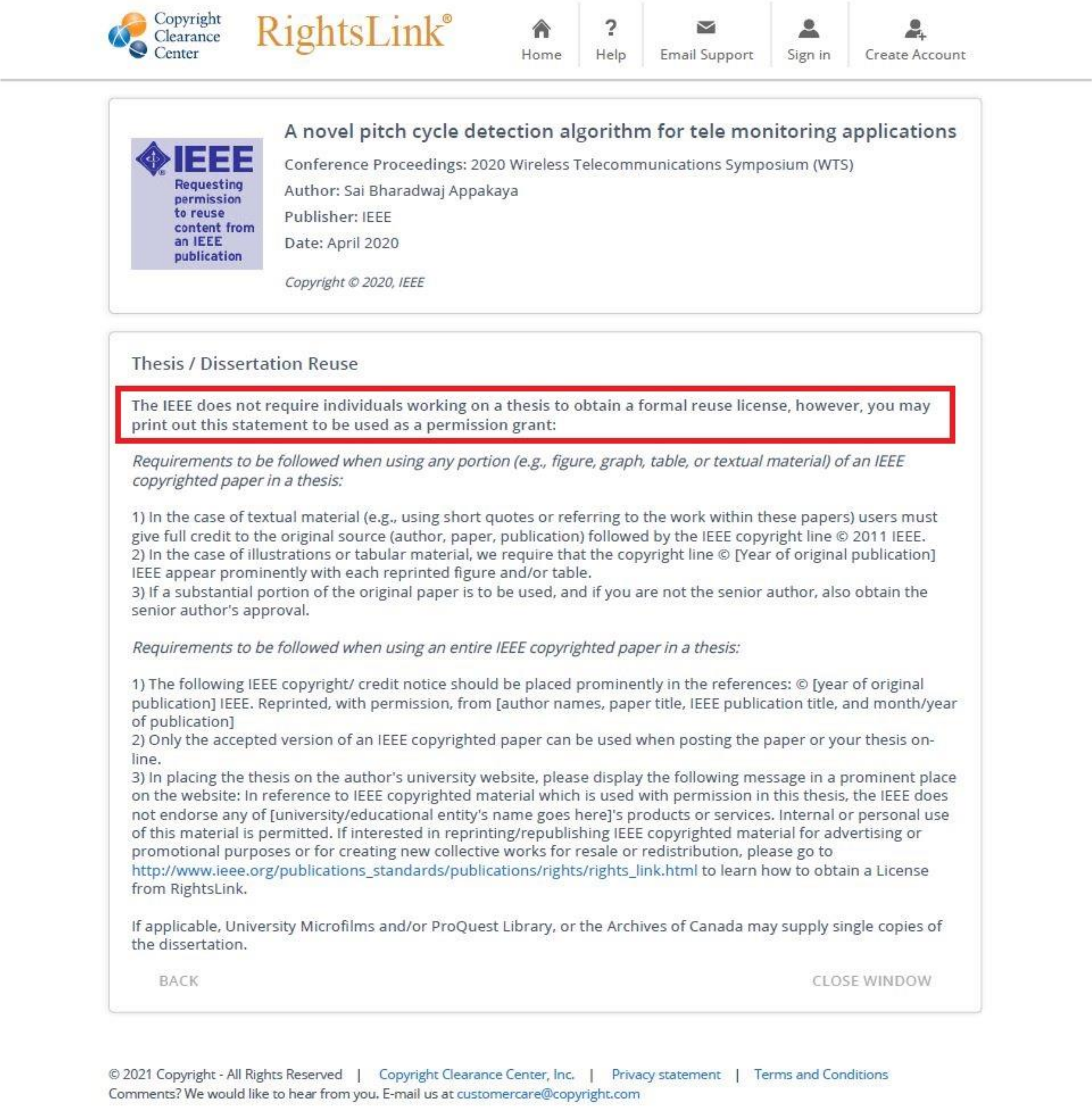
1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis online.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

The permission below is for the use of published content in Chapter 3, Section 3.3


Copyright Clearance Center
RightsLink®
Home
Help
Email Support
Sign in
Create Account
IEEE
Requesting permission to reuse content from an IEEE publication

A novel pitch cycle detection algorithm for tele monitoring applications

Conference Proceedings: 2020 Wireless Telecommunications Symposium (WTS)
Author: Sai Bharadwaj Appakaya
Publisher: IEEE
Date: April 2020

Copyright © 2020, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.